ENHANCEMENT OF PARALLEL K-MEANS ALGORITHM FOR CLUSTERING BIG DATASETS

ARDAVAN ASHABI

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia

DECEMBER 2022

DEDICATION

To my family

ACKNOWLEDGEMENT

First and foremost, thank Allah, the compassionate and the merciful, for providing me the opportunity and the ability to reach this point.

I am grateful to my family for all their love, care, and encouragement. Their endless support and patience were my companion in every aspect of my life. I am also very grateful to my friends who became my family.

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed to my understanding and thoughts. In particular, I wish to express my sincere appreciation to my supervisor Professor Dr. Shamsul Bin Sahibuddin for his guidance throughout my entire PhD. His comments provided great insight into the current research.

I also owe an immense gratitude to Dr. Mehdi Salkhordeh Haghighi for his kind information sharing, support, and advice.

To each of the above, I extend my deepest appreciation.

ABSTRACT

Big Data encompasses huge amounts of complex data which is generated in different areas such as business, marketing, educational systems, IoT, and healthcare. For instance, in the healthcare domain, huge amounts of data are generated daily from different sources such as health monitoring and medical diagnosis systems by health service providers. Data mining aims to extract meaningful and valuable patterns from a set of raw data to transform data into meaningful information for better decision-making. However, Big Data is very complex and voluminous, and traditional methods of Data Mining are not capable to process and analyze this data efficiently. Data clustering, one of the main methods of data mining, eases the extraction of information from each cluster separately. Since 1960s, K-means algorithm has been known as one of the most classical techniques of data clustering. Even though there has been an extremely rich bibliography about improving the efficiency of K-means for years now, traditional K-means still suffers from some weaknesses, especially in dealing with Big Data. Despite many attempts to optimize K-means algorithm to handle Big Data using different techniques such as parallelization, the proposed methods are still not able to cluster Big Datasets efficiently due to lack of improvement in some effective parameters such as the number of clusters and the initial clusters' centroids. This study aims to understand the current limitations of K-means algorithm and to overcome the limitations in order to produce more efficient performance in clustering big datasets from healthcare domain. To develop the optimized extension of K-means algorithm, a systematic literature review (SLR) was conducted to investigate the current limitations and existing solutions for the K-means limitations over Big Data. Based on the the SLR, this study proposed an enhanced parallel version of K-means clustering algorithm to reduce the execution time of the clustering process over the big datasets with the minimum negative impact on the clustering's accuracy. Determining the optimum number of clusters, obtaining the suitable initial centroids, and improving the process of parallelization were the three steps of the optimization process. To avoid any random results, the proposed hybrid solution defined the optimum number of clusters by using elbow method. In addition, the proposed algorithm obtained the ideal initial centroids by utilizing a careful seed selection method, performing K-means with a fuzzy technique to increase the precision of the clustering, and parallelizing the clustering process by using Hadoop platform with the optimized Map and Reduce functions to reduce the execution time of the process. The evaluation of the proposed algorithm revealed that the new method performed the clustering process over multiple big datasets with shorter execution time compared to the study's benchmarks: Apache Mahout K-means, K-means++, and Fuzzy K-means. Also, the results of the three selected cluster validity indices - Silhouette, Dunn, and Davies-Bouldin - verified that there was no negative impact on the quality of the clusters.

ABSTRAK

Data Raya merangkumi sejumlah besar data kompleks yang dihasilkan dalam pelbagai bidang seperti perniagaan, pemasaran, sistem pendidikan, IoT, dan penjagaan kesihatan. Sebagai contoh, dalam domain penjagaan kesihatan, sejumlah besar data dihasilkan setiap hari daripada sumber yang berbeza seperti sistem pemantauan kesihatan dan diagnosis perubatan oleh penyedia perkhidmatan kesihatan. Pengumpulan data ini bertujuan untuk mengekstrak corak yang bermakna dan berharga dari sekumpulan data mentah untuk mengubah data menjadi maklumat yang bermakna untuk membuat keputusan yang lebih baik. Walau bagaimanapun, Data Raya sangat kompleks, terlalu banyak dan kaedah tradisional pengumpulan data tidak mampu memproses dan menganalisis data ini dengan cekap. Pengelompokan data merupakan salah satu kaedah pengumpulan data utama, dan ia memudahkan pengekstrakan maklumat dari setiap kelompok secara berasingan. Sejak tahun 1960-an, algoritma K-*means* telah dikenali sebagai salah satu teknik pengelompokan data yang paling klasik. Walaupun telah ada bibliografi yang sangat luas tentang meningkatkan K-*means* selama bertahun-tahun sekarang, K-*means* tradisional masih mengalami beberapa kekurangan, terutamanya dalam menangani Data Raya. Kajian ini bertujuan untuk memahami kekurangan algoritma K-*means* semasa dan untuk mengatasi kekurangan ini bagi menghasilkan prestasi yang lebih cekap dalam mengelompokkan kumpulan data besar dari domain penjagaan kesihatan. Untuk membangunkan perkembangan algoritma K-*means* secara optimum, tinjauan literatur sistematik (SLR) telah dilakukan untuk mengkaji batasan semasa dan penyelesaian sedia ada untuk batasan K-*means* ke atas Data Raya. Berdasarkan penemuan SLR, kajian ini mencadangkan versi selari yang dipertingkatkan bagi algoritma pengelompokan K-*means* untuk mengurangkan masa pelaksanaan proses pengelompokan ke atas kumpulan Data Raya dengan kesan negatif minimum terhadap ketepatan pengelompokan. Menentukan bilangan pengelompokan yang optimum, mendapatkan pusat jisim awal yang sesuai dan menambah baik proses penyelarasan adalah tiga langkah proses pengoptimuman. Untuk mengelakkan sebarang hasil rawak, penyelesaian hibrid yang dicadangkan telah menentukan bilangan kelompok yang optimum dengan menggunakan kaedah *Elbow*. Di samping itu, algoritma yang dicadangkan memperoleh pusat jisim yang ideal dengan menggunakan kaedah pemilihan data yang teliti, melakukan kaedah K-*means* dengan teknik *Fuzzy* untuk meningkatkan ketepatan pengelompokan dan menyelaraskan proses pengelompokan dengan menggunakan platform *Hadoop* dengan fungsi *Map* dan *Reduce* yang dioptimumkan untuk mengurangkan waktu pelaksanaan proses. Penilaian algoritma yang dicadangkan menunjukkan bahawa kaedah baharu melakukan proses pengelompokan ke atas banyak kumpulan Data Raya dengan masa pelaksanaan yang lebih singkat berbanding dengan penanda aras kajian: *Apache Mahout K-means*, *K-means++*, dan *Fuzzy K-means*). Selain itu, keputusan tiga indeks kelompok yang dipilih - *Silhouette*, *Dunn*, dan *Davies-Bouldin* - mengesahkan bahawa tidak ada kesan negatif terhadap kualiti pengelompokan tersebut.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PSEUDOCODES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

AWS	-	Amazon Web Services
BA	-	Business Analytics
BDM	-	Big Data Mining
BI	-	Business Intelligence
CLARA	-	Clustering LARge Applications
CVI	-	Cluster Validity Indices
DBI	-	Davies-Bouldin Index
DI	-	Dunn Index
DT	-	Decision Tree
DM	-	Data Mining
DSS	-	Decision Support Systems
EBS	-	Evidence-Based Medicine
HER	-	Electronic Health Records
EMR	-	Elastic MapReduce
ETL	-	Extract, Transform, Load
GPS	-	Global Positioning System
HDFS	-	Hadoop distributed file system
IoT	-	Internet of Thing
KDD	-	Knowledge Discovery in Databases
KNN	-	K-Nearest Neighbor
OCD	-	Optimized Cluster
PSM	-	Probabilistic Subtyping Model
RDBMS	-	Relational Data Base Management Systems
SAS	-	Statistical Analysis System

SI	-	Silhouette Index
SLR	-	Systematic Literature Review
SVM	-	Support Vector Machine
WCSS	-	Within a cluster Sum of Square

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Big Data is described as a massive amount of complex data. For Industries, it is considered a valuable opportunity to gain insights for improving their services. To extract such insights, using different Data Mining (DM) techniques is required. Clustering is one of the essential methods for DM and Big Data analysis. However, because of the issues and challenges that have been recently raised in manipulating large volumes of data as well as its complexity, applying the clustering methods to big datasets is faced with some obstacles. The question is how to tackle this issue and how to apply clustering techniques to Big Data more efficiently.

One of the fields which has been faced with the rapid rate of generating Big Data is healthcare domain. Healthcare domain includes the comprehensive practices of prevention, diagnosis, and treatment of diseases, damages, and mental and physical impairments in people (Yang et al., 2015). Life sciences have always been a discerning area that requires fast-growing innovations for better wellbeing of the people (Sreedevi et al., 2022). With the growth of healthcare technologies and the growing digitization in healthcare domain, huge amounts of data have been generating every day from different sources, such as medical record devices, patient care, and monitoring devices (Piri, 2017). This sizable amount of data has been collecting through the healthcare systems regularly.

The vast access to the data which has been generated in healthcare domain, as well as rapid progress in DM and machine learning tools and techniques, has generated an improving area of healthcare analytics. The improvement of Decision Support Systems (DSS) by data scientists and using medical scientists' knowledge in this scope has simplified the clinical procedures, also simplified the clinicians and physicians' tasks as well. Analyzing the data which is generated in healthcare domain and applying DM and machine learning techniques to them has several benefits. For instance, Medical staffs are able to categorize the patients in accordance to the level of disease severity or patients' condition and, as a consequence, suitable treatments can be provided for each group; risk factors of different diseases can be identified, leading potentially to better health management; and diseases can be detected at early stages, allowing for appropriate interventions and treatments (W. Raghupathi & Raghupathi, 2014).

