K-MEANS ALGORITHM VIA PREPROCESSING TECHNIQUE AND SINGULAR VALUE DECOMPOSITION FOR HIGH DIMENSION DATASETS

DAUDA USMAN

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

Faculty of Science
UniversitiTeknologi Malaysia

DECEMBER 2014

To my beloved mother and father

ACKNOWLEDGEMENT

I would like to express my greatest appreciation to my advisor, Assoc. Prof. Dr. Ismail Bin Mohamad, for introducing me to the world of clustering and provided me with a steady stream of his insights into data visualization and manipulation. I feel blessed to have him as my advisor. He helped me not only accomplish my dream of becoming a professional statistician but also develop a more mature personality. I thank him for every piece of his intensive efforts that have been put into this research work.

I would also like to thank the management team of Umaru Musa Yar'adua University, Katsina- Nigeria for giving me the opportunity to further my study.

I owe many thanks to my collaborators for this and other material: Aishah Bint Mohd Noor Universiti Malaysia Perlis, Malaysia and Abdulrahman from Syria. There is much that never would have been discovered without their insight and their diligence. I would also like to thank my colleagues at UTM; it was their presence that made UTM the great research environment it was.

Last but not least, I am deeply grateful to my family for their patience, love, motivation and encouragement from the beginning of this long story.

ABSTRACT

Data clustering is an unsupervised classification method aimed at creating groups of objects, or clusters that are distinct. Among the clustering techniques, *K*-means is the most widely used technique. Two issues are prominent in creating a *K*-means clustering algorithm; the optimal number of clusters and the center of the clusters. In most cases, the number of clusters is pre-determined by the researcher, thus leaving out the challenge of determining the cluster centers so that scattered points can be grouped properly. However, if the cluster centers are not chosen correctly computational complexity is expected to increase, especially for high dimensional data set. In order to obtain an optimum solution for *K*-means cluster analysis, the data needs to be pre-processed. This is achieved by either data standardization or using principal component analysis on rescaled data to reduce the dimensionality of the data. Based on the outcomes of the preprocessing carried out on the data, a hybrid *K*-means clustering method of center initialization is developed for producing optimum quality clusters which makes the algorithm more efficient. This research investigates and analyzes the performance behavior of the basic *K*-means clustering algorithm when three different standardization methods are used, namely decimal scaling, *z*-score and min-max. The results show that, *z*-score perform the best, judging from the sum of square error. Further experiments on the hybrid algorithm are conducted using uncorrelated and correlated simulated data sets having low, moderate and high dimension and it is observed that the method presented in this thesis gives a good and promising performance. It is also observed that, the sum of the total clustering errors reduced significantly whereas inter-distances between clusters are preserved to be as large as possible for better clusters identification. The results and findings are validated using life data on infectious diseases.

ABSTRAK

Pengkelompokan data adalah kaedah pengkelompokan tak terselia yang bertujuan membentuk kumpulan objek atau kluster yang berbeza. Dalam banyak kaedah pengkelompokan, kaedah *K-means* adalah kaedah yang paling kerap digunakan. Dua isu utama dalam membentuk algoritma *K-means* adalah penentuan bilangan kluster yang optimum dan pusat kluster. Dalam kebanyakan kes, bilangan kluster telah ditentukan terlebih dahulu oleh pengkaji, dan cabaran seterusnya ialah menentukan kedudukan pusat kluster supaya titik data dapat dikluster dengan sempurna. Jika pusat kluster tidak dipilih dengan betul ia akan meningkatkan kerumitan pengiraan terutama bagi data berdimensi tinggi. Bagi memperoleh penyelesaian optimum *K-means*, data perlu diproses terlebih dahulu. Matlamat ini boleh dicapai dengan piawaian data atau menggunakan analisis komponen prinsipal terhadap data yang diskala semula bagi mengurangkan dimensi data. Kaedah hibrid *K-means* ini seterusnya digunakan terhadap data terturun yang menghasilkan kluster optimum berkualiti yang membuatkan algoritma ini lebih efisen. Kajian ini meninjau dan menganalisis keupayaan algoritma pengkelompokan apabila tiga kaedah piawai iaitu *K-means* kaedah *decimal scaling*, *z-score* dan *min-max* digunakan. Keputusan menunjukkan, *z-score* adalah yang terbaik berdasarkan kepada jumlah ralat kuasa dua. Kajian lanjut mengenai algoritma hibrid ini terhadap data berkorelasi dan tidak berkorelasi dalam keadaan dimensi rendah, sederhana dan tinggi menunjukkan bahawa kaedah yang dibentangkan dalam tesis ini mempunyai pencapaian yang memuaskan. Keputusan ini juga mendapati jumlah ralat kuasa dua dikurangkan dan jarak antara satu kluster ke kluster lain dijadikan semaksimum mungkin yang memisahkan kluster dengan jelas. Keputusan dan dapatan ini ditentusahkan dengan menggunakan data penyakit berjangkit.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	DECLARATION	ii
	DEDICATION	iii
	ACKNOWLEDGEME	iv
	ABSTRACT	v
	ABSTRAK	vi
	TABLE OF CONTENTS	vii
	LIST OF TABLES	xi
	LIST OF FIGURES	xiii
	LIST OF ABBREVIATIONS	xv
	LIST OF APPENDICES	xvi
1	**INTRODUCTION**	**1**
	1.1 Background of the Study	1
	1.2 Problem Statement	3
	1.3 Objectives of the Study	4
	1.4 Scope of the Study	5
	1.5 Contribution of the Study	6
	1.6 Thesis Organization	6
	1.7 Chapter Summary	7
2	**LITERATURE REVIEW**	**8**
	2.1 Introduction	8
	2.2 Cluster Analysis	9
	2.3 The Clustering Process	9
	2.4 Similarity Measures	11

