UNIVERSITI PUTRA MALAYSIA

STATISTICAL DATA PREPROCESSING METHODS IN DISTANCE FUNCTIONS TO ENHANCE K-MEANS CLUSTERING ALGORITHM

PAUL INUWA DALATU

STATISTICAL DATA PREPROCESSING METHODS IN DISTANCE FUNCTIONS TO ENHANCE K-MEANS CLUSTERING ALGORITHM

By

PAUL INUWA DALATU

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

January 2018

DEDICATIONS

This thesis is dedicated to:
My late elder brother Retired Superintendent of Police;
Rtd Mohammed Inuwa Dalatu
My wife;
Mrs Rebecca Paul, and
My children;
Usaku,
Nachamada,
Chimda, and
Biyama.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

STATISTICAL DATA PREPROCESSING METHODS IN DISTANCE FUNCTIONS TO ENHANCE K-MEANS CLUSTERING ALGORITHM

By

PAUL INUWA DALATU

January 2018

Chairman : **Professor Habshah Midi, PhD**
Faculty : **Science**

Clustering is an unsupervised classification method with major aim of partitioning, where objects in the same cluster are similar, and objects belong to different clusters vary significantly, with respect to their attributes. The K-Means algorithm is the commonest and fast technique in partitional cluster algorithms, although with unnormalized datasets it can achieve local optimal.

We introduced two new approaches to normalization techniques to enhance the K-Means algorithms. This is to remedy the problem of using the existing Min-Max (MM) and Decimal Scaling (DS) techniques, which have overflow weakness. The suggested approaches are called new approach to min-max (NAMM) and decimal scaling (NADS).

The Hybrid mean algorithms which are based on spherical clusters is also proposed to remedy the most significant limitation of the K-Means and K-Midranges algorithms. It is attained successfully by combining the mean in K-Means algorithm, minimum and maximum in K-Midranges algorithm and compute their average as mean cluster of Hybrid mean.

The problem of using range function in Heterogeneous Euclidean-Overlap Metric (HEOM) is addressed by replacing the range with interquartile range function called Interquartile Range-Heterogeneous Metric (IQR-HEOM). Dividing the HEOM with range allows outliers to have big effect on the contribution of attributes. Hence, We proposed interquartile range which is more resistance against outliers in data pre-processing. It shows that the IQR-HEOM method is more efficient to rectify the problem caused by using range in HEOM.

The Standardized Euclidean distance which uses standard deviation to down weight maximum points of the ith features on the distance clusters are being criticized in the literature by many researchers that the method is prone to outliers and has 0% breakdown points. Therefore, to remedy the problem, we introduced two statistical estimators called Q_n and S_n estimator, both have 50% breakdown points, with their efficiency as 58% and 82% for S_n and Q_n, respectively. The empirical evidences show that the two suggested methods are more efficient compared to the existing methods.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

KAEDAH DATA BERSTATISTIK PRAPEMPROSESAN DALAM FUNGSI JARAK UNTUK MENINGKATKAN ALGORITMA K-MEANS KLUSTER

Oleh

PAUL INUWA DALATU

January 2018

Pengerusi : **Professor Habshah Midi, PhD**
Fakulti : **Sains**

Pengelompokan adalah kaedah pengelasan tanpa pengawasan dengan tujuan utama pembahagian, dengan objek dalam kluster yang sama adalah serupa, dan objek kepunyaan kluster berbeza, perbezaannya adalah ketara, dengan sifat mereka masing-masing. Algoritma *K-Means* adalah teknik yang paling biasa dan cepat dalam algoritma kluster terpetak, walaupun dengan set data yang tidak dipiawaikan ia boleh mencapai optimum setempat.

Kami memperkenalkan dua pendekatan baru untuk teknik normalisasi untuk meningkatkan algoritma *K-Means*. Ini adalah untuk memperbaiki masalah penggunaan teknik sedia ada ia-itu Min-Max (MM) dan penskalaan perpuluhan (DS), yang mempunyai banyak kelemahan. Pendekatan yang dicadangkan dipanggil pendekatan baru untuk min-max (NAMM) dan pendekatan baru untuk penskalaan perpuluhan (NADS).

Algoritma Hibrid min yang berdasarkan kluster sfera juga di cadangkan untuk menyelesaikan batasan paling signifikan bagi algoritma *K-Means* dan *K-Midranges*. Ia berjaya dicapai dengan menggabungkan min dalam algoritma *K-Means*, minimum dan maksimum dalam algoritma *K-Midranges* dan mengira purata nya sebagai min kluster purata hibrid.

Masalah menggunakan fungsi renj dalam *Heterogen Euclidean-Overlap Metric* (HEOM) di tangani dengan menggantikan renj dengan fungsi renj interkuantil yang

dinamakan *Interquartile Range-Heterogen Metric* (IQR-HEOM). Membahagikan HEOM dengan renj membenarkan titik terpencil mempunyai kesan besar terhadap sumbangan atribut. Oleh yang demikian kami mencadangkan renj interkuantil yang lebih teguh terhadap titik terpencil bagi data prapemprosesan. Hasil kajian menunjukkan bahawa kaedah IQR-HEOM lebih efisien untuk memperbetulkan masalah yang disebabkan oleh penggunaan renj dalam HEOM.

Jarak Euclidan Terpiawai yang menggunakan sisihan piawai untuk menurinkan pemberat titik maximum ciri-ciri i pada kluster jarak telah dikritik oleh banyak penyelidik dalam literatur di mana kaedah ini terdedah kepada titik terpencil dan mempunyai 0% titik musnah. Oleh itu, untuk menyelesaikan masalah ini, kami telah memperkenalkan dua penganggar statistik yang dinamakan penganggar Qn dan Sn, kedua-duanya mempunyai 50% titik musnah, dengan kecekapan mereka sebanyak 58% dan 82$ masing-masing bagi Sn dan Qn. Bukti empirik menunjukkan bahawa kaedah yang dicadangkan adalah lebih efisien dibandingkan dengan kaedah yang sedia ada.

ACKNOWLEDGEMENTS

First of all, I am grateful to the Almighty God for his love, mercy ,guidance, protection, direction, the good health, wellbeing granted and the great opportunity that were necessary for me to complete this study.