As mentioned earlier, clustering as a primary task in DM, is described as finding heterogeneous groups of data using some dissimilarity criterion (Khanmohammadi et al., 2017). In healthcare domain, clustering, as an evidence-based medicine (EBM) analysis system, is used in order to reduce medical errors. Moreover, clustering can help in finding the patterns of different diseases from stored medical data. This data can be gathered from different sources and during various activities such as monitoring screening, therapy, biomedical and biological analysis, epidemiological studies, hospital management, medical instruction, etc. Efficient clustering tools reduce the demands on costly healthcare resources. These tools can assist physicians to cope with information overload, and also can help in future planning for enhanced services (Purandhar et al., 2021). Clustering results are used to study independence or correlation among diseases, and also to have more comprehensive insight into medical survey data. The mentioned benefits encourage data scientists and researchers to propose more efficient techniques for medical data clustering (Kalyani, 2012).

However, the massive and complex data which has been generating every day has a negative impact on the performance of data analysis techniques as well as clustering methods, therefore despite improving the clustering techniques over decades, these methods still need to be improved further.

1.2 Background of Study

With the new trends in data generation and collection, there exist larger numbers of data streams being transferred or created each day. According to IBM report, in 2017, every day almost 2.5 exabytes of data has been generated (IBM, 2017). With the rapid growth of generating data, it is expected that in 2025 this amount reach 463 exabytes per day (Vuleta, 2020). These data streams come from virtually everywhere, from sensors used in commonly used devices like cellphones, cars, and houses, to contents posted online on blogs and websites. Furthermore, some sources of data include multimedia content being uploaded and downloaded such as pictures, videos, movies, and songs, to the sensors used in gathering information for weather forecasts, purchase and transaction records, school and academic records, and GPS information amongst others (Mayer-Schonberger & Cukier, 2013). Therefore, the term of Big Data was introduced to refer to such volumes of data. From this view, Big Data can be defined as the massive amount of complex data that is generated in various domains and continue to grow rapidly (Kumar, 2017).

Essentially, with the increase in applications of big data, the need to analyze, process, and extract significant and practical knowledge from these data streams has also been growing in demand. By emerging new application areas which produce huge amounts of data, finding solutions to develop new techniques that can help to make sense of these data and turn them into useful knowledge is also needed. DM includes techniques that are becoming invaluable in helping to extract patterns from existing data. By DM techniques, it is possible to interpret the pattern and use it effectively during the decision-making process (Hand et al., 2001).

The significance of DM cannot go unnoticed, especially when it indirectly contributes to decision-making in many aspects of our daily lives. It is essential to learn from the past and study the history of patterns in order to make decisions for a better future based on these past patterns. These patterns, in the form of data, can be observed in all areas of professional and private lives, in sectors such as finance and banking, marketing, retail sales, and healthcare (Alsayat, 2016). Information is also gathered for many activities as well, such as population study, human migration, health, science, and education. By exploring these data, it is feasible to establish and analyze any pattern and plan ahead for a better future (Bifet, 2009).

According to (Gupta et al., 2014), Big Data Mining (BDM) describes as the process of looking for significant and practical information in huge volumes of data. Big data examples can be found in various areas such as social networking sites, sensor networks, atmospheric science, astronomy, life sciences, natural disaster and resource management, medical science, mobile phones, government data, weblogs, scientific research, and telecommunications (Haoxiang & Smys, 2021).

We consider the need for DM techniques in the healthcare domain. The healthcare domain has been presumed as a scope with numerous valuable data. In most countries, the healthcare sector is evolving rapidly and every day, a huge volume of data such as administrative reports, electronic medical records, and other benchmarking findings has been generated in this area (Jothi et al., 2015). The healthcare industry is responsible for the production of large data streams such as patient medical records, doctors' reports, and information on different drugs, diseases, and treatments. In these application areas, huge amounts of data are being produced continuously, therefore, intense analysis is also needed to establish trends, which could help to enhance better treatment methods, disease prevention, and guide overall healthcare practices. The necessity when it comes to the identification and classification of unidentified important information in the medical field which has many patients ensures that DM techniques are effective in healthcare, treatment, and diagnosis. These DM techniques enable health caregivers and researchers in the field to make effective policies in healthcare, come up with drug-recommending structures, and compile health profiles of various individuals (Koh & Tan, 2011).

According to (Jothi et al., 2015) DM is being used to search and find important and practical information among the massive amounts of data in the healthcare area. Furthermore, it is used to predict numerous diseases and also assist doctors in diagnosing and making their clinical decisions.

All medical data which is related to both patients and healthcare service providers are valuable. The researchers use DM tools and techniques in distributed medical environments to deliver improved and advanced medical services to a large number of people with more efficient resources management, more efficient customer relationship management, healthcare resources management, lower cost, etc. This significant knowledge which is provided by these tools and techniques may assist managements to make better decisions, such as the selection of treatments, disease prediction, decisions regarding health insurance policy, evaluation of medical staff, etc., (McGregor et al., 2012) (Bellazzi & Zupan, 2008) (Harper, 2005) (Stel et al., 2008). There are many issues and challenges of DM in the healthcare application area (Hosseinkhah et al., 2009) (Bellazzi & Zupan, 2008). In order to predict various diseases in a population area, effective DM methods may be used (Ahmad et al., 2015) (W. Li et al., 2021).

There are different methods for DM. Classification is one of the most known DM methods employed in the healthcare domain. It predicts the targets such that patients may be categorized as infected or normal or infected based on the patterns extracted from the patient's data (Helma et al., 2004). To add on, there are different algorithms and methods for data classification, some of them are classifier ensemble, k-nearest neighbor (KNN), support vector machine (SVM), and decision tree (DT).

Another notable DM method is clustering. In the clustering process, a set of data is partitioned into some smaller subsets of data, each subset is called cluster. The objects in a cluster have similar properties, yet dissimilar properties to objects in other clusters (Han et al., 2012). Clustering algorithms have been widely used in different scopes such as machine learning, image processing, information technology, artificial intelligence, pattern recognition, medical science, psychology, biology, business, and marketing (Gan et al., 2007) (V. R. Patel & Mehta, 2011). Clustering can also be categorized as an unsupervised learning technique used in the DM field. Unlike classification, there is no pre-defined category in the data used for clustering. Over the years, and through various research experiences, various clustering techniques have been developed to be used in different applications. The concept of clustering involves the division of big datasets into smaller subsets, where each subset has some similar characteristics that are measurable (Kulis & Jordan, 2012).

Data clustering, the same as other DM techniques, has been using in healthcare domain; for example, Rui Veloso (Veloso et al., 2014) used the vector quantization method as a clustering approach in predicting readmissions in intensive medicine. (Ashok Kumar, 2012) proposed a new clustering method for dichotomous healthcare data which was usable for determining the correlation of health disorders and symptoms observed in big medical and health binary databases.

One of the most widely used and the most popular clustering techniques is K-means (Mao et al., 2022). This technique was initially proposed by MacQueen in 1967 (MacQueen, 1967) and further improved by others over the decades. In a research by (X. Wu et al., 2008) they indicated that this K-means is one of the top ten most popular algorithms in DM. a research study (Kalia & Gupta, 2021) also verified that after five decades this method is still one of the popular methods which is used to cluster the large datasets.

In their paper, (Nithya et al., 2013) stated that K-means is very popular as it is simple and easily implemented. It is a simple iterative method to recover the user specified number of clusters determined by their centroids.

In spite of k-means is one the most popular clustering algorithms, but it has several issues. Although the time complexity is linear to data size, the standard k-means algorithm is not able to handle large-scale data efficiently. In some specific scenarios, the running time of k-means could be even exponential in the worst case (Ailon et al., 2009) (Vattani, 2011). Therefore, the latest researches have intended to enhance the quality or efficiency (Arthur & Vassilvitskii, 2007) (Shindler et al., 2011) (Avrithis et al., 2015) (Kanungo et al., 2002) (Elkan, 2003) (Sculley, 2010) (Bahmani et al., 2014) (J. Wang et al., 2015). K-means also was adapted in order to execute web-scale image clustering (Avrithis et al., 2015) (Gong et al., 2015).

(Rao & Rao, 2014) discussed about k-means performance, and some merits for k-means is also introduced in this research. They stated that: k-means algorithm works well for compact datasets; but it is not efficient and has poor performance for large datasets. The computational complexity is $O(n \times m \times k \times t)$, where n is number of data objects, m is the number of attributes, k is number of clusters and t is number of iterations; The number of iterations always is less than or equal to n i.e., $k \leq n$ and $t \leq n$.

Similarly, in a notable recent study (Melnykov & Michael, 2020) examined K-means performance and argued that the capability of K-means in its traditional form based on Euclidean distances is limited for analyzing high dimensional datasets.

Likewise, in their paper, (Zhao et al., 2018) explained that in the last decades the researchers have proposed several clustering algorithms. However, between these algorithms, k-means is still considered as one of the favorite choices because of its ease. During the past five decades, k-means algorithm has been widely used in various areas and industries; Moreover, the researchers have attempted continuously to optimize the k-means method in order to increase its efficiency.

In 2015 (Rajalakshmi et al., 2015) did a comparative analysis on k-means algorithm in disease prediction. According to their findings, K-means has been applied in cancer, diabetes, liver disease and heart disease predication systems. This algorithm may be faster than hierarchical clustering if k is small.

Also, (Kalyani, 2012) enhanced some known clustering algorithms that can efficiently partition medical big datasets into a number of clusters. k-means algorithm was one which was studied and enhanced in that research. The performance evaluation showed that all the optimized version of clustering models which have been proposed are able to produce the clusters with higher quality compared to the standard algorithms. However, the new k-means model which was proposed had the lower speed as compared to the standard algorithms.

(Xu et al., 2019) also highlighted some shortcomings of K-means clustering algorithm, especially in dealing with large datasets which may generate inaccurate results in the clustering process.

Likewise, there is a research in 2017 that stated the standard k-means algorithm would be quite slow for clustering millions of data into thousands of or even tens of thousands of clusters (Hu et al., 2017).

Similarly, a recent research (Xiong et al., 2020) argued that the traditional version of K-means clustering algorithm usually cannot perform the clustering process over large-scale datasets effectively as it occupies a sizeable amount of memory resources and computing costs when dealing with massive data.