2.4.1 Euclidean Distance 13
2.4.2 Manhattan Distance 13
2.4.3 L_∞ *norm* 14
2.4.4 Mahalanobis Distance 14
2.5 *K*-means Cluster Analysis 15
2.6 Principal Component Analysis (PCA) 17
2.7 Cluster Initialization Technique 18
2.8 Data Standardization 20
2.9 Detecting Outlier and Cleaning Data 21
2.10 Cluster Validity 22
2.11 Fundamental Concepts of Cluster Validity 23
2.12 Some Related Work on *K*-means 24
2.13 Rerearch Framework 29
2.14 Chapter Summary 31

3 METHODOLOGY 32
3.1 Introduction 32
3.2 Pre-processing Method 32
3.2.1 Data Standardization and Transformation 33
3.2.1.1 Z-score Standardization 34
3.2.1.2 Min-Max Standardization 34
3.2.1.3 Decimal Scaling Standardization 35
3.2.2 Principal Component Analysis 35
3.3 Singular Value Decomposition 38
3.3.1 Components To be Retained 40
3.4 *K*-means Cluster Analysis 41
3.5 The Proposed Hybrid *K*-means Method 42
3.6 How To Obtain Initial Cluster Centers 48
3.7 Chapter Summary 49

4 METHOD IMPLEMENTATION 50
4.1 Introduction 50
4.2 Data Pre-processing 50
4.2.1 Data Standardization 52

4.2.2 *K*-means with Preprocessed Data 59
4.2.3 Discussion 63
4.3 Chapter Summary 64

5 EVALUATION OF THE HYBRID *K*-MEANS METHOD 65
5.1 Introduction 65
5.2 Covariance Matrix Structure 65
5.3 Simulation Experiment with Uncorrelated Dataset 67
5.4 Simulation Experiment with Correlated Dataset 70
5.4.1 Simulation from generated dataset where correlation is controlled 70
5.4.2 Simulation from generated data set with fixed correlation 78
5.4.3 Discussion 81
5.5 Separable Cases 81
5.5.1 Discussion 85
5.6 Comparative Analysis for the Basic *K*-means and Hybrid *K*-means 85
5.6.1 Discussion 91
5.7 Z-test for $SSE_D < SSE_S$ 91
5.7.1 Discussion 96
5.8 An Example to Illustrate the Hybrid *K*-means Method 97
5.9 Chapter Summary 104

6 CASE STUDY 105
6.1 Introduction 105
6.2 Pattern of Infection Across Countries 106
6.3 Cluster Formations 106
6.4 Discussion 111
6.5 Recommendation 112