Foremost, I would like to express my profound gratitude to my supervisor, Prof.Dr. Habshah Midi, for the continuous support of my Ph.D. study and research, whose expertise ,great effort, understanding, motivation, enthusiasm, and patience, added significantly to my graduate experience. I appreciate her vast knowledge and skills in many areas, for much of her assistance she provided at all levels of this study most especially her assistance in writing this thesis.

I would like to express my very great appreciation and thanks to Dr. Aida Mustapha for her valuable, direction, recommendations and constructive suggestions most especially outlining some modern techniques in writing academic article. Her willingness to give her time so generously despite lots of academic schedules has been very much appreciated.

I would like to extend my sincere thanks to Dr. Alihossein Aryanfar for his help in offering me resources for the formulation of codes in programs, which necessitated the evaluations and analysis of all my data sets results.

I would also like to thank my family for their patience and the moral support they provided me through my entire life and in particular this study, I must acknowledge my wife, Rebecca for her love and prayer and my best friend, Mathias Ibrahim, who takes the whole responsibility to see that my children are comfortable and sound at school.

In conclusion, I recognize that this research would not have been possible without the financial assistance of Tertiary Education Trust Fund (TETFund) which was established as an intervention agency under the TETFund ACT-Tertiary Education Trust Fund (Establishment, etc.,) Act, 2011; charged with the responsibility for managing, disbursing and monitoring the education tax, and Adamawa State University, Mubi, Nigeria, who approved the grant, I express my gratitude and appreciation to this agency and the institution.

I certify that a Thesis Examination Committee has met on 3 January 2018 to conduct the final examination of Paul Inuwa Dalatu on his thesis entitled "Statistical Data Preprocesing Methods in Distance Functions to Enhance K-Means Clustering Algorithm" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Thesis Examination Committee were as follows:

Fudziah binti Ismail, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Noor Akma binti Ibrahim, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Azami Zaharim, PhD
Professor
Universiti Kebangsaan Malaysia
Malaysia
(External Examiner)

A.H.M. Rahmatullah Imon, PhD
Professor
Ball State University
United States
(External Examiner)

NOR AINI AB. SHUKOR, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 28 March 2018

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Habshah Midi, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Chairperson)

Jayanthi Arasan, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

Ibragimov Gafurjan, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

ROBIAH BINTI YUNUS, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature:____________________ Date:____________________

Name and Matric No: Paul Inuwa Dalatu, GS 38183

Declaration by Members of Supervisory Committee

This is to confirm that:
- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: ______________________________
Name of Chairman of Supervisory Committee:
Professor Dr. Habshah Midi

Signature: ______________________________
Name of Member of Supervisory Committee:
Associate Professor Dr. Jayanti Arasan

Signature: ______________________________
Name of Member of Supervisory Committee:
Associate Professor Dr. Ibragimov Gafurjan

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DS	Decimal Scaling
HEOM	Heterogeneous Euclidean-Overlap Metric
IQR	Interquartile Range
MDP	Minimum Diameter Partitioning
MEMS	Micro-Electro-Mechanical Systems
MM	Min-Max
NADS	New Approach to Decimal Scaling
NAMM	New Approach to Min-Max
NP	Nondeterministic Polynomial Time
ODM	Outliers Detection Model
rn-diff	range-normalized difference
ROC	Receiver Operating Characteristic curve
TETFund	Tertiary Education Trust Fund
TNR	True Negative Rate
TPR	True Positive Rate
UCI	University of California, Irvine
Z	Z-score

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Data clustering is a general method for statistical data analysis, which is most commonly used in numerous areas such as image analysis, pattern recognition and bioinformatics (Sundararajan and Karthikeyan, 2014). According to Sarma et al. (2013), clustering can be considered as an essential instrument in numerous applications like biology, marketing, information retrieval, remote sensing, pattern recognition, image processing, and text mining. Clustering groups data instances into subsets in such a way that similar instances are grouped together, while dissimilar instances belong to different groups. The instances are ordered into an efficient illustration that describes the population being sampled. Clustering of points or objects started as early as the human requirement for labeling the significant features of men and objects, classifying them with a type (Rokach and Maimon, 2014).

Unsupervised clustering processes are important tools in exploratory data analysis. As clustering conditions are usually based on some distance measures between individual data vectors, they are extremely sensitive to the scale, or dispersion of the variables (Vesanto, 2001).

The aim of feature selection in clustering is to classify a subset of significant features from the unique illustration space. The recognized important features are useful for data clustering that targets to maximize the between-cluster scatter and minimize within-cluster scatter (Chen, 2015). It is also important to note that the measurement of distance is essential in the cluster analysis process as most clustering methods start with the computation of a matrix of distances (Doherty et al., 2004).

Though clustering is a valuable and challenging problem with unlimited potential in applications, its presentation must be carefully controlled. Else, the method can simply be abused or misused. The number of clusters and distance measures are the two most important rules of clustering analysis, which affect the general quality of the outcomes (Mok et al., 2012). Therefore, pre-processing the datasets is crucial especially in terms of normalization.

The most common clustering method is the K-Means algorithm (Reddy et al., 2012). While it is very simple and strong in clustering large datasets, the technique suffers from a few drawbacks. The user needs to ascertain the number of clusters which is difficult to know in advance for many real world data sets. Nonetheless, the main problems it suffers is that, it is very sensitive for the selection of initial cluster centers.

Equally, it may result not always yielding global optimum outcomes.

Consequently, in order to overcome these aforementioned problems, many researchers had proposed new algorithms and some new distance functions to overcome the weakness in K-Means (Jain, 2010). The best appropriate measures to use in practice stay unidentified. Certainly, there are many inspiring validation matters which have not been completely addressed in the clustering works. For example, the position of normalizing validation measures has not been entirely proven.

Similarly, the relationship between dissimilar validation measures is not clear (Wu et al., 2009). Clustering validation, which calculates the goodness of clustering outcomes, has long been known as one of the vital problems critical to the achievement of clustering applications (Liu et al., 2010).

1.2 Significance of the Study

The major purpose of clustering approaches is to partition a set of objects into dissimilar groups, called clusters. These groups may be consistent in terms of similarity of its members. As the name implies, the representative-based clustering approaches apply some procedures of representation for each cluster. Consequently, each group has a member that signifies it. The word cluster analysis does not identify a specific statistical method or model, as do discriminant analysis, factor analysis, and regression. One does not have to make frequently any assumptions about the fundamental distribution of the data. K-Means clustering is a kind of unsupervised learning, which is used when one has unlabeled data.