In their paper, (Arora et al., 2016) listed four drawbacks of k-means algorithm. they stated that: in this algorithm finding the most suitable number of clusters (value of k) is a difficult task; using k-means with large data is not effective; since initial values of cluster centers are randomly selected, different initial cluster centers may change the result of clustering; different density and size of clusters are not handled by the algorithm. Furthermore, a notable research revealed that k-means suffers from some problems when it is applied on large amount of data (Lutz et al., 2018). (Broder et al., 2014) stated that the algorithm had a running time of $O(n \times k \times d)$ per iteration, which could become large as either the number of samples, or clusters, or the dimensionality of the dataset increased. The running time is dominated by the computation of the nearest cluster center to every sample, which is a process taking $O(k \times d)$ time per point.

Another recent research, (Mononteliza, 2020) mentioned that when the size of the dataset is very big, the performance of the algorithm will be reduced. Furthermore, the random selection of the algorithm's parameters will generate the random result.

In order to tackle the K-means limitations, different methods have been proposed. for instance, in a recent paper, (Y. Liu et al., 2021) proposed an enhanced version of the parallel process of K-means using MapReduce. MapReduce is a model which is introduced by Google in 2004. It is a simple and efficient model for managing a huge amount of data parallelly in a distributed environment. MapReduce model consists of two key components of “Map” and “Reduce” functions. Beefily, Map retrieves input data and generates key-value pairs and in Reducer the sorted key is processed and stored in the output file.

Likewise, another significant issue which the scholars intended to solve is the determination of the optimum number of clusters. various solutions have been proposed to deal with this matter. Elbow method is one of the solutions which was used in many researches such as (Sammouda & El-Zaart, 2021). Elbow method is a simple and efficient method which is applied to define the optimum number of clusters (value of K). Using the "elbow" as a cutoff point is a commonly used method in mathematical optimization which applies to select a point where diminishing returns are no longer worth the additional cost.

Another notable example of the attempt to overcome K-means limitation is using fuzzy clustering instead of normal clustering. Generally, the process of hard clustering and fuzzy clustering of K-means is the same. But in fuzzy clustering, the assignment is soft; which means, each object is assigned to all clusters with certain membership degrees varying in the unit interval which helps to increase the precision of the clustering process(Ferraro, 2021).

1.3 Problem Statement

As mentioned in the previous section, k-means is one of the most popular clustering techniques which is widely used at the present time as it is quite fast, yet, simple to understand, relatively efficient and easy to implement. Despite k-means clustering algorithm is still considered as one of the most classical data mining techniques and this algorithm has been improved over decades, but it still has some issues, especially in dealing with big data. Indeed, the traditional K-means algorithm is suitable and efficient for analyzing normal datasets. However, by rapid growth of data and the increase of the data’s volume, variety, and velocity, the current versions of K-means are not efficient enough to handle these kinds of big and complex datasets in different domains and some enhancement is required to increase the efficiency of K-means is dealing with big data.

Despite there have been a lot of improvements suggested for K-means algorithm and utilizing different techniques, K-means still suffers from some shortcomings, especially in dealing with massive and complex data. The optimum number of clusters and the initial centroid of each cluster are two parameters that have a significant impact on the quality of the clusters, since initial values of cluster centers are randomly selected, different initial cluster centers may change the result of clustering; different density and size of clusters are not handled by the algorithm.

Likewise, with regards to the time complexity of K-means, even a small increase in the number of clusters, size of data, or the number of algorithms' iterations can significantly increase the algorithm's process time and naturally, this adverse impact is more substantial in processing larger datasets. In summary, the large-scale datasets, the diversity of data types, and the high-speed stream of data the data are the main challenges of applying K-means to Big data. There are variety of parameters which have been using to measure the efficiency of clustering algorithms, but the algorithm's time complexity and the clusters' accuracy are considered as the most important parameters. Since k-means algorithm has some major weaknesses in working with large data including high computation time, this study attempts to propose an optimized extension of k-means clustering algorithm to reduce the execution time of the algorithm over the big datasets with the minimum negative impact on the clustering's accuracy. In this thesis, besides applying some techniques to optimize different steps of K-means process, the MapReduce method is used with k-means algorithm to increase the efficiency of the k-means in big data applications. The effectiveness of the proposed method in clustering big data is compared with some other clustering methods used in big data applications.

1.4 Research Questions

In this study Based on the research background and the problem statement, the main questions of the research are:

1. How to increase the efficacy of k-mean in clustering big datasets by reducing the algorithm’s process time?

2. How to provide a new version for k-mean clustering and apply it in big data clustering in healthcare domain?

3. How effective is the new technique in reducing the cluster processing time?

1.5 Research Objectives

The optimization of k-means to reduce the clustering process time in handling big data is the main contribution of this study. To describe precisely, this study attempts to propose an optimal hybrid version of k-means algorithm tested on big datasets of healthcare domain. To achieve this, the following objectives are identified:

1. To design a new version of k-means in order to reduce the process time of the algorithm in clustering big data.

2. To develop an optimal version of k-means algorithm to make this algorithm more efficient in clustering big datasets.

3. To evaluate the performance of the proposed algorithm in handling big data from healthcare domain.

1.6 Scope of the Study

The focus of this thesis is on presenting an enhanced version of k-means algorithm to be used in Big Data clustering. To do so, the current state of the art in Big Data, KDD, DM, Data Clustering, and K-means is reviewed.

Briefly, Big data is a term which is used for defining the data sets which is not possible to be stored, handled, and analyzed with traditional relational databases because of their size and complexity. Knowledge Discovery in Databases (KDD) is the procedure of uncovering meaningful and practical knowledge from a dataset. Data Mining (DM) is one of the main phases of KDD, it includes descriptive and predictive tasks and of the main descriptive tasks is clustering. K-means algorithm is one of the partitioning-based clustering techniques which has been widely used in DM for many years.

However, details of the datasets and the internal technology and criteria of providing these datasets are out of the scope of this study. The scope of this study is illustrated in Figure 1-1:

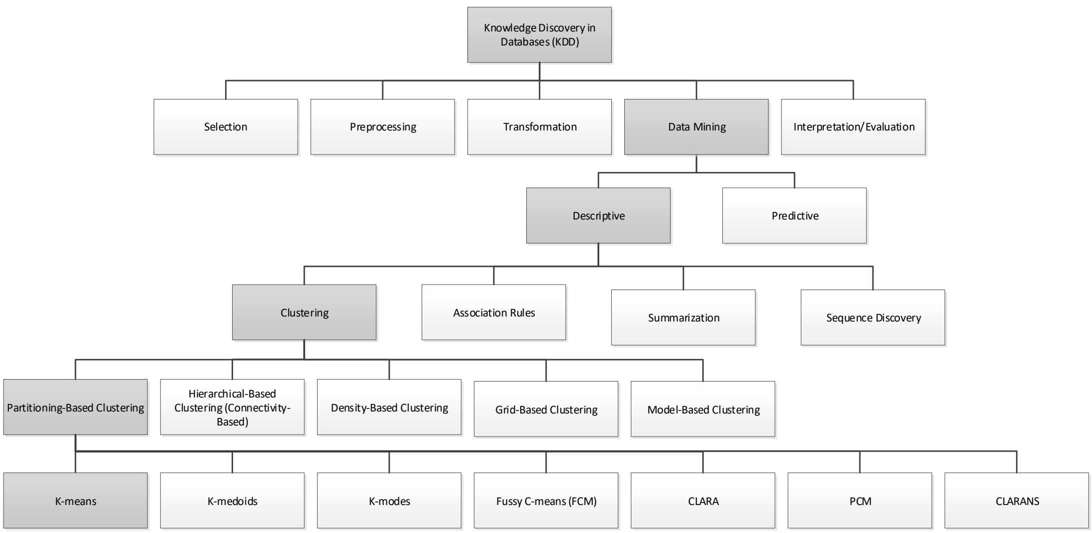

Figure 1-1 Study Scope

1.7 Significance of the Study

Due to the rapid and uninterrupted increasing volume of data in various domains, the data complexity also increases as well. By using the traditional DM methods and tools, extracting useful information from massive data is complicated. Hence, to analyze and obtain significant and practical knowledge from this complex data, powerful computing tools and techniques are required. In such a situation that huge amount of data is available, using DM to extract useful information is beneficial (Ahmad et al., 2015). For instance, healthcare data mostly includes all the information related to the patients as well as the parties involved in the treatment process. With improvements of tmedical devices and healthcare systems, the rate of storing large amounts of such data is increasing rapidly. In healthcare domain, DM techniques and tools assist to diagnose the unknown diseases, causes of diseases, and identification of medical treatment methods. It also helps medical scientists and researchers to establish efficient healthcare policies, constructing drug recommendation systems, and developing health profiles of individuals (Haraty et al., 2015).

Various DM methods such as association, clustering and classification has been utilizing by data analysts. Clustering, which is one of the methods of DM, is an unsupervised learning method. In clustering, large data sets are split into some small sub-groups and each sub-group will be analyzed individually (Kulis & Jordan, 2012). K-means is one of the most popular clustering algorithms which has been widely used for over half a century to divide the data into small sub-groups in order to ease the analysis of data. There are many researches and studies about the application of K-means algorithm in healthcare domain. For example (Narmadha et al., 2016) did a survey on clustering methods which is used in healthcare area in order to group the patients' records.

In order to reach the most suitable result, the data should be stored, integrated, cleaned, stored, analyzed, and interpreted with the more capable methods. The complexity, timeliness, noise, heterogeneity, and incompleteness of big data obstruct the process of extracting valuable knowledge from them. It is not possible for traditional clustering techniques to handle this massive amount of data smoothly due to their computational costs and high complexity (Shirkhorshidi & Aghabozorgi, 2014). Because of the importance of healthcare in societies and the abundant application of different DM techniques, such as clustering in healthcare, researchers are extremely influenced by the abilities of these techniques and the improvement of their performance in healthcare industry.

There is still a gap between the potential and usability of k-means in DM applications in healthcare domain in practice (Lee, Luo, Ngiam, Zhang, Zheng, Chen, Ooi, & Yip, 2017). The major purpose of this study is to speed up k-means clustering algorithm with minimum negative impact on the clustering accuracy. The clustering algorithms' scalability and speed were always an important aims for scholars in this field, however, big data issues and challenges highlight these weaknesses and require further consideration and study in this area.

1.8 Organization of the Thesis

This thesis has been structured into seven chapters. The first chapter introduces the topic of the research. In this chapter the background of the study, the existing problem, the research questions and objective, and the scope and significance and the study are being discussed.