7 CONCLUSION 113
7.1 Introduction 113
7.2 Conclusion 113

7.3 Achievement 114

7.4 Further Research 115

REFERENCES 117

APPENDICES A - B 125-137

LIST OF TABLES

TABLE NO.	TITLE	PAGE
4.1	The Rafindadi Clinic data	51
4.2	The Standardized z-score for Rafindadi Clinic data	54
4.3	The Standardized decimal scaling for Rafindadi Clinic data	56
4.4	The Standardized min-max for Rafindadi Clinic data	58
4.5	The PCs variances and cumulative percentage	60
4.6	The reduced principal components of the Rafindadi Clinic data	61
4.7	Projected data	62
4.8	Summary of cluster formation results	63
5.1	The total sum of squares error and time taken for uncorrelated datasets	70
5.2	The total sum of squares errors and time taken for computed correlated dataset	78
5.3	The total sum of squares errors and time taken for correlated datasets	80
5.4	Error sum of squares for the separable non separable, half Separable and Separable	85
5.5	The number of time and proportions of $SSE_D < SSE_S$ for $k = 2$ from 500 simulations (The numbers in brackets are the proportions and * symbols indicates the proportion is significantly large than 50% at $\alpha = 0.05$)	93
5.6	The number of time and proportions of $SSE_D < SSE_S$ for $k = 3$ from 500 simulations (The numbers in brackets are the proportions and * symbols indicates the proportion is significantly large than 50% at $\alpha = 0.05$)	93
5.7	The number of time and proportions of $SSE_D < SSE_S$ for $k = 5$ from 500 simulations (The numbers in brackets are the proportions and * symbols indicates the proportion is significantly large than 50% at $\alpha = 0.05$)	93

5.8	The number of time and proportions of $SSE_D < SSE_S$ for $k = 7$ from 500 simulations (The numbers in brackets are the proportions and * symbols indicates the proportion is significantly large than 50% at $\alpha = 0.05$)	94
5.9	The original sample data	97
5.10	The Standardized z-score for the sample data	99
5.11	The The PCs variances and cumulative percentage of the standardized sample data	99
5.12	The principal components of the standardized sample data	100
5.13	The reduced principal components of the standardized sample data	100
5.14	Reduced projected data	101
5.15	Cases number and clusters	102
6.1	Cluster cases number and their distances of KGHK data	109
6.2	Cluster cases number and their distances of FMCK data	109

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
2.1	Clustering Process	10
2.2	Research Framework	31
4.1	Basic *K*-means centroids	52
4.2	*Z*-score basic K-means centroids	55
4.3	Decimal scaling basic *K*-means centroids	57
4.4	Min-Max basic *K*-means centroids	59
4.5	PCA/SVD basic *K*-means centroids	62
5.1	Basic *K*-means centroids for uncorrelated data (5, 500)	68
5.2	Hybrid *K*-means centroids for uncorrelated data (5, 500)	68
5.3	Basic *K*-means centroids for uncorrelated data (20, 500)	69
5.4	Hybrid *K*-means centroids for uncorrelated data (20, 500)	69
5.5	Basic *K*-means centroids V_1 (5, 50)	72
5.6	Hybrid *K*-means centroids V_1 (5, 50)	72
5.7	Basic *K*-means centroids V_2 (5, 50)	73
5.8	Hybrid *K*-means centroids V_2 (5, 50)	74
5.9	Basic *K*-means centroids V_3 (5, 50)	75
5.10	Hybrid *K*-means centroids V_3 (5, 50)	76
5.11	Basic *K*-means centroids V_4 (5, 50)	77
5.12	Hybrid *K*-means centroids V_4 (5, 50)	77
5.13	Basic *K*-means centroids for correlated data (5, 500)	79
5.14	Hybrid *K*-means centroids for correlated data (5, 500)	79
5.15	Basic *K*-means centroids for correlated data (20, 500)	79
5.16	Hybrid *K*-means centroids for correlated data (20, 500)	80
5.17	Basic *K*-means centroids in the not separable case	82
5.18	Hybrid *K*-means centroids in the not separable case	82

5.19	Basic *K*-means centroids in the half separable case	83
5.20	Hybrid *K*-means centroids in the half separable case	83
5.21	Basic *K*-means centroids in the separable case	84
5.22	Hybrid *K*-means centroids in the separable case	84
5.23	Clustering results for (5, 500) and *K*=2	87
5.24	Clustering results for (10, 500) and k=3	88
5.25	Clustering results for (20, 500) and *K*=5	89
5.26	Clustering results for (50, 500) and *K*=7	90
5.27	Basic *K*-means method	103
5.28	Hybrid *K*-means method	103
6.1	Classification of KGHK data into two clusters	110
6.2	Classification of FMCK data into two clusters	110