The aim of this algorithm is to find groups in the data, with the number of groups represented by variables k. The algorithm processes iteratively to allocate each data points to one of k groups established on the features that are delivered. Data points are clustered founded on feature similarity. Therefore, knowledge about the cluster analysis that can occur in numerous data sets will assist researchers to choose on the actual situations when considering such characteristics like no assumptions should be made and the data sets are unlabeled. It will provide policy makers in different sectors of life with a better comprehension of many approaches, while, giving more rooms to researchers to decide about better data accuracy in meeting the present days challenges.

The K-Means clustering, to be specific while using heuristics such as Lloyd‘s algorithm (1957 but only published in (Lloyd, 1982)), is reasonably easy to implement and use even on large data sets. Clustering approaches have extensive use and are significance currently. This significance tends to increase as the volume of data grows and the processing power of the computer increases. Clustering applications are used extensively and successfully in several fields such as artificial intelligence, pattern recognition, ecology, psychiatry and marketing.

1.3 The Problem Statement

The main aim of data preparation is to get total assurance that the quality of the data before it is applied to any learning algorithms. The types of the data preparation according to Ogasawara et al. (2010), includes data cleaning, integration and transformation, and reduction. Therefore, our study is limited on data transformation methods, which are basically focused on min-max and decimal scaling respectively. Normalization means scaling down the value of the magnitudes to some appreciable low values, for instance, among the features, if there is frequently large difference between the maximum and minimum values, for example 1000 and 1.

Consequently, the most popular normalization methods used in the literature for data transformation are the min-max (where the data inputs are transformed into a predefined range 0 or -1 to 1), the z-score (where the values of an attribute A are normalized agreeing to its mean and standard deviation), and the decimal scaling (where the decimal point of the data values of an attribute A are moved according to its maximum absolute value). Furthermore, Liu et al. (2011) and Jain et al. (2005) have identified one of the weaknesses of using both the min-max and decimal scaling in data transformation. They stated that both of the techniques will have overflow problem, this makes the two technique not robust. However, Jain et al. (2005) and Zumel and Mount (2013) suggested that, in order to remedy this problem in decimal scaling approach, we have to apply $\log_{10}\max(x_i)$. While, in min-max approach Milligan (1989) and Liu et al. (2011) suggested to down weighting the technique so that irrelevant variables approach near zero. Therefore, we are motivated by lack of robustness of the two methods to adopt the ideas suggested by (Zumel and Mount (2013), Liu et al. (2011), Jain et al. (2005), and Milligan (1989)) to improve the methods of min-max (Jayalakshmi and Santhakumaran, 2011), and decimal scaling (Han et al., 2011). Therefore, our proposed methods are called new approach to min-max (NAMM) and new approach to decimal scaling (NADS). Hence, to our best knowledge, nothing have yet been done to improve the robustness and down weighting of the normalization by min-max and decimal scaling.

However, for spherical clusters, the most common algorithm popularly known for is K-Means, which minimizes the sum of squared Euclidean distances of the objects to the mean of the cluster (MacQueen, 1967). Furthermore, de Amorim and Makarenkov (2016) added that this problem of spherical shapes may lead to no assurance for K-Means algorithms will reach global optimum. In Rousseeuw and Hubert (2011), they also stated that , this particular method is not robust as it applies group means. However, Shanmugavadivu and Rajeswari (2012) also stated that, the major important limitation of K-Means clustering algorithms is its concept which is based on spherical clusters that are distinguishable in a way that the mean value converges towards the cluster center. Brusco and Steinley (2014), suggested using closely related to the classic problem of minimum diameter partitioning (MDP), where the diameter of a cluster is the largest distance between any pair of points within cluster. Therefore, we were motivated by the ideas of Brusco and Steinley (2014) and the work of Shanmugavadivu and Rajeswari (2012), where they combined the mean in K-Means and the maximum in K-Midrange and divided it by two to form the modified mean to remedy the problem of spherical shapes, whereby the approach depends on means as cluster centers. Hence, on our part we combined the mean in K-Means with the minimum and maximum in K-Midranges to form hybrid mean. This suggested algorithm will improve the dependence on means from K-Means and added to the potential of K-Midrange in cluster analysis. To our knowledge, nothing yet has been done to address the spherical concept in K-Means algorithm by using hybrid mean as a center for each cluster centers.

However, it is important to mention that the Heterogeneous Euclidean-Overlap Metric (HEOM) needs no normalization as it executes local normalization using range function (ChitraDevi et al., 2012). However, according to Singh and Leavline (2016) the procedure applied in HEOM, by dividing it with range tolerates outliers to have intense effect on the contribution of the attributes. Furthermore, Rousseeuw and Hubert (2011) pointed out the breakdown points for range is 0% (meaning that it can be contaminated by single point). Therefore, Singh and Leavline (2016) recommended using interquartile range which is more robust to range against outliers in data preprocessing. Hence, Rousseeuw and Croux (1993), pointed out that the interquartile range has 25% breakdown point compared to range which has 0%. This problem motivates us to propose IQR-HEOM, by replacing the range function in the existing HEOM (ChitraDevi et al., 2012) with interquartile range function. Therefore, to the best of our knowledge, no research has been done to study the interquartile range as an alternative to range in HEOM for data preprocessing.

Furthermore, Xu and Tian (2015), used another Weighted Euclidean called Standardized Euclidean (see Equation 6.1), they claimed that the larger *si* (denotes the standard deviation of the dataset) the smaller is the effect of the ith feature on the distance. Which they believed that the reason behind the method is the assumption that both normal and anomalous may appear from different cluster in feature space. Hence, the data may contain outliers which do not belong to a bigger cluster, yet the K-Means clustering algorithm functions as long as the number of outliers is small. Recently, Gerstenberger and Vogel (2015) criticized the method, that as far as using standard deviation to down weight maximum points, its prone to outliers and lack robustness.

Therefore, this weakness motivated us to replace the standard deviation which has 0% breakdown point (Rousseeuw and Hubert, 2011) and its lack of robustness. It is also susceptible to outliers and its low efficiency at heavy-tailed distribution (Gerstenberger and Vogel, 2015). We introduced two statistical estimators called Q_n and S_n estimators, both have 50% breakdown points and with their efficiency as; S_n is 58% and Q_n is 82% (Rousseeuw and Croux, 1993). The two proposed methods are called Q_n-Weighted Euclidean distance and S_n-Weighted Euclidean distance, which both will improve (Xu and

Tian, 2015) of lack robustness, low breakdown points and also low efficiency. However, to the best of our knowledge, we are the first researchers in distance-based clustering analysis to apply some statistical estimators to improve the efficiency and accuracy of K-Means clustering algorithm.