Chapter two includes review of the literature on the topic, this chapter covers the historical and technical information and the current state of research on big data, data mining, especially with consideration of data mining in healthcare domain, data clustering, and big data clustering, and K-means algorithm.

The third chapter presents the general methodology which is used in this study to propose and evaluate the optimized version of k-means. The research design includes the operational framework. Moreover, all aspects of the required processes for study, design & develop, and verification & validation steps is explained in this chapter.

Chapter four discusses about the idea of design the optimized method, also different stages of designing the new algorithm are described in detail.

In the fifth chapter, the process of developing the new algorithm is presented.

Chapter six represents the process of evaluation of the presented algorithm in accordance with the evaluation protocol of the study.

And finally, chapter seven provides a conclusion of the study, issues, and challenges, also discusses future works as well.

REFERENCES

Ahmad, P., Qamar, S., & Afser Rizvi, S. Q. (2015). Techniques of Data Mining In Healthcare: A Review. *International Journal of Computer Applications*, *120*(15), 38–50. https://doi.org/10.1016/j.procs.2015.12.145

Ailon, N., Jaiswal, R., & Monteleoni, C. (2009). Streaming k-means approximation. *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*.

Akhanli, S. E., & Hennig, C. (2020). Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. *Statistics and Computing*, *30*(5), 1523–1544. https://doi.org/10.1007/s11222-020-09958-2

Akthar, N., Ahamad, M. V., & Ahmad, S. (2016). MapReduce model of improved K-means clustering algorithm using hadoop MapReduce. *Proceedings - 2016 2nd International Conference on Computational Intelligence and Communication Technology, CICT 2016*, 192–198. https://doi.org/10.1109/CICT.2016.46

Alia, O. M. d., Mandava, R., & Aziz, M. E. (2011). A hybrid harmony search algorithm for MRI brain segmentation. *Evolutionary Intelligence*, *4*(1), 31–49. https://doi.org/10.1007/s12065-011-0048-1

Alsayat, A. M. (2016). *Efficient Genetic K-Means Clustering Algorithm And Its Application To Data Mining On Different Domains*. Bowie State University.

Althaf, R. S. K., Sai, R. K., & Girija, R. K. (2018). Challenging tools on Research Issues in Big Data Analytics. *International Journal of Engineering Development and Research*, *6*(1), 637–644.

Apache Hadoop. (2019). *Apache Hadoop*. https://hadoop.apache.org/

Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm for Big Data. *Physics Procedia*, *78*, 507–512. https://doi.org/10.1016/j.procs.2016.02.095

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035.

Ashkartizabi, M., & Aminghafari, M. (2018). Functional data clustering using K-means and random projection with applications to climatological data. *Stochastic Environmental Research and Risk Assessment*, *32*(1), 83–104. https://doi.org/10.1007/s00477-017-1441-9

Ashok Kumar, D. (2012). Clustering Dichotomous Data for Health Care. *International Journal of Information Sciences and Techniques*, *2*(2), 235–33. https://doi.org/10.5121/ijist.2012.2203

Aubaidan, B., Mohd, M., Albared, M., & Author, F. (2014). Comparative study of k-means and k-means++ clustering algorithms on crime domain. *Journal of*

Computer Science, *10*(7), 1197–1206. https://doi.org/10.3844/jcssp.2014.1197.1206

Avrithis, Y., Kalantidis, Y., Anagnostopoulos, E., & Emiris, I. Z. (2015). Web-scale image clustering revisited. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 1502–1510. https://doi.org/10.1109/ICCV.2015.176

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining*, *January*, 182–185. https://doi.org/ISBN: 978-972-8924-63-8

B., S., J.P., D., C., G., J.L., O., S., V., K.J., C., & J.N., C. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/781670

Bahmani, B., Kumar, R., & Vassilvitskii, S. (2014). Scalable K-Means ++. *Proceedings of the VLDB Endowment*, 622–633. https://doi.org/10.14778/2180912.2180915

Balouchestani, M., & Krishnan, S. (2016). Advanced K-means clustering algorithm for large ECG data sets based on a collaboration of compressed sensing theory and K-SVD approach. *Signal, Image and Video Processing*, *10*(1), 113–120. https://doi.org/10.1007/s11760-014-0709-5

Béjar, J. (2013). *Strategies and Algorithms for Clustering Large Datasets: A Review*. 1–20. https://upcommons.upc.edu/bitstream/handle/2117/23415/R13-11.pdf

Belciug, S. (2009). Patients length of stay grouping using the hierarchical clustering algorithm. *Annals Math. Comp. Sci. Ser*, *36*(2), 79–84. http://inf.ucv.ro/~ami/index.php/ami/article/viewFile/288/279

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, *77*(2), 81–97. https://doi.org/10.1016/j.ijmedinf.2006.11.006

Benmammar, B., Taleb, M. H., & Krief, F. (2017). Diffusing-CRN k-means: an improved k-means clustering algorithm applied in cognitive radio ad hoc networks. *Wireless Networks*, *23*(6), 1849–1861. https://doi.org/10.1007/s11276-016-1257-4

Beyer, M., & Laney, D. (2012). *The Importance of "Big Data": A Definition*. Gartner, Analysis Report G00235055,. https://www.gartner.com/doc/2057415/importance-big-data-definition

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2–3), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Bhatia, N., Sojan, J. M., Simonovic, S., & Srivastav, R. (2020). Role of cluster validity indices in delineation of precipitation regions. *Water (Switzerland)*, *12*(5), 1–28. https://doi.org/10.3390/W12051372

Bifet, A. (2009). Adaptive learning and mining for data streams and frequent patterns. *ACM SIGKDD Explorations Newsletter*, *11*(1), 55. https://doi.org/10.1145/1656274.1656287

Borne, K. (2014). *Top 10 List – The V's of Big Data - Data*. Data Scien Cecentral, The Online Resource for Big Data Practitioners. https://www.datasciencecentral.com/profiles/blogs/top-10-list-the-v-s-of-big-data

Broder, A., Garcia-Pueyo, L., Josifovski, V., Vassilvitskii, S., & Venkatesan, S. (2014). Scalable K-Means by ranked retrieval. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining - WSDM '14*, 233–242. https://doi.org/10.1145/2556195.2556260

Bu, Y., Howe, B., Balazinska, M., & Ernst, M. D. (2010). HaLoop: : efficient iterative data processing on large clusters. *36th International Conference on Very Large Data Bases*, *3*(1–2), 285–296. https://doi.org/10.14778/1920841.1920881

Cackett, D. (2013). Information Management and Big Data A Reference Architecture. *Oricale White Paper*, *February*, 28. https://doi.org/10.1.1.398.7632

Chaudhary, A., & Bhattacharjee, V. (2020). An efficient method for brain tumor detection and categorization using MRI images by K-means clustering & DWT. *International Journal of Information Technology*, *12*(1), 141–148. https://doi.org/10.1007/s41870-018-0255-4

Chen, Y., Hu, P., & Wang, W. (2019). Improved K-Means Algorithm and its Implementation Based on Mean Shift. *Proceedings - 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2018*, 1–5. https://doi.org/10.1109/CISP-BMEI.2018.8633100

Cinaroglu, S. (2020). Integrated k-means clustering with data envelopment analysis of public hospital efficiency. *Health Care Management Science*, *23*(3), 325–338. https://doi.org/10.1007/s10729-019-09491-3

Cortada, J. W., Gordan, D., & Lenihan, B. (2012). The value of analytics in healthcare. *IBM Institute for Business Value*.

Crowley, J., & Pauler Ankerst, D. (Eds.). (2005). *Handbook of Statistics in Clinical Oncology, Second Edition*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420027761

Cui, M. (2020). Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. *Accounting, Auditing and Finance*, *2020*(1), 5–8. https://doi.org/10.23977/accaf.2020.010102

Dalatu, P. I. (2016). Time Complexity of K-Means and K-Medians Clustering Algorithms in Outliers Detection. *Global Journal of Pure and Applied Mathematics.*, *12*(5), 4405–4418.

Das, A. K., Kedia, A., Sinha, L., Goswami, S., Chakrabarti, T., & Chakrabarti, A.

(2016). Data mining techniques in Indian healthcare: A short review. *Proceedings - 2015 International Conference on Man and Machine Interfacing, MAMI 2015*. https://doi.org/10.1109/MAMI.2015.7456611

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Dean, J., & Ghemawat, S. (2008). MapReduce, Simplified Data Processing on Large Clusters. *Communications of the ACM*, *51*(1), 107. https://doi.org/10.1145/1327452.1327492

Debelee, T. G., Schwenker, F., Rahimeto, S., & Yohannes, D. (2019). Evaluation of modified adaptive k-means segmentation algorithm. *Computational Visual Media*, *5*(4), 347–361. https://doi.org/10.1007/s41095-019-0151-2

Dhanasekaran, S., Sundarrajan, R., Murugan, B. S., Kalaivani, S., & Vasudevan, V. (2019). Enhanced Map Reduce Techniques for Big Data Analytics based on K-Means Clustering. *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2019*, 0–4. https://doi.org/10.1109/INCOS45849.2019.8951368

Dunn, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, *4*(1), 95–104. https://doi.org/10.1080/01969727408546059

El-Mandouh, A. M., Mahmoud, H. A., Abd-Elmegid, L. A., & Haggag, M. H. (2019). Optimized K-means clustering model based on gap statistic. *International Journal of Advanced Computer Science and Applications*, *10*(1), 183–188. https://doi.org/10.14569/IJACSA.2019.0100124

Elkan, C. (2003). Using the Triangle Inequality to Accelerate k-Means. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, 147–153. https://doi.org/10.1016/0026-2714(92)90278-S

Eren, B., Karabulut, E. C., Alptekin, S. E., & Alptekin, G. I. (2015). A K-Means Algorithm Application on Big Data. *World Congress on Engineering and Computer Science, Wcecs 2015, Vol II*, *II*, 814–818.