LIST OF ABBREVIATIONS

CPU	-	Central Processsor Unit
DVSCAN	-	Density-Based Clustering Algorithm
FE	-	Feature Extraction
FMCK	-	Federal Medical Centre, Katsina
FR	-	Feature Reduction
GA	-	Genetic Algorithm
GAIK	-	Genetic Algorithm Initialize K-means
GCA	-	Genetic-Baseb Clustering Algorithm
IGK	-	Improved Genetic K-means
KGHK	-	Katsina General Hospital, Katsina
KIGA	-	K-means Initializes the Genetic Algorithm
LOF	-	Local Outlier Factor
PC	-	Principal Component
PCA	-	Principal Component Analysis
PD	-	Partial Distance
SCP	-	Spectral Constaraint Prototype
SSE	-	Sum of Squares Error
SVD	-	Singular Value Decomposition

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Publications	125
B	Data Collected	127
C	Matlab Programmes	131

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The *K*-means clustering algorithm is one of the most popular methods for clustering multivariate observations (Tsai and Chiu, 2008). It is a system ordinarily used to directly segment sets of data into *k* groups. *K*-means algorithm generates a fast and efficient solution. The basic *K*-means algorithm works with the objective to minimize the mean square distance from each data point to its nearest center.

There are two important issues in creating a *K*-means clustering algorithm: the optimal number of clusters and the center of the cluster. In many cases, the number of clusters is given, thus the important issue is where to put the cluster center so that scattered points can be grouped appropriately. Center of the cluster can be obtained by first assigning any random point and then optimizing the mean distance to the center. The process is repeated until all the mean square distances are optimized.

The drawback of the basic *K*-means algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the criterion function value if the initial centroids are not properly chosen. A local minimum is the least value that is located within a set of points which may or may not be a global minimum and it is not the lowest value in the entire set. Its computational complexity is also very high, especially for large data set. In addition the number of distance

calculations increases exponentially with the increase of the dimensionality of the data. An ad-hoc solution to these problems is by choosing a set of different initial partition and the initial partition that gives the smallest sum of squares error is taken as the solution but this ad-hoc solution does not guarantee the solution will give the smallest sum of square error (SSE) because it is just a mere guessing approach.

When a random initialization of centroids is used, different runs of *K*-means typically produce different total SSEs, therefore choosing the proper initial centroids is the key step of the basic *K*-means procedure (Zhu *et al.* 2009). The result of the *K*-means algorithm is highly dependent upon its initial selection of cluster centers and before clustering it must be previously known and fixed (Tsai and Chiu, 2008). Fahim *et al*. (2009) proposed a method to select a good initial solution by partitioning data set into blocks and applying *K*-means to each block. But here the time complexity is slightly more.

Tajunisha and Saravanan (2010) proposed a method to improve the performance of the *K*-means algorithm, using principal component analysis (PCA) for dimension reduction and to find the initial centroid for *K*-means. The method partitioned the data set into *K* sets and the median of each set were used as initial cluster centers and then assign each data point to its nearest cluster centroid. Heuristic approach was also used to reduce the number of distance calculation in the standard *K*-means algorithm to assign the data point to the cluster.

Mohammed and Wesam (2012) proposed a visual clustering framework using C++ Builder 2009, for initialization of the *K*-means clustering algorithm. The method generates *K* points using semi random technique. It makes the diagonal of the data as a starting line and selects the points randomly around it. But this method did not suggest any improvement to the time complexity for *K*-means algorithm. The above algorithms are quite complex and used the *K*-means algorithm as part of their algorithm, which still need to use the random method for cluster center initialization.

To obtain an optimum solution for *K*-means clustering, the data need to be pre-processed before the *K*-means clustering analysis (Chandrasekhar *et al*. 2011). This pre-processing process consists of data standardization method to rescale the dataset and principal component analysis method for outliers' detection. Outliers are the data that are numerically distant from the rest of the data. If they are not properly detected and handled, the clustering result will be affected in a great manner (Sairam *et al*. 2011).

An approach to handle outlier is data standardization, it rescale the data set to fall within a specified range of values so that any attribute with larger value will not dominate the attribute with a smaller value. However, for a very high dimensional data set, PCA can be used to initially reduce this dimension (Chris and Xiaofeng, 2006). They further proved that principal components are the continuous solutions to the discrete cluster membership indicators for *K*-means clustering and showed that unsupervised dimension reduction is closely related to unsupervised learning. On dimension reduction, the result provides new insights to the observed effectiveness of PCA-based data reductions, beyond the conventional noise-reduction explanation.