1.4 Research Objectives

The main goal of this study is to improve the performance of a K-Means clustering algorithm via statistical approach. In order to achieve the goal, the following objectives are required:

1. To propose new approaches to normalization techniques in cluster analysis.
2. To propose hybrid mean algorithms from K-Means and K-Midranges clustering algorithms.
3. To introduce statistical interquartile range into heterogeneous distance function.
4. To introduce Q_n estimator and S_n estimator into Standardized Euclidean distance function.

1.5 Scope and Limitation of the Study

Cormack (1971) proposed that clusters should be internally cohesive and externally isolated, entailing a certain degree of homogeneity within clusters and heterogeneity between clusters. Generally, clustering does not provide any statistical assumptions to data (Cao et al., 2009). In the past, many researchers tried to operationalize this meaning by minimizing within-group variation (see (Cox, 1957), (Engelman and Hartigan, 1969), (Fisher, 1958), and (Thorndike, 1953). Subsequently, these prompt efforts at maximizing within-group homogeneity (Sebestyen, 1962). MacQueen (1967) individually established the K-Means method as an approach that tries to find optimal partitions. Therefore, this type of classification is known as unsupervised learning (clustering), it is an exploratory or descriptive in nature, meaning that the investigator does not have pre- specified models or hypotheses but wants to know the general characteristic or arrangement of the high-dimensional data (Jain, 2010). Clustering has been used in a widespread diversity of fields, such as; engineering (machine learning, artificial intelligence, pattern recognition, mechanical engineering, electrical engineering), computer sciences (web mining, spatial database analysis, textual document collection, image segmentation), life and medical sciences (genetics, biology, microbiology, palcontology, psychiatry, clinic, pathology), earth sciences (geography, geology, remote sensing), social sciences (sociology, psychology, archeology, education), and economics (marketing, business) (Xu and Wunsch, 2008).

The K-Means clustering algorithm is generally applied in data clustering. The most essential unsupervised learning problem can be considered as data clustering. It deals with finding a structure or organization in a collection of unlabeled data (Su et al., 2009).

In statistical clustering problems, there are different categories of measures for the similarity or difference between objects. It is well-known that Euclidean distance is the popular used as a measure of difference, and minimization within clusters is equally to minimizing within group mean square error. Hence, the size of the Euclidean distribution between two objects is dependent on the scales of measurement of the characteristics of the objects. No definite or acceptable rule for weighting characteristics has been suggested (Matthews, 1979), though some many statisticians recommend normalizing each characteristics by some measure of its variability, to give the characteristics equal weight. A potential benefit of a variable weighting algorithm is the possibility that such a procedure would assign near zero weights to variables which are irrelevant to the clustering that exists in the remaining data. A variable weighting algorithm could reduce or eliminate this masking effect, which would be a useful contribution to classification technology (Milligan, 1989). Therefore, the measurement of similarity or distance is fundamental in the cluster analysis process as most clustering techniques begin with the calculation of a matrix distances (or dissimilarities) (Doherty et al., 2004).

In order to learn a new object or understand a new phenomenon, people always try to seek the features that can describe it, and further compare it with other known objects or phenomena, based on the similarity or dissimilarity, generalized as proximity, according to some certain standards or rules (Xu and Wunsch, 2005). Normally, there are three types of testing criteria: external indices, internal indices, and relative indices. The three indices are defined on the three major categories of clustering organizations, well-known as partitional clustering, hierarchical clustering , and individual clusters (Cadez et al., 2000).

Therefore, our scope are limited to external and internal indices; although, the internal indices had only one chapter in the current dispensations. However, external indices are based on some pre specified arrangement, which is the likeness of prior information on the data, and used as a rule to validate the clustering solutions. While, internal indices are not dependent on external information (prior knowledge). Differently, they test the clustering organization right from the original data.

However, Jain and Dubes (1988) referred to cluster validity as the formal processes that evaluate the results of cluster analysis in quantitative and objective approach. Although, Jain and Dubes (1988) stated that, clustering validation has long been acknowledged as one of the vibrant problems important to the achievement of clustering applications. However, Wu et al. (2009) pointed out that, in spite of the enormous amount of professional struggle spent on this problem, there is no reliable and definite solution to cluster validation. The best appropriate measures to apply in practice remain unidentified. They added that, certainly , there are many challenging validation problems which have not been fully addressed in the clustering literature.

For example, the significance of normalizing validation measures has not been fully recognized. There is no universally defined rule for normalizing datasets and thus, the choice of a particular normalization rule is largely left to the discretion of the user (Singh et al., 2015). It is worthwhile to enhance clustering quality by normalizing the dynamic range of input data objects into specific range (de Souto et al., 2008).

1.6 Methodology

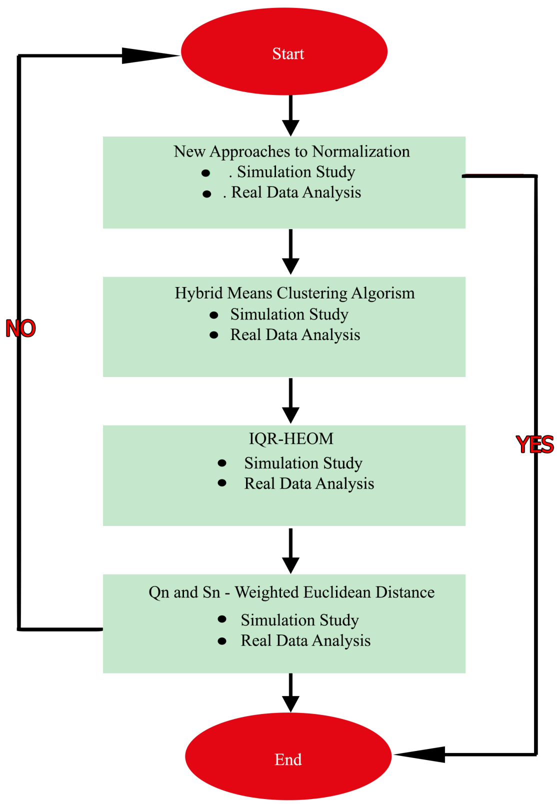

Figure 1.1: Flow Chart showing Flow of the Methodology

Note: Interquartile Range-Heterogeneous Euclidean-Overlap Metric (IQR-HEOM). Figure 1.1, presents flow chart showing flow of the methodology. The methodology comprises of four contributing chapters, starts from Chapter 3 which has two suggested normalization techniques called New Approaches to Min-Max (NAMM)and Decimal Scaling (NADS). Chapter 4 has proposed algorithm called Hybrid Means Algorithms. This proposed algorithm was combined from K-Means and K-Midranges algorithms. Chapter 5, interquartile range was introduced into Heterogeneous Euclidean-Overlap Metric (HEOM) to replace range as local normalization and the proposed method is called Interquartile Range-Heterogeneous Euclidean-Overlap Metric (IQR-HEOM). Chapter 6, two statistical estimators Qn and Sn was introduced into Standardized Euclidean distance to replace standard deviation as a local normalization, the suggested methods are Qn and Sn-Weighted Euclidean distance.