Esteves, R. M., Hacker, T., & Rong, C. (2013). Competitive K-means: A new accurate and distributed K-means algorithm for large datasets. *Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom*, *1*, 17–24. https://doi.org/10.1109/CloudCom.2013.89

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 267–279. https://doi.org/10.1109/TETC.2014.2330519

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 37–54. https://doi.org/10.1145/240455.240463

Ferraro, M. B. (2021). Fuzzy k-Means: history and applications. *Econometrics and Statistics*, *xxxx*, 1–14. https://doi.org/10.1016/j.ecosta.2021.11.008

Finances online. (2019). *20 Best Data Analytics Software for 2019 - Financesonline.com*. https://financesonline.com/data-analytics/

Firican, G. (2017). *The 10 Vs of Big Data*. TDWI Business Intelligence and Data Warehousing Education and Research. https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Fraley, C. (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, *41*(8), 578–588. https://doi.org/10.1093/comjnl/41.8.578

Gan, G., Ma, C., & Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. *January 2007*. https://doi.org/10.1137/1.9780898718348

Gentle, J. E., Kaufman, L., & Rousseuw, P. J. (1991). Finding Groups in Data: An Introduction to Cluster Analysis. *Biometrics*, *47*(2), 788. https://doi.org/10.2307/2532178

Ghadiri, N., Ghaffari, M., & Nikbakht, M. A. (2017). BigFCM: Faast, precise and scalable FCM on hadoop. *Future Generation Computer Systems*, *77*, 29–39. https://doi.org/10.1016/j.future.2017.06.010

Golberg, D. E. (1989). *Genetic Algorithms in Search Optimization & Machine Learning* (p. 412). https://doi.org/10.1007/3-540-44673-7

Gong, Y., Pawlowski, M., Yang, F., Brandy, L., Boundev, L., & Fergus, R. (2015). Web Scale Photo Hash Clustering on A Single Machine. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 19–27.

Guerra-hern, A., & Mondrag, R. (2008). *Explorations of the BDI Multi-agent support for the Knowledge Discovery in Databases Process Explorations of the BDI Multi-Agent support for the Knowledge Discovery in Databases Process*. *January*.

Gupta, R., Gupta, S., & Singhal, A. (2014). Big Data: Overview. *International Journal of Computer Trends and Technology*, *9*(5), 266–268. https://doi.org/10.14445/22312803/IJCTT-V9150

Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis* (7th ed.). https://www.pearson.com/us/higher-education/program/Hair-Multivariate-Data-Analysis-7th-Edition/PGM263675.html

Hamaainen, J., Karkkainen, T., & Rossi, T. (2020). Improving scalable k-means++. *Algorithms*, *14*(1), 1–20. https://doi.org/10.3390/a14010006

Hamalainen, J., Jauhiainen, S., & Karkkainen, T. (2017). Comparison of internal clustering validation indices for prototype-based clustering. *Algorithms*, *10*(3). https://doi.org/10.3390/a10030105

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In *San Francisco, CA, itd: Morgan Kaufmann*. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Hand, D., Hand, D., Mannila, H., Mannila, H., Smyth, P., & Smyth, P. (2001). Principles of data mining. In *Drug safety : an international journal of medical toxicology and drug experience* (Vol. 30). https://doi.org/10.2165/00002018-200730070-00010

Haoxiang, W., & Smys, S. (2021). Big Data Analysis and Perturbation using Data Mining Algorithm. *Journal of Soft Computing Paradigm*, *3*(1), 19–28. https://doi.org/10.36548/jscp.2021.1.003

Haraty, R. A., Dimishkieh, M., & Masud, M. (2015). An enhanced k-means clustering algorithm for pattern discovery in healthcare data. *International Journal of Distributed Sensor Networks, 2015*. https://doi.org/10.1155/2015/615740

Harper, P. R. (2005). A review and comparison of classification algorithms for medical decision making. *Health Policy, 71*(3), 315–331. https://doi.org/10.1016/j.healthpol.2004.05.002

Helma, C., Cramer, T., Krame, S., & Luc, D. R. (2004). Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds Noncongeneric Compounds. *Journal of Chemical Information and Computer Sciences*. https://doi.org/10.1021/CI034254Q

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big Data, 1*(1). https://doi.org/10.1186/2196-1115-1-2

Hinneburg, A., & Keim, D. A. (1998). An Efficient Approach to Clustering in Large Multimedia Databases with Noise. *Proceedings of the 4th ACM SIGKDD Knowledge Discovery and Data Mining AAAI Press, Menlo Park, 5865*(c), 58–65. https://doi.org/10.1.1.44.3961

Hinneburg, A., & Keim, D. a. (1999). Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering. *International Conference on Very Large Databases (VLDB)*, 506–517. https://doi.org/1-55860-615-7

Hosseini, B., & Kiani, K. (2018). FWCMR: A scalable and robust fuzzy weighted clustering based on MapReduce with application to microarray gene expression. *Expert Systems with Applications*, *91*, 198–210. https://doi.org/10.1016/j.eswa.2017.08.051

Hosseinkhah, F., Ashktorab, H., Veen, R., & Owrang O, M. M. (2009). Challenges in Data Mining on Medical Databases. *Database Technologies: Concepts, Methodologies, Tools, and Applications*, 502–511. https://doi.org/10.4018/978-1-60566-058-5.ch083

Hou, X. (2020). An Improved K-means Clustering Algorithm Based on Hadoop Platform. In *Advances in Intelligent Systems and Computing* (Vol. 928, pp. 1101–1109). https://doi.org/10.1007/978-3-030-15235-2_146

Hu, Q., Wu, J., Bai, L., Zhang, Y., & Cheng, J. (2017). Fast K-means for Large Scale

Clustering. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*, 2099–2102. https://doi.org/10.1145/3132847.3133091

Huang, S. Y., & Zhang, B. (2019). Research on improved k-means clustering algorithm based on hadoop platform. *Proceedings - 2019 International Conference on Machine Learning, Big Data and Business Intelligence, MLBDBI 2019*, 301–303. https://doi.org/10.1109/MLBDBI48998.2019.00067

Huang, Z. (1997). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. *Research Issues on Data Mining and Knowledge Discovery*, 1–8. https://doi.org/10.1.1.6.4718

Humaira, H., & Rasyidah, R. (2020). Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm. *An Integrated Workflow for MICP-Based Rock Typing: A Case Study of a Tight-Gas Sandstone Reservoir in the Baltic Basin (Poland).*, *January*. https://doi.org/10.4108/eai.24-1-2018.2292388

IBM. (2017). *10 Key Marketing Trends for 2017*. https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN

IBM. (2018a). *Big Data Analytics*. https://www.ibm.com/analytics/hadoop/big-data-analytics

IBM. (2018b). *The Four V's of Big Data*. IBM Big Data & Analytics Hub. https://www.ibmbigdatahub.com/infographic/four-vs-big-data

Islam, M., Hasan, M., Wang, X., Germack, H., & Noor-E-Alam, M. (2018). A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining. *Healthcare*, *6*(2), 54. https://doi.org/10.3390/healthcare6020054

J, S. S., & Pandya, S. (2016). An Overview of Partitioning Algorithms in Clustering Techniques. *International Journal of Advanced Research in Computer Engineering & Technology*, *5*(6), 2278–1323. http://ijarcet.org/wp-content/uploads/IJARCET-VOL-5-ISSUE-6-1943-1946.pdf

Jain, M., & Verma, C. (2014). Adapting k-means for Clustering in Big Data. *International Journal of Computer Applications*, *101*(1), 19–24.

Jin, S., Cui, Y., & Yu, C. (2016). A New Parallelization Method for K-means. *ArXiv Distribution Service and Open-Access Archive for Scholarly Articles, Cornell University*. http://arxiv.org/abs/1608.06347

Jony, R. I., Rony, R. I., Rahat, A., & Rahman, M. (2016). Big Data Characteristics , Value Chain and Challenges. *1st International Conference on Advanced Information and Communication Technology 2016, At Chittagong Independent University, Bangladesh.*, *May*, 1–6.

Jothi, N., Rashid, N. A., & Husain, W. (2015). Data Mining in Healthcare - A Review. *Procedia Computer Science*, *72*, 306–313. https://doi.org/10.1016/j.procs.2015.12.145

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. *2013 46th Hawaii International Conference on System Sciences*, 995–1004. https://doi.org/10.1109/HICSS.2013.645

Kalia, K., & Gupta, N. (2021). Analysis of hadoop MapReduce scheduling in heterogeneous environment. *Ain Shams Engineering Journal*, *12*(1), 1101–1110. https://doi.org/10.1016/j.asej.2020.06.009

Kalyani, P. (2012). Approaches to Partition Medical Data using Clustering Algorithms. *International Journal of Computer Applications*, *49*(23), 975–8887. https://doi.org/10.5120/7941-1102

Kant, S., & Ansari, I. A. (2016). An improved K means clustering with Atkinson index to classify liver patient dataset. *International Journal of Systems Assurance Engineering and Management*, *7*(1), 222–228. https://doi.org/10.1007/s13198-015-0365-3

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 881–892. https://doi.org/10.1109/TPAMI.2002.1017616

Kapil, G., Agrawal, A., & Khan, R. A. (2016). A study of big data characteristics. *2016 International Conference on Communication and Electronics Systems (ICCES)*, 1–4. https://doi.org/10.1109/CESYS.2016.7889917

Khanali, H., & Vaziri, B. (2020). An improved approach to fuzzy clustering based on FCM algorithm and extended VIKOR method. *Neural Computing and Applications*, *32*(2), 473–484. https://doi.org/10.1007/s00521-019-04035-w

Khanmohammadi, S., Adibeig, N., & Shanehbandy, S. (2017). An improved overlapping k-means clustering method for medical applications. *Expert Systems with Applications*, *67*, 12–18. https://doi.org/10.1016/j.eswa.2016.09.025

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated. *Science*, *220*(4598), 671–680.

Kitchenham, B. A., & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. *Citeseer*. https://doi.org/10.1145/1134285.1134500

Koh, H. C., & Tan, G. (2005). Data mining applications in healthcare. *J Healthc Inf Manag*, *19*(2), 64–72. https://doi.org/10.4314/ijonas.v5i1.49926

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management : JHIM*, *19*(2), 64–72. http://www.ncbi.nlm.nih.gov/pubmed/15869215

Kolen, J. F., & Hutcheson, T. (2002). Reducing the Time Complexity of the Fuzzy. *Brain*, *10*(2), 263–267.