As *K*-means is highly dependent on its initial center position (Rana *et al*. 2010), an alternative way of center initialization method for *K*-means cluster analysis is also required to make the algorithm more effective and efficient. To overcome the above drawback the current research focused on developing a *K*-means clustering technique by data preprocessing and the initialization of the center points for high dimensional datasets.

1.2 Problem Statement

As mentioned in the research background, *K*-means clustering has shortcomings especially when implemented on large dataset. This is best seen in that the basic *K*-means algorithm for cluster analysis developed for low dimensional data, often do not work well for high dimensional data and most of the times the results

may not be accurate due to noise and outliers associated with the initial dataset. This brings about an increase in computational complexity and also resulting in some attributes with larger domain dominating those attributes with lower domain. Outliers are the data that are numerically distant from the rest of the data and if they are not detected and handled, they tend to affect the clustering result.

In the creation of a *K*-means clustering algorithm two main issues are prominent, these are: the optimal number of clusters and the cluster center points. In most cases, the number of clusters is given, thus leaving the challenge where to put the cluster centers so that scattered points can be grouped properly and to avoid its convergence to a local minimum of the objective function. Furthermore, the random initialization results in different total SSEs value from several runs of the *K*-means. This makes the result from the algorithm of the *K*-means to depend greatly on the initial selection of the cluster centers which must be known and fixed beforehand. Therefore, the choice of proper initial centroids is pertinent to the basic *K*-means procedure.

As fallout from the above, a new technique of centers initialization for *K*-means clustering is required to make the algorithm more effective and efficient. Hence this research focus on an alternative way of handling the *K*-means clustering technique by data pre-processing and the initialization of the center points for high dimensional datasets.

1.3 Objectives of the Study

The objectives of this study consist of three parts, the computational (data pre-processing), centre points initialization, and practical aspects. The main objective of the first aspect is to come up with a suitable data preprocessing method for *K*-means cluster analysis that is able to obtain good clustering with reduced complexity, and also provides better accuracy.

The second part consists of the followings:

i. Centers initialization using singular value decomposition (SVD) to avoid random initialization and convergence to a local minimum, this will make the algorithm more effective and efficient.

ii. Simulation experiment and comparison of the optimality for the basic and the proposed techniques.

The third part is the application of the technique to Nigerian data of infectious diseases to demonstrate the use of this technique.

1.4 Scope of the Study

This research covers the following three aspects, computational (data pre-processing), initialization of centre points, and practical aspects. Furthermore, the number of clusters required is always pre-determined throughout the thesis.

i. Computational aspect (Data pre-processing)
   In computational aspects, we used a real data of infectious diseases consisting of seven variables and a sample size of 20 to come up with a good method of data pre-processing for *K*-means cluster analysis and a simulation experiments to validate and test the proposed hybrid *K*-means technique.

ii. Centre point initialization aspect
   In order to have a better understanding about *K*-means, we give an overview of clustering. Then we show its drawback. This motivates us to propose a technique of center point initialization. Afterwards, we come up with a hybrid *K*-means algorithm. We equally investigate the CPU time taken and the SSE for the basic and new techniques.

iii. Practical aspect

Application in real data of infectious diseases from hospital was used to show the performance and advantage of the hybrid *K*-means method developed in this thesis.

1.5 Contribution of the Study

This thesis offers a contribution in two aspects:

1. The main contributions are

i. New method in choosing the initial centroids for *K*-mean cluster analysis.

ii. A technique of *K*-means clustering algorithm for high dimension dataset. Simulation experiments proved that the proposed technique is optimum and converges faster than the basic method.

iii. The computational complexity of the proposed *K*-means clustering technique is far lower than the basic *K*-means algorithm. This is another advantage of the proposed method especially when the data sets are of higher dimensions.

2. This thesis used infectious diseases data sets. Therefore we hope that the finding results will be useful and assist the government to embark on health policy formulations that will address the problems of the intensity of diseases in the states and to plan adequately for the provision of welfare services such as good pipe borne water, good drainage system and environmental protection.