1.7 Organization of Thesis

The following is a brief description of the contents of each chapter. This chapter serves as an essential introduction of this study by presenting background of the study, statement of problem / motivation of study, significance of the study, research objectives, definition of terms, scope and limitation of the study. In accordance with the objectives and the scope of the study, the contents of this dissertation are organized as follows.

Chapter 2: Literature Review. This comprises of some reviews on the development of clustering analysis from published materials on clustering and its outcomes,types of clustering analysis, and some applications of clustering analysis in different fields of sectors. We also, provided K-Means clustering algorithm, general proximity measures through distance functions, proximity measures for numerical data, proximity measures for discrete data, and as well as proximity measures for mixed data.

Chapter 3: New Approaches to Normalization Techniques for External Validity Measures in K-Means Clustering Algorithm. The main subject in this chapter is that, we proposed new approaches to normalization techniques using the two most prominent data preprocessing such as; min-max, and decimal scaling. Consequently, we had comparison of the approaches through some outcomes from real datasets and generated data set applying simulated annealing clustering analysis method.

Chapter 4: Introduction of Hybrid Mean Algorithms from K-Means and K-Midranges Clustering Algorithms. We proposed a hybrid mean algorithms by combining the effectiveness of K-Means algorithm and K-Midranges algorithm; then averaging mean from K-Means and minimum, maximum from K-Midranges. However, we evaluated the two conventional algorithms and the suggested algorithm using nine distance functions testing on three benchmark data sets and simulated data set.

Chapter 5: Statistical Approaches for Data Preprocessing in Enhancing Heterogeneous Distance Functions. In this chapter, we are able to use three UCI datasets; supported by generated data set. The conventional method used in this section is called "Heterogeneous Euclidean-Overlap Metric $(HEOM)$" and from the ideas of this $HEOM$ we suggested $IQR-HEOM$ method. We applied internal validity measures such as silhouette coefficients and cohesion values to examine the capability and accuracy of the conventional method against the proposed method through the results obtained.

Chapter 6: K-Means Algorithms based on Weighted Euclidean Distance Here we proposed two approaches such as Q_n weighted Euclidean distance, and S_n weighted Euclidean distance. We used the ideas from Standardized weighted Euclidean distance (sometimes called Normalized weighted Euclidean distance). We experimented the two suggested methods on three real data sets from benchmark datasets and generated data set. However, the two proposed methods introduced from weighted Euclidean distance have shown better results compared to the existing traditional methods.

Chapter 7: Conclusions and Recommendations for Future Research. This serves as the last chapter, which consists the conclusions from the outcomes of real data sets and from simulated data set. Hence, we recommended and suggested some possibilities for future research.

BIBLIOGRAPHY

Abubaker, M. and Ashour, W. (2013). Efficient data clustering algorithms: Improvements over K-means. *International Journal of Intelligent Systems and Applications*, 5(3):37.

Aha, D. W. (1992). Tolerating noisy, irrelevant and novel attributes in *Instance*-based learning algorithms. *International Journal of Man-Machine Studies*, 36(2):267–287.

Aksoy, S. and Haralick, R. M. (2001). Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern recognition letters*, 22(5):563–582.

Ali, B. B. and Massmoudi, Y. (2013). K-means clustering based on gower similarity coefficient: A comparative study. In *Modeling, Simulation and Applied Optimization (ICMSAO), 2013 5th International Conference on*, pages 1–5.

Bache, K. and Lichman, M. (2013). Uci machine learning repository. *University* of *California*, *School* of *Information* and *Computer Science*, irvine, ca. *Retrieved from the World Wide Web October*, 27:2014.

Ball, G. H. and Hall, D. J. (1965). Isodata, a novel method of data analysis and pattern classification. Technical report.

Barakbah, A. R. and Kiyoki, Y. (2009). A pillar algorithm for k-means optimization by distance maximization for initial centroid designation. In *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*, pages 61–68.

Behera, H., Lingdoh, R. B., and Kodamasingh, D. (2011). An improved hybridized k-means clustering algorithm (ihkmca) for highdimensional dataset & its performance analysis. *Int. J. Comput. Sci. Eng.(IJCSE)*, 3(3):1183–1190.

Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013). Evaluation measures for modelsassessment over imbalanced datasets. *Journal Of Information Engineering and Applications*, 3(10).

Belhaouari, S. B., Ahmed, S., and Mansour, S. (2014). Optimized K-means algorithm. *Mathematical Problems in Engineering*, 2014.

Benson-Putnins, D., Bonfardin, M., Magnoni, M. E., and Martin, D. (2011). Spectral clustering and visualization: a novel clustering of fishers iris data set. *SIAM Undergraduate Research Online*, 4.

Bordogna, G. and Pasi, G. (2012). A quality driven hierarchical data divisive soft clustering for information retrieval. *Knowledge-based systems*, 26:9–19.

Brown, A. and Hill, K. (2009). *Tasks and Criteria in Performance Assessment: Proceedings of the 28th Language Testing Research Colloquium*, volume 13.

Brusco, M. J. and Steinley, D. (2014). Model selection for minimum-diameter partitioning. *British Journal of Mathematical and Statistical Psychology*, 67(3):471–495.

Cadez, I. V., Gaffney, S., and Smyth, P. (2000). A general probabilistic framework for clustering individuals and objects. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 140–149.

Cao, F., Liang, J., and Jiang, G. (2009). An initialization method for the k-means algorithm using neighborhood model. *Computers & Mathematics with Applications*, 58(3):474–483.