Kulis, B., & Jordan, M. I. (2012). Revisiting k-means: New Algorithms via Bayesian Nonparametrics. *In Proceedings of the 29th International Conference*

OnMachine Learning, 513–520. https://doi.org/10.1002/int.20069

Kumar, V. (2017). *Big Data Facts | AnalyticsWeek*. https://analyticsweek.com/content/big-data-facts/

Kurinjivendhan, N., & Thangadurai, K. (2016). Modified k-means algorithm and genetic approach for cluster optimization. *Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016*, 53–56. https://doi.org/10.1109/SAPIENCE.2016.7684130

L. Kaufman, & P. J. Rousseuw. (1991). Finding Groups in Data : An Introduction to Cluster Analysis. *Society, International Biometric*, *47*(2), 788.

Laccetti, G., Lapegna, M., Mele, V., Romano, D., & Szustak, L. (2020). Performance enhancement of a dynamic K-means algorithm through a parallel adaptive strategy on multicore CPUs. *Journal of Parallel and Distributed Computing*, *145*, 34–41. https://doi.org/10.1016/j.jpdc.2020.06.010

Lasheng, C., & Yuqiang, L. (2017). Improved initial clustering center selection algorithm for K -means. *Algorithms, Architectures, Arrangements, and Applications*, 275–279.

Le, T. L., Huynh, T. T., Lin, L. Y., Lin, C. M., & Chao, F. (2019). A K-means Interval Type-2 Fuzzy Neural Network for Medical Diagnosis. *International Journal of Fuzzy Systems*, *21*(7), 2258–2269. https://doi.org/10.1007/s40815-019-00730-x

Lee, C., Luo, Z., Ngiam, K. Y., Zhang, M., Zheng, K., Chen, G., Ooi, B. C., & Yip, W. L. J. (2017). Big Healthcare Data Analytics: Challenges and Applications. *Large-Scale Distributed Computing in Smart Healthcare*, 11–41. https://doi.org/10.1007/978-3-319-58280-1_2

Lee, C., Luo, Z., Ngiam, K. Y., Zhang, M., Zheng, K., Chen, G., Ooi, C., Luen, W., & Yip, J. (2017). Big Healthcare Data Analytics: Challenges and Applications. In *Handbook ofLarge-Scale Distributed Computing in Smart Healthcare, Scalable Computing and Communications, DOI*. Springer International Publishing AG 2017. https://doi.org/10.1007/978-3-319-58280-1

Li, W., Chai, Y., Khan, F., Jan, S. R. U., Verma, S., Menon, V. G., Kavita, & Li, X. (2021). A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System. *Mobile Networks and Applications*, *26*(1), 234–252. https://doi.org/10.1007/s11036-020-01700-6

Li, Y., Yang, G., He, H., Jiao, L., & Shang, R. (2016). A study of large-scale data clustering based on fuzzy clustering. *Soft Computing*, *20*(8), 3231–3242. https://doi.org/10.1007/s00500-015-1698-1

Liao, Q., Yang, F., & Zhao, J. (2013). An improved parallel K-means clustering algorithm with MapReduce. *International Conference on Communication Technology Proceedings, ICCT*, 764–768. https://doi.org/10.1109/ICCT.2013.6820477

Liu, Y., Du, X., & Ma, S. (2021). Innovative study on clustering center and distance measurement of K-means algorithm: mapreduce efficient parallel algorithm based on user data of JD mall. In *Electronic Commerce Research* (Issue 0123456789). Springer US. https://doi.org/10.1007/s10660-021-09458-z

Liu, Z., Bao, J., & Ding, F. (2018). An improved k-means clustering algorithm based on semantic model. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3148453.3306269

Lu, F., & Boritz, J. E. (2005). Detecting Fraud in Health Insurance Data: Learning to Model Incomplete Benford's Law Distributions. *European Conference on Machine Learning*, 633–640. https://doi.org/10.1007/11564096_63

Lu, W. (2020). Improved K-Means Clustering Algorithm for Big Data Mining under Hadoop Parallel Framework. *Journal of Grid Computing*, *18*(2), 239–250. https://doi.org/10.1007/s10723-019-09503-0

Luna-Romera, J. M., Gutierrez, J. G., Martinez-Ballesteros, M., & Riquelme Santos, J. C. (2018). An approach to validity indices for clustering techniques in Big Data. *Progress in Artificial Intelligence*, *7*(2), 81–94. https://doi.org/10.1007/s13748-017-0135-3

Lutz, C., Breb, S., Rabl, T., Zeuch, S., & Mark, V. (2018). Efficient and Scalable k-Means on GPUs. *Datenbank Spektrum*, 157–169.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*(233), 281–297. https://doi.org/citeulike-article-id:6083430

Madhulatha, T. S. (2012). an Overview on Clustering Methods. *IOSR Journal of Engineering*, *02*(04), 719–725. https://doi.org/10.9790/3021-0204719725

Maimon, O., & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook. In *Springer* (Second Edi). Springer. https://doi.org/10.2333/jbhmk.26.46

Manochandar, S., Punniyamoorthy, M., & Jeyachitra, R. K. (2020). Development of new seed with modified validity measures for k-means clustering. *Computers and Industrial Engineering*, *141*(July 2018), 106290. https://doi.org/10.1016/j.cie.2020.106290

Mao, Y. M., Gan, D. J., Mwakapesa, D. S., Nanehkaran, Y. A., Tao, T., & Huang, X. Y. (2022). A MapReduce-based K-means clustering algorithm. *Journal of Supercomputing*, *78*(4), 5181–5202. https://doi.org/10.1007/s11227-021-04078-8

Marutho, D., Hendra Handaka, S., Wijaya, E., & Muljono. (2018). The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. *Proceedings - 2018 International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life, ISemantic 2018*, 533–538. https://doi.org/10.1109/ISEMANTIC.2018.8549751

Mayer-Schonberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

https://books.google.com.my/books/about/Big_Data.html?id=HpHcGAkFEjkC&printsec=frontcover&source=kp_read_button&redir_esc=y#v=onepage&q&f=false

McGregor, C., Catley, C., & James, A. (2012). A process mining driven framework for clinical guideline improvement in critical care. *CEUR Workshop Proceedings*, *765*.

McGuire, T. (2013). *Making data analytics work: Three key challenges*. McKinsey. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/making-data-analytics-work

Mehar, A. M., Maeder, A., Matawie, K., Ginige, A., Mehar, A. M., Maeder, A., Matawie, K., Ginige, A., & Clustering, B. (2014). Blended Clustering for Health Data Mining. *IFIP Advances in Information and Communication Technology*.

Melnykov, V., & Michael, S. (2020). Clustering Large Datasets by Merging K-Means Solutions. *Journal of Classification*, *37*(1), 97–123. https://doi.org/10.1007/s00357-019-09314-8

Min, Z., & Kai-Fei, D. (2015). Improved Research to K-means Initial Cluster Centers. *Proceedings - 2015 9th International Conference on Frontier of Computer Science and Technology, FCST 2015*, 349–353. https://doi.org/10.1109/FCST.2015.61

Mohebi, A., Aghabozorgi, S., Wah, T. Y., Herawan, T., & Yahyapour, R. (2015). Iterative big data clustering algorithms: a review. *Software - Practice and Experience*, *39*(7), 701–736. https://doi.org/10.1002/spe.2341

Mononteliza, J. (2020). Improved K-means Clustering Algorithm based on Dynamic Clustering. *Asia-Pacific Journal of Convergent Research Interchange*, *6*(4), 1–12. https://doi.org/10.21742/apjcri.2020.04.01

Moorman, C. (2013). *The Utilization Gap: Big Data’s Biggest Challenge*. Forbes. https://www.forbes.com/sites/christinemoorman/2013/03/17/the-utilization-gap-big-datas-biggest-challenge/#60aecea33563

Munandar, T., & Musdholifah, A. (2014). Comparative Study Between Primitive Operation Complexity Against Running Time Application On Clustering Algorithm. *International Journal of Advanced Research in Computer Science*, *5*(5).

Murali K, P., Salehi Amini, M., Jayasimha R., K., Xie, Y., & Raghavan, V. (2016). Massive Data Analysis: Tasks, Tools, Applications, and Challenges. *Big Data Analytics: Methods and Applications*, 1–276. https://doi.org/10.1007/978-81-322-3628-3

Murtagh, F., & Fionn. (1992). Comments on "Parallel algorithms for hierarchical clustering and cluster validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *14*(10), 1056–1057. https://doi.org/10.1109/34.159908

Murthy, A. (2013). *Tez: Accelerating processing of data stored in HDFS*.

https://hortonworks.com/blog/introducing-tez-faster-hadoop-processing/

Murthy, C. A., & Chowdhury, N. (1996). In search of optimal clusters using genetic algorithms. *Pattern Recognition Letters*, *17*(8), 825–832. https://doi.org/10.1016/0167-8655(96)00043-8

Naik, D. S. B., Kumar, S. D., & Ramakrishna, S. V. (2013). *Parallel Processing Of Enhanced K-Means Using OpenMP*. 1–4.

Nair, A. (2019). *Beginner's Guide To K-Means Clustering*. Developers Corner. https://analyticsindiamag.com/beginners-guide-to-k-means-clustering/

Narmadha, D., Balamurugan, A. alias, Sundar, G. N., & Priya, S. J. (2016). Survey of clustering algorithms for categorization of patient records in healthcare. *Indian Journal of Science and Technology*, *9*(8). https://doi.org/10.17485/ijst/2016/v9i8/87971

Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, *14*(5), 1003–1016. https://doi.org/10.1109/TKDE.2002.1033770

Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015). Classification of healthcare data using genetic fuzzy logic system and wavelets. *Expert Systems with Applications*, *42*(4), 2184–2197. https://doi.org/10.1016/j.eswa.2014.10.027

Nist, D., & Stew, H. (2006). Scalable Recognition with a Vocabulary Tree. *CVPR '06 Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2*, 2161–2168.