Carroll, J. D. and Chaturvedi, A. (1998). k-midranges clustering. In *Advances in data science and classification*, pages 3–14.

Cerioli, A. (2005). K-means cluster analysis and mahalanobis metrics: a problematic match or an overlooked opportunity. *Statistica Applicata–Italian Journal of Applied Statistics*, 17(1):61–73.

Chan, E. Y., Ching, W. K., Ng, M. K., and Huang, J. Z. (2004). An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern recognition*, 37(5):943–952.

Chaturvedi, A., GREEN, P., and CARROLL, J. (1996). Market segmentation via k-modes clustering. In *Invited paper presented at the American Statistical Association conference held in Chicago*.

Chen, C.-H. (2015). Feature selection for clustering using instance-based learning by exploring the nearest and farthest neighbors. *Information Sciences*, 318:14–27.

ChitraDevi, N., Palanisamy, V., Baskaran, K., and Prabeela, S. (2012). A novel distance for clustering to support mixed data attributes and promote data reliability and network lifetime in large scale wireless sensor networks. *Procedia Engineering*, 30:669–677.

Chitradevi, N., Palanisamy, V., Baskaran, K., and Swathithya, K. (2013). Efficient density based techniques for anomalous data detection in wireless sensor networks. *Ł*, 16(2):211–223.

Chu, C.-W., Holliday, J. D., and Willett, P. (2008). Effect of data standardization on chemical clustering and similarity searching. *Journal of chemical information and modeling*, 49(2):155–161.

Climent, J. and Hexsel, R. A. (2012). Iris recognition using adaboost and levenshtein distances. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(02):1266001.

Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society. Series A (General)*, pages 321–367.

Cox, D. R. (1957). Note on grouping. *Journal of the American Statistical Association*, 52(280):543–547.

Davis, J. J. and Clark, A. J. (2011). Data preprocessing for anomaly based network intrusion detection: A review. *Computers & Security*, 30(6):353–375.

de Amorim, R. C. and Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, 324:126–145.

de Amorim, R. C. and Makarenkov, V. (2016). Applying subclustering and l p distance in weighted k-means with distributed centroids. *Neurocomputing*, 173:700–707.

de Souto, M. C., de Araujo, D. S., Costa, I. G., Soares, R. G., Ludermir, T. B., and Schliep, A. (2008). Comparative study on normalization procedures for cluster analysis of gene expression datasets. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 2792–2798.

Dhaliwal, D. S., Sandhu, P. S., and Panda, S. (2011). Enhanced k-nearest neighbor algorithm. *World Academy of Science, Engineering and Technology*, 73:681–685.

Doherty, K., Adams, R., and Davey, N. (2004). Non-euclidean norms and data normalisation. In *ESANN*, pages 181–186.

Eckes, T. and Orlik, P. (1993). An error variance approach to two-mode hierarchical clustering. *Journal of Classification*, 10(1):51–74.

El Agha, M. and Ashour, W. M. (2012). Efficient and fast initialization algorithm for k-means clustering. *International Journal of Intelligent Systems and Applications*, 4(1):21.

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., and Buydens, L. M. (2013). Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, 50:96–106.

Engelman, L. and Hartigan, J. A. (1969). Percentage points of a test for clusters. *Journal of the American Statistical Association*, 64(328):1647–1648.

Firdaus, S. and Uddin, M. A. (2015). A survey on clustering algorithms and complexity analysis. *International Journal of Computer Science Issues (IJCSI)*, 12(2):62.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American statistical Association*, 53(284):789–798.

Galili, T. (2015). *Dendextend*: an r package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, 31(22):3718–3720.

Gan, G., Ma, C., and Wu, J. (2007). *Data clustering: theory, algorithms, and applications*.

García, S., Luengo, J., and Herrera, F. (2015). *Data preprocessing in data mining*.

García, S., Luengo, J., and Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, 98:1–29.

Gerstenberger, C. and Vogel, D. (2015). On the efficiency of ginis mean difference. *Statistical Methods & Applications*, 24(4):569–596.

Giancarlo, R., Bosco, G. L., and Pinello, L. (2010). Distance functions, clustering algorithms and microarray data analysis. In *Learning and Intelligent Optimization*, pages 125–138.

Gnanadesikan, R., Kettenring, J. R., and Tsao, S. L. (1995). Weighting and selection of variables for cluster analysis. *Journal of Classification*, 12(1):113–136.

Goldberg, D. (2005). Genetic algorithm in search, optimization and machine learning, pearson education. *Low Price Edition, Delhi*.

Guha, S., Rastogi, R., and Shim, K. (2001). Cure: an efficient clustering algorithm for large databases. *Information Systems*, 26(1):35–58.

Gürbüz, F., Özbakir, L., and Yapici, H. (2011). Data mining and preprocessing application on component reports of an airline company in turkey. *Expert Systems with Applications*, 38(6):6618–6626.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*.

Hand, D. J., Mannila, H., and Smyth, P. (2001). *Principles of data mining*.

Haripriya, H., Amrutha, S., Veena, R., and Nedungadi, P. (2015). Integrating apriori with paired k-means for cluster fixed mixed data. In *Proceedings of the Third International Symposium on Women in Computing and Informatics*, pages 10–16.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337):123–129.

Hernández-Torruco, J., Canul-Reich, J., Frausto-Solís, J., and Méndez-Castillo, J. J. (2014). Feature selection for better identification of subtypes of guillain-barré syndrome. *Computational and mathematical methods in medicine*, 2014.

Hsu, C.-C. and Huang, Y.-P. (2008). Incremental clustering of mixed data based on distance hierarchy. *Expert Systems with Applications*, 35(3):1177–1185.

Huang, J. Z., Ng, M. K., Rong, H., and Li, Z. (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):657–668.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, pages 21–34.

Isa, N. A. M., Salamah, S. A., and Ngah, U. K. (2009). Adaptive fuzzy moving k-means clustering algorithm for image segmentation. *IEEE Transactions on Consumer Electronics*, 55(4).

Islam, M. Z. and Brankovic, L. (2011). Privacy preserving data mining: A noise addition framework using a novel clustering technique. *Knowledge-Based Systems*, 24(8):1214–1223.

Jain, A. and Dubes, R. (1988). Algorithms for clustering data, prentice-hall, inc. upper saddle river. *NJ, USA*.

Jain, A., Nandakumar, K., and Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12):2270–2285.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Jayalakshmi, T. and Santhakumaran, A. (2011). Statistical normalization and back propagationfor classification. *International Journal of Computer Theory and Engineering*, 3(1):89.

Ji, J., Pang, W., Zhou, C., Han, X., and Wang, Z. (2012). A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. *Knowledge-Based Systems*, 30:129–135.

Jiang, L., Cai, Z., Zhang, H., and Wang, D. (2013). Naive bayes text classifiers: a locally weighted learning approach. *Journal of Experimental & Theoretical Artificial Intelligence*, 25(2):273–286.

Jiang, L., Li, C., Zhang, H., and Cai, Z. (2014). A novel distance function: Frequency difference metric. *International Journal of Pattern Recognition and Artificial Intelligence*, 28(02):1451002.

Jo, B. and Baloch, Z. (2017). Internet of things-based arduino intelligent monitoring and cluster analysis of seasonal variation in physicochemical parameters of jungnangcheon, an urban stream. *Water*, 9(3):220.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344.

Khan, F. (2012). An initial seed selection algorithm for k-means clustering of georeferenced data to improve replicability of cluster assignments for mapping application. *Applied Soft Computing*, 12(11):3698–3700.

Kirchner, K., Zec, J., and Delibašić, B. (2016). Facilitating data preprocessing by a generic framework: a proposal for clustering. *Artificial Intelligence Review*, 45(3):271–297.

Kotsiantis, S. and Pintelas, P. (2004). Recent advances in clustering: A brief survey. *WSEAS Transactions on Information Science and Applications*, 1(1):73–81.

Kou, G., Peng, Y., and Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using mcdm methods. *Information Sciences*, 275:1–12.

Krishnasamy, G., Kulkarni, A. J., and Paramesran, R. (2014). A hybrid approach for data clustering based on modified cohort intelligence and k-means. *Expert Systems with Applications*, 41(13):6009–6016.

Lee, L. C., Liong, C.-Y., and Jemain, A. A. (2017). A contemporary review on data preprocessing (dp) practice strategy in atr-ftir spectrum. *Chemometrics and Intelligent Laboratory Systems*.

Liu, W., Liang, Y., Fan, J., Feng, Z., and Cai, Y. (2014). Improved hierarchical k-means clustering algorithm without iteration based on distance measurement. In *International Conference on Intelligent Information Processing*, pages 38–46.

Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 911–916.

Liu, Z. et al. (2011). A method of svm with normalization in intrusion detection. *Procedia Environmental Sciences*, 11:256–262.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Loohach, R. and Garg, K. (2012). Effect of distance functions on simple k-means clustering algorithm. *International Journal of Computer Applications*, 49(6).

Ma, M., Luo, Q., Zhou, Y., Chen, X., and Li, L. (2015). An improved animal migration optimization algorithm for clustering analysis. *Discrete Dynamics in Nature and Society*, 2015.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297.

Makarenkov, V. and Legendre, P. (2001). Optimal variable weighting for ultrametric and additive trees and k-means partitioning: Methods and software. *Journal of Classification*, 18(2):245–271.

Mashor, M. Y. (2000). Hybrid training algorithm for rbf network. *International Journal of the computer, the Internet and Management*, 8(2):50–65.

Matthews, A. (1979). Standardization of measures prior to cluster-analysis. In *Biometrics*, volume 35, pages 892–892.

Merendino, S. and Celebi, M. E. (2013). A simulated annealing clustering algorithm based on center perturbation using gaussian mutation.

Meyer, B. and Glenz, A. (2013). Team faultline measures: A computational comparison and a new approach to multiple subgroups. *Organizational Research Methods*, 16(3):393–424.

Milligan, G. W. (1989). A validation study of a variable weighting algorithm for cluster analysis. *Journal of Classification*, 6(1):53–71.

Milligan, G. W. and Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of classification*, 5(2):181–204.

Mirkin, B. (2012). *Clustering: a data recovery approach*. CRC Press.

Mogotsi, I. (2010). Christopher d. manning, prabhakar raghavan, and hhinrich schütze: Introduction to information retrieval. *Information Retrieval*, 13(2):192–195.

Mohamad, I. and Usman, D. (2013). Standardization and its effects on k-means clustering algorithm. *Res. J. Appl. Sci. Eng. Technol*, 6(17):3299–3303.

Mohd, W. M. W., Beg, A., Herawan, T., and Rabbi, K. (2012). An improved parameter less data clustering technique based on maximum distance of data and lioyd k-means algorithm. *Procedia Technology*, 1:367–371.

Mok, P.-Y., Huang, H., Kwok, Y., and Au, J. (2012). A robust adaptive clustering analysis method for automatic identification of clusters. *Pattern Recognition*, 45(8):3017–3033.

Münz, G., Li, S., and Carle, G. (2007). Traffic anomaly detection using k-means clustering.

Napoleon, D. and Pavalakodi, S. (2011). A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set. *International Journal of Computer Applications*, 13(7):41–46.

Nayak, S., Misra, B., and Behera, H. (2014). Impact of data normalization on stock index forecasting. *Int. J. Comp. Inf. Syst. Ind. Manag. Appl*, 6:357–369.

Nazeer, K. A. and Sebastian, M. (2009). Improving the accuracy and efficiency of the k-means clustering algorithm. In *Proceedings of the World Congress on Engineering*, volume 1, pages 1–3.

Noorbehbahani, F., Mousavi, S., and Mirzaei, A. (2015). An incremental mixed data clustering method using a new distance measure. *Soft Computing*, 19(3):731–743.

Ogasawara, E., Martinez, L. C., De Oliveira, D., Zimbrão, G., Pappa, G. L., and Mattoso, M. (2010). Adaptive normalization: A novel data normalization approach for non-stationary time series. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8.

Orloci, L. (1967). An agglomerative method for classification of plant communities. *The Journal of Ecology*, pages 193–206.

Pappas, T. N. (1992). An adaptive clustering algorithm for image segmentation. *IEEE Transactions on signal processing*, 40(4):901–914.

Paramasivam, M., Dasgupta, S., Ajjarapu, V., and Vaidya, U. (2015). Contingency analysis and identification of dynamic voltage control areas. *IEEE Transactions on Power Systems*, 30(6):2974–2983.

Patel, V. R. and Mehta, R. G. (2011). Impact of outlier removal and normalization approach in modified k-means clustering algorithm. *IJCSI International Journal of Computer Science Issues*, 8(5).