Nithya, N. S., Duraiswamy, K., & Gomathy, P. (2013). A Survey on Clustering Techniques in Medical Diagnosis. *International Journal of Computer Science Trends and Technology*, *1*(2), 17–22. www.ijcstjournal.org

Normandeau, K. (2013). *Big data volume, variety, velocity and veracity*. Inside Big Data. https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/

Oliveira, G. V., Coutinho, F. P., Campello, R. J. G. B., & Naldi, M. C. (2017). Improving k-means through distributed scalable metaheuristics. *Neurocomputing*, *246*, 45–57. https://doi.org/10.1016/j.neucom.2016.07.074

Oyelade, J., Isewon, I., Oladipupo, O., Emebo, O., Omogbadegun, Z., Aromolaran, O., Uwoghiren, E., Olaniyan, D., & Olawole, O. (2019). Data Clustering: Algorithms and Its Applications. *Proceedings - 2019 19th International Conference on Computational Science and Its Applications, ICCSA 2019*, *ii*, 71–81. https://doi.org/10.1109/ICCSA.2019.000-1

P., D., & Ahmed, K. (2016). A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools. *International Journal of Advanced Computer Science and Applications*, *7*(2). https://doi.org/10.14569/IJACSA.2016.070267

Patel, S., & Patel, H. (2016). Survey of Data Mining Techniques used in Healthcare Domain. *International Journal of Information Sciences and Techniques*, *6*(1/2), 53–60. https://doi.org/10.5121/ijist.2016.6206

Patel, V. R., & Mehta, R. G. (2011). Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm. *IJCSI International Journal of Computer Science Issues ISSN*, *8*(2), 1694–1814. www.IJCSI.org

Pfitzner, D., Leibbrandt, R., & Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. In *Knowledge and Information Systems* (Vol. 19, Issue 3). https://doi.org/10.1007/s10115-008-0150-6

Piri, S. (2017). *Developing and Deploying Data Mining Techniques In Healthcare*. (Doctoral dissertation, Oklahoma State University).

Press, G. (2016). *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says*. Forbes. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#68ef600b6f63

Prodel, M., Prodel, M., & Augusto, M. V. (2017). *Process discovery , analysis and simulation of clinical pathways using health-care data*. Université de Lyon.

Purandhar, N., Ayyasamy, S., & Saravanakumar, N. M. (2021). Clustering healthcare big data using advanced and enhanced fuzzy C-means algorithm. *International Journal of Communication Systems*, *34*(1), 1–12. https://doi.org/10.1002/dac.4629

Qi, J., Yu, Y., Wang, L., & Liu, J. (2016). K∗-means: An effective and efficient k-means clustering algorithm. *Proceedings - 2016 IEEE International Conferences on Big Data and Cloud Computing, BDCloud 2016, Social Computing and Networking, SocialCom 2016 and Sustainable Computing and Communications, SustainCom 2016*, 242–249. https://doi.org/10.1109/BDCloud-SocialCom-SustainCom.2016.46

Raghupathi, K. (2018). *10 Interesting Use Cases for the K-Means Algorithm*. DZone AI. https://dzone.com/articles/10-interesting-use-cases-for-the-k-means-algorithm

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 3. https://doi.org/10.1186/2047-2501-2-3

Rajalakshmi, K., Dhenakaran, S. S., & Roobini, N. (2015). Comparative Analysis of K-Means Algorithm in Disease Prediction. *International Journal of Science, Engineering and Technology Research*, *4*(7), 2697–2699.

Rammal, A., Perrin, E., Vrabie, V., Bertrand, I., & Chabbert, B. (2017). Classification of lignocellulosic biomass by weighted-covariance factor fuzzy C-means clustering of mid-infrared and near-infrared spectra. *Journal of Chemometrics*, *31*(2), 1–10. https://doi.org/10.1002/cem.2865

Rao, P. V., & Rao, S. K. M. (2014). Performance Issues on K-Mean Partitioning Clustering Algorithm. *International Journal of Computer (IJC)*, *14*(1), 41–51.

Rendon, E., Abundez, I., Arizmendi, A., & Quiroz, E. (2011). Internal versus external cluster validation indexes. *International Journal of Computers and Communications*, *5*(1), 27–34.

Riccomini, C. (2013). *Apache Samza: Real-time Stream Processing at LinkedIn*. https://www.infoq.com/presentations/samza-linkedin

Rose, K., Gurewitz, E., & Fox, G. (1990). A deterministic annealing approach to clustering. *Pattern Recognition Letters*, *11*(9), 589–594. https://doi.org/10.1016/0167-8655(90)90010-Y

Rosencrance, L., Vaughan, J., & Preslar, E. (2021). *Hadoop Distributed File System (HDFS) Definition*. TechTarget. https://searchdatamanagement.techtarget.com/definition/Hadoop-Distributed-File-System-HDFS

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1177/003754977702900403

Rukmi, A. M., & Iqbal, I. M. (2017). Using k-means++ algorithm for researchers clustering. *AIP Conference Proceedings, 1867*. https://doi.org/10.1063/1.4994455

Sai Krishna, T. V., Yesu Babu, A., & Kiran Kumar, R. (2018). Determination of optimal clusters for a non-hierarchical clustering paradigm k-means algorithm. *Lecture Notes on Data Engineering and Communications Technologies*, *9*, 301–316. https://doi.org/10.1007/978-981-10-6319-0_26

Sammouda, R., & El-Zaart, A. (2021). An Optimized Approach for Prostate Image Segmentation Using K-Means Clustering Algorithm with Elbow Method. *Computational Intelligence and Neuroscience*, *2021*. https://doi.org/10.1155/2021/4553832

Sander, J. (2010). Density-Based Clustering. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 270–273). Springer US. https://doi.org/10.1007/978-0-387-30164-8_211

Sardar, T. H., & Ansari, Z. (2018). Partition based clustering of large datasets using MapReduce framework: An analysis of recent themes and directions. *Future Computing and Informatics Journal*, *3*(2), 247–261. https://doi.org/10.1016/j.fcij.2018.06.002

Sayad, S. (2018). *K-Means Clustering*. http://www.saedsayad.com/clustering_kmeans.htm

Schulam, P., Wigley, F., & Saria, S. (2015). Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data : Applications to Phenotyping and Endotype Discovery. *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, 2956–2964.

Sculley, D. (2010). Web-scale K-means Clustering. *Proceedings of the Nineteenth In- Ternational Conference on World Wide Web*, 1177–1178.

Selim, S. Z., & Ismail, M. A. (1984). K-Means-Type Algorithms: A Generalized

Convergence Theorem and Characterization of Local Optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(1), 81–87. https://doi.org/10.1109/TPAMI.1984.4767478

Shang, R., Ara, B., Zada, I., Nazir, S., Ullah, Z., & Khan, S. U. (2021). Analysis of Simple K- Mean and Parallel K- Mean Clustering for Software Products and Organizational Performance Using Education Sector Dataset. *Scientific Programming*, *2021*. https://doi.org/10.1155/2021/9988318

Sharma, P. K., & Holness, G. (2017). L2-norm transformation for improving k-means clustering. *International Journal of Data Science and Analytics*, *3*(4), 247–266. https://doi.org/10.1007/s41060-017-0054-1

Shim, J. P., French, A. M., Guo, C., & Jablonski, J. (2015). Big data and analytics: Issues, solutions, and ROI. *Communications of the Association for Information Systems*, *37*(1), 797–810. http://www.scopus.com/inward/record.url?eid=2-s2.0-84945364853&partnerID=40&md5=8e3026ebc07265f5da8519771bdf8f31

Shindler, M., Wong, A., & Meyerson, A. (2011). Fast and accurate κ-means for large datasets. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS*, 2375–2383.

Shirkhorshidi, A. S., & Aghabozorgi, S. (2014). Big Data Clustering: A Review. *Computational Science and Its Applications – ICCSA 2014*, *8583*(June). https://doi.org/10.1007/978-3-319-09156-3

Simpao, A. F., Ahumada, L. M., Gálvez, J. A., & Rehman, M. A. (2014). A review of analytics and clinical informatics in health care. *Journal of Medical Systems*, *38*(4). https://doi.org/10.1007/s10916-014-0045-x

Sindol, D. (2016). *Big Data Basics - Introduction to Big Data*. Edgewood Solutions. https://www.mssqltips.com/sqlservertip/3132/big-data-basics--part-1--introduction-to-big-data/

Sivaguru, M., & Punniyamoorthy, M. (2020). Performance-enhanced rough k -means clustering algorithm. *Soft Computing*, *2*(25). https://doi.org/10.1007/s00500-020-05247-2

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286. https://doi.org/10.1016/j.jbusres.2016.08.001

Smith, M., Szongott, C., Henne, B., & Voigt, G. Von. (2012). Big Data Privacy Issues in Public Social Media. *Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on. IEEE*, 1–6. https://doi.org/10.1109/DEST.2012.6227909

Soliman, T. H. A., Sewissy, A. A., & AbdelLatif, H. (2010). A gene selection approach for classifying diseases based on microarray datasets. *2010 2nd International Conference on Computer Technology and Development*, *lCCTD*, 626–631. https://doi.org/10.1109/ICCTD.2010.5645975

Spini, G., van Heesch, M., Veugen, T., & Chatterjea, S. (2020). Private Hospital Workflow Optimization via Secure k-Means Clustering. *Journal of Medical*

Systems, *44*(1). https://doi.org/10.1007/s10916-019-1473-4

Sreedevi, A. G., Nitya Harshitha, T., Sugumaran, V., & Shankar, P. (2022). Application of cognitive computing in healthcare, cybersecurity, big data and IoT: A literature review. *Information Processing and Management*, *59*(2), 102888. https://doi.org/10.1016/j.ipm.2022.102888

Sreedhar, C., Kasiviswanath, N., & Chenna Reddy, P. (2017). Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop. *Journal of Big Data*, *4*(1). https://doi.org/10.1186/s40537-017-0087-2

Stel, V. S., Pluijm, S. M. F., Deeg, D. J. H., Smit, J. H., Bouter, L. M., & Lips, P. (2008). A classification tree for predicting recurrent falling in community-dwelling older persons. *Journal of the American Geriatrics Society*, *51*(10), 1356–1364. https://doi.org/10.1046/j.1532-5415.2003.51452.x

Stonebraker, M., Abadi, D., DeWitt, D. J., Madden, S., Paulson, E., Pavlo, A., & Rasin, A. (2010). MapReduce and parallel DBMS. *Communications of the ACM*, *53*(1), 64. https://doi.org/10.1145/1629175.1629197

Subbalakshmi, C., Rama Krishna, G., Krishna Mohan Rao, S., & Venketeswa Rao, P. (2015). A method to find optimum number of clusters based on fuzzy silhouette on dynamic data set. *Procedia Computer Science*, *46*(Icict 2014), 346–353. https://doi.org/10.1016/j.procs.2015.02.030

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, *336*(1). https://doi.org/10.1088/1757-899X/336/1/012017

Tapia, J. J., Morett, E., & Vallejo, E. E. (2009). A Clustering Genetic Algorithm for Genomic Data Mining. *Foundations of Computational Intelligence*, *4*, 249–275. https://doi.org/10.1007/978-3-642-01088-0_11

Tariq RS, N. T. (2015). Big Data Challenges. *Computer Engineering & Information Technology*, *04*(03). https://doi.org/10.4172/2324-9307.1000133

Taylor-Sakyi, K. (2016). Understanding big data. *ArXiv.Org > Cs > ArXiv:1601.04602*, *January*, 166. https://doi.org/1 0 9 8 7 6 5 4 3 2 1

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of data clusters via the gap statistic. In *Journal of the Royal Statistical Society: Series B* (Vol. 63, Issue Part 2, pp. 411–423).

Troester, M. (2015). Big Data Meets Big Data Analytics. *SAS White Paper*. http://www.informatik-aktuell.de/entwicklung/methoden/big-data-meets-big-data.html?utm_source=CleverReach&utm_medium=email&utm_campaign=02-11-2015+Newsletter+19-2015&utm_content=Mailing_6367514

Tsapanos, N., Tefas, A., Nikolaidis, N., & Pitas, I. (2015). A distributed framework for trimmed Kernel k-Means clustering. *Pattern Recognition*, *48*(8), 2685–2698. https://doi.org/10.1016/j.patcog.2015.02.020

Twinkle, B. (2012). *What is DBMS? What is RDBMS? DBMS vs RDBMS | My Tec*

Bits. Oracle Database, SQL Server. https://www.mytecbits.com/microsoft/sql-server/what-is-dbms-what-is-rdbms

Umargono, E., Suseno, J. E., & Vincensius, G. S. K. (2020). K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean and Median. *Advances in Social Science, Education and Humanities Research*, *474*(Isstec 2019), 234–240. https://doi.org/10.5220/0009908402340240

Vattani, A. (2011). k-means Requires Exponentially Many Iterations Even in the Plane. *Discrete and Computational Geometry*, *45*(4), 596–616. https://doi.org/10.1007/s00454-011-9340-1

Velmurugan, T., & Santhanam, T. (2011). A survey of partition based clustering algorithms in data mining:An Experimental Approach. *Information Technology Journal*, *10*(3), 478–484.

Veloso, R., Portela, F., Santos, M. F., Silva, Á., Rua, F., Abelha, A., & Machado, J. (2014). A Clustering Approach for Predicting Readmissions in Intensive Medicine. *Procedia Technology*, *16*(December), 1307–1316. https://doi.org/10.1016/j.protcy.2014.10.147

Vuleta, B. (2020). *Big Stats and Facts About Big Data*. Seed Scientific. https://seedscientific.com/how-much-data-is-created-every-day/

Wang, J., Wang, J., Ke, Q., Zeng, G., & Li, S. (2015). Fast Approximate K-Means via Cluster Closures. In *Multimedia Data Mining and Analytics* (pp. 373–395). Springer International Publishing. https://doi.org/10.1007/978-3-319-14998-1_17

Wang, Q., Zhang, T., Ma, F., Wang, Y., & Yue, D. (2018). Improved Fuzzy K-means Clustering Based on Imbalanced Measure of Cluster Sizes. *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), Nanjing, China*, 548–551.

Wang, T., & Gao, J. (2019). An Improved K-Means Algorithm Based on Kurtosis Test. *Journal of Physics: Conference Series*, *1267*(1). https://doi.org/10.1088/1742-6596/1267/1/012027

Wang, X., & Bai, Y. (2016). The global Minmax k-means algorithm. *SpringerPlus*, *5*(1). https://doi.org/10.1186/s40064-016-3329-4

Wang, X., Li, Y., Wang, M., Yang, Z., & Dong, H. (2018). An Improved K _ means Algorithm for Document Clustering Based on Knowledge Graphs. *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1–5.

Wei, X., & Li, Y. (2017). Research on improved k-means algorithm based on hadoop. *Proceedings - 2017 4th International Conference on Information Science and Control Engineering, ICISCE 2017*, 593–598. https://doi.org/10.1109/ICISCE.2017.129

White, T. (2012). *Hadoop : The Definitive Guide* (3rd Editio). O'Reilly Press.

Wu, K., Zeng, W., Wu, T., & An, Y. (2015). Research and improve on K-means algorithm based on hadoop. *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, *2015-Novem*, 334–337. https://doi.org/10.1109/ICSESS.2015.7339068

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. In *Knowledge and Information Systems* (Vol. 14, Issue 1). Springer-Verlag London Limited. https://doi.org/10.1007/s10115-007-0114-2

Wu, Z., & Flintsch, G. W. (2015). Optimal Selection of Pavement Maintenance & Rehabilitation Projects. *7th International Conference on Managing Pavement Assets*, *January 2015*, 12. https://www.researchgate.net/publication/265352631

Xhafa, F., Bogza, A. M., Caballe, S., & Barolli, L. (2016). Apache Mahout's k−Means vs. Fuzzy k−Means Performance Evaluation. *Eigth International Conference on Intelligent Networking and Collaborative System*.

Xiao, B., Wang, Z., Liu, Q., & Liu, X. (2018). SMK-means: An improved mini batch k-means algorithm based on mapreduce with big data. *Computers, Materials and Continua*, *56*(3), 365–379. https://doi.org/10.3970/cmc.2018.01830

Xie, H., Zhang, L., Lim, C. P., Yu, Y., Liu, C., Liu, H., & Walters, J. (2019). Improving K-means clustering with enhanced Firefly Algorithms. *Applied Soft Computing Journal*, *84*, 105763. https://doi.org/10.1016/j.asoc.2019.105763

Xing, R., & Li, C. (2019). Fuzzy c-means algorithm automatically determining optimal number of clusters. *Computers, Materials and Continua*, *60*(2), 767–780. https://doi.org/10.32604/cmc.2019.04500

Xiong, Y., Peng, Q., & Zhang, Z. (2020). Research on MapReduce parallel optimization method based on improved K-means clustering algorithm. *ACM International Conference Proceeding Series*, 0–5. https://doi.org/10.1145/3414274.3414282

Xu, H., Fu, C., Yao, S., & Zong, X. (2019). An improved k-means algorithm based on intersection over union for network security. *2019 IEEE 11th International Conference on Communication Software and Networks, ICCSN 2019*, 514–517. https://doi.org/10.1109/ICCSN.2019.8905279

Yamashita, N., & Adachi, K. (2020). A Modified k-Means Clustering Procedure for Obtaining a Cardinality-Constrained Centroid Matrix. *Journal of Classification*, *37*(2), 509–525. https://doi.org/10.1007/s00357-019-09324-6

Yang, J.-J., Li, J., Mulder, J., Wang, Y., Chen, S., Wu, H., Wang, Q., & Pan, H. (2015). Emerging information technologies for enhanced healthcare. *Computers in Industry*, *69*(C), 3–11. https://doi.org/10.1016/j.compind.2015.01.012

Yaramala, D. (2016). Health Care Data Analytics Using Hadoop. In *San Diego State University*.

Yenkar, V., & Bartere, M. (2014). Review on " Data Mining with Big Data ." *International Journal of Computer Science and Mobile Computing*, *3*(4), 97–

102.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, *36*(4), 2431–2448. https://doi.org/10.1007/s10916-011-9710-5

Yu, S. S., Chu, S. W., Wang, C. M., Chan, Y. K., & Chang, T. C. (2018). Two improved k-means algorithms. *Applied Soft Computing Journal*, *68*, 747–755. https://doi.org/10.1016/j.asoc.2017.08.032

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*(2), 226–235. https://doi.org/10.3390/j2020016

Zakir, J. (2015). Big Data Analytics. *Issues in Information Systems*, *16*(2), 81–90. https://doi.org/10.1007/978-3-662-55469-2_14

Zhang, G., Zhang, C., & Zhang, H. (2018). Improved K-means algorithm based on density Canopy. *Knowledge-Based Systems*, *145*, 289–297. https://doi.org/10.1016/j.knosys.2018.01.031

Zhang, K., & Wang, H. (2015). Cancer Genome Atlas Pan-cancer analysis project. *Chinese Journal of Lung Cancer*, *18*(4), 219–223. https://doi.org/10.3779/j.issn.1009-3419.2015.04.02

Zhang, L., Qu, J., Gao, M., & Zhao, M. (2019). Improvement of k-means algorithm based on density. *Proceedings of 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, ITAIC 2019*, *Itaic*, 1070–1073. https://doi.org/10.1109/ITAIC.2019.8785550

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An Efficient Data Clustering Method for Very Large Databases. *ACM SIGMOD International Conference on Management of Data*, *1*, 103–114. https://doi.org/10.1145/233269.233324

Zhang, Y. L., & Wang, Y. N. (2018). An improved sampling K-means clustering algorithm based on MapReduce. *ICNC-FSKD 2017 - 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 1934–1939. https://doi.org/10.1109/FSKD.2017.8393064

Zhao, W. L., Deng, C. H., & Ngo, C. W. (2018). k-means: A revisit. *Neurocomputing*, *291*, 195–206. https://doi.org/10.1016/j.neucom.2018.02.072

LIST OF PUBLICATIONS

Ashabi, A., Sahibuddin, S. B., & Haghighi, M. S. (2020, April). Big data: Current challenges and future scope. In *2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)* (pp. 131-134). IEEE.

Ashabi, A., Sahibuddin, S. B., & Haghighi, M. S. (2020, December). The Systematic Review of K-Means Clustering Algorithm. In *2020 The 9th International Conference on Networks, Communication and Computing* (pp. 13-18). ACM.