Pena, J. M., Lozano, J. A., and Larranaga, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. *Pattern recognition letters*, 20(10):1027–1040.

Priness, I., Maimon, O., and Ben-Gal, I. (2007). Evaluation of gene-expression clustering via mutual information distance measure. *BMC bioinformatics*, 8(1):1.

Rajput, D. S., Singh, P., and Bhattacharya, M. (2011). Feature selection with efficient initialization of clusters centers for high dimensional data clustering. In *Communication Systems and Network Technologies (CSNT), 2011 International Conference on*, pages 293–297.

Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., and Herrera, F. (2017). A survey on data preprocessing for data stream mining: current status and future directions. *Neurocomputing*, 239:39–57.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Raykov, Y. P., Boukouvalas, A., Baig, F., and Little, M. A. (2016). What to do when k-means clustering fails: a simple yet principled alternative algorithm. *PloS one*, 11(9):e0162259.

Reddy, D., Jana, P. K., and Member, I. S. (2012). Initialization for k-means clustering using voronoi diagram. *Procedia Technology*, 4:395–400.

Redfern, N. (2010). Robust measures of scale for shot length distributions.

Rokach, L. and Maimon, O. (2014). *Data mining with decision trees: theory and applications*.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283.

Rousseeuw, P. J. and Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79.

Rousseeuw, P. J. and Kaufman, L. (1990). *Finding Groups in Data*.

Ryu, T.-W. and Eick, C. F. (2005). A database clustering methodology and tool. *Information Sciences*, 171(1):29–59.

Saad, M., Ahmad, M., Abu, M., and Jusoh, M. (2014). Hamming distance method with subjective and objective weights for personnel selection. *The Scientific World Journal*, 2014.

Santhanam, T. and Padmavathi, M. (2014). Comparison of k-means clustering and statistical outliers in reducing medical datasets. In *Science Engineering and Management Research (ICSEMR), 2014 International Conference on*, pages 1–6.

Sarma, T. H., Viswanath, P., and Reddy, B. E. (2013). A hybrid approach to speed-up the k-means clustering method. *International Journal of Machine Learning and Cybernetics*, 4(2):107–117.

Sarstedt, M. and Mooi, E. (2014). A concise guide to market research. *The Process, Data, and Methods using IBM SPSS Statistics*.

Sebestyen, G. S. (1962). Decision-making processes in pattern recognition (acm monograph series).

Selim, S. Z. and Ismail, M. A. (1984). K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (1):81–87.

Shanmugavadivu, P. and Rajeswari, R. S. (2012). Identification of microcalcifications in digital mammogram using modified k-mean clustering. In *Emerging Trends in Science, Engineering and Technology (INCOSET), 2012 International Conference on*, pages 216–221.

Shi, W. and Zeng, W. (2013). Genetic k-means clustering approach for mapping human vulnerability to chemical hazards in the industrialized city: a case study of shanghai, china. *International journal of environmental research and public health*, 10(6):2578–2595.

Shirkhorshidi, A. S., Aghabozorgi, S., and Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, 10(12):e0144059.

Singh, B. K., Verma, K., and Thoke, A. (2015). Investigations on impact of feature normalization techniques on classifier's performance in breast tumor classification. *International Journal of Computer Applications*, 116(19).

Singh, D. and Leavline, E. J. (2016). Model-based outlier detection system with statistical preprocessing. *Journal of Modern Applied Statistical Methods*, 15(1):39.

Sola, J. and Sevilla, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science*, 44(3):1464–1468.

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1):1–34.

Su, C., Zhan, J., and Sakurai, K. (2009). Importance of data standardization in privacy-preserving k-means clustering. In *Database Systems for Advanced Applications*, pages 276–286.

Sundararajan, S. and Karthikeyan, S. (2014). An efficient hybrid approach for data clustering using dynamic k-means algorithm and firefly algorithm. *ARPN Journal of Engineering and Applied Sciences*, 9(8).

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). Introduction to data mining. 1st.

Tan, P.-N., Steinbach, M., and Kumar, V. (2013). Data mining cluster analysis: basic concepts and algorithms. *Introduction to data mining*.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, 18(4):267–276.

Tomar, D. and Agarwal, S. (2015). Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes. *Advances in Artificial Neural Systems*, 2015:1.

Uddin, J., Ghazali, R., and Deris, M. M. (2017). An empirical analysis of rough set categorical clustering techniques. *PloS one*, 12(1):e0164803.

Velardi, P., Navigli, R., Faralli, S., and Ruiz-Martinez, J. M. (2012). A new method for evaluating automatically learned terminological taxonomies. In *LREC*, pages 1498–1504.

Vesanto, J. (2001). Importance of individual variables in the k-means algorithm. In *Advances in Knowledge Discovery and Data Mining*, pages 513–518.

Vinod, H. D. (1969). Integer programming and the theory of grouping. *Journal of the American Statistical Association*, 64(326):506–519.

Visalakshi, K. and Thangavel, K. (2009). Impact of normalization in distributed k-means clustering. *International Journal of Soft Computing*, 4(4):168–172.

Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., et al. (2001). Constrained k-means clustering with background knowledge. In *ICML*, volume 1, pages 577–584.

Wang, H. and Dubitzky, W. (2005). A flexible and robust similarity measure based on contextual probability. In *International Joint Conference on Artificial Intelligence*, volume 19, page 27.

Wilson, D. R. and Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of artificial intelligence research*, 6:1–34.

Wu, J., Xiong, H., and Chen, J. (2009). Adapting the right measures for k-means clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 877–886.

Wu, J., Xiong, H., Chen, J., and Zhou, W. (2007). A generalization of proximity functions for k-means. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 361–370.

Xiaojun, L. (2017). An improved clustering-based collaborative filtering recommendation algorithm. *Cluster Computing*, pages 1–8.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678.

Xu, R. and Wunsch, D. (2008). *Clustering*.

Yang, J., Jiang, H., and Zhang, H. (2011). Teaching assistant evaluation based on support vector machines with parameters optimization. *Information Technology Journal*, 10(11):2140–2146.

Yang, L. and Jin, R. (2006). Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2(2).

Zhang, W., Yoshida, T., Tang, X., and Wang, Q. (2010). Text clustering using frequent itemsets. *Knowledge-Based Systems*, 23(5):379–388.

Zumel, N. and Mount, J. (2013). Log transformations for skewed and wide distributions. 18.