1.6 Thesis Organization

This thesis is organized into six chapters. Chapter 1 briefly overviews the two issues in creating a *K*-means clustering algorithm: The optimal number of

clusters and the center of the cluster. The drawback of the basic *K*-means algorithm is also stated, the research objectives are defined, and the scope of this research is also presented. However, the chapter presents the significance of this study and ends with the contribution of the study.

Chapter 2 presents a comprehensive literature review. The target is to review current literatures to identify what has previously been attained and recognize the gap in our research. The existing theories of the *K*-means clustering technique, similarity measures, principal component analysis, data standardization methods and the evolution of ideas in the initialization of the *K*-means cluster centers are presented, this ends with a research framework.

Chapter 3 focuses on an in depth explanation of the methodologies. It concentrates on how this research is carried out in order to arrive at the findings and conclusions. This chapter describes the data pre-processing methods and how to implement the proposed *K*-means technique.

Chapter 4 concentrated on the data pre-processing methods, implementation of the proposed hybrid *K*-means clustering method and validation of the method using simulation experiment with correlated and uncorrelated in Chapter 5. After that, Chapter 6 demonstrates the use of this method to Nigerian data of infectious diseases. This thesis closes with the conclusion and recommendation for future research in Chapter 7.

1.7 Chapter Summary

This chapter overviews the drawback of a basic *K*-means algorithm and addresses the two issues when creating a *K*-means clustering, that is, the optimal number of clusters and the center of the cluster. We also highlight the problem statement, research objectives, the research scope and significance of the research. The Chapter ends with the thesis organization.

REFERENCES

Agusti, L. E. B., Salcedo, S. S., Jimenez, S. F., Carro, L. C., Del, S. J. and Portilla, J. A. F (2012). A new Grouping Genetic Algorithm for Clustering Problems. *Expert Systems with Applications Elsevier.* 39:9695-9703.

Al Hasan, M., Chaoji, V., Salem, S. and Zaki, M. (2009). Robust Partitional Clustering by Outlier and Density Insensitive Seeding. *Pattern Recognition Letters*. 30(11):994‑1002.

Alshalabi, L., Shaaban, Z. and Kasasbeh, B. (2006). Data Mining: A Preprocessing Engine. *Journal of Computer Science*. 2(9):735-739

Al-Shboul, B. and Myaeng, S. (2009). Initializing *K*-means Using Genetic Algorithms. *World Academy of Science, Engineering and Technology*, 54:114-118.

Anderberg, M. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Andrews, H. and Patterson, C. (1976). Singular Value Decompositions and Digital Image Processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 24:26–53.

Angel, A. and Sathurappan, N. (2012). *K*-means Clustering with Careful Seeding for Large Cluster Number. *International Journal of Communications and Engineering*, 2(4): 71-74.

Arai, K. and Barakbah, A. R. (2007). Hierarchical *K*-means: An Algorithm for Centroids Initialization for *K*-means. *Reports of the Faculty of Science and Engineering*, Saga University, 36(1): 25-31.

Behera, H. S., Ghosh, A. S. (2012). A New Improved Hybridized *K*-means Clustering Algorithm with Improved PCA Optimized with PSO for High Dimensional Data Set. *International Journal of Soft Computing and Engineering.* 2(2):121-126.

Belal, A. M., Hudaib, M., Huneiti., A. and Hammo, B. (2008). New Efficient Strategy to Accelerate *K*-means Clustering Algorithm. *American Journal of Applied Sciences*, 5(9): 1247-1250.

Berry, M. J. A. and Linoff, G. S. (1997). *Data Mining Techniques for Marketing, Sales and Customer Support*. John Wiley & Sons, Inc., New York.

Besset, D. H. (2001). *Object-Oriented Implementation of Numerical Methods: An Introduction with Java and Smalltalk*, California: Morgan Kaufmann.

Bock, R. K. and Krischer, W. (1998). *The Data Analysis Briefbook*, New York: Springer-Verlag.

Chandrasekhar, T., Thangavel, K. and Elayaraja, E. (2011). Effective Clustering Algorithms for Gene Expression Data. *International Journal of Computer Applications*, 32(4): 25-29.

Chen, M. S., Han, J. and Yu, P. S. (1996). Data Mining: an Overview from a Database Perspective. *IEEE Trans. On Knowledge And Data Engineerin*. 8:866–883.

Chris, D. and Xiaofeng, H. (2006). *K*-means Clustering Via Principal Component Analysis. *Proc. of the 21st International Conference on Machine Learning.* Banff, Canada.

Cios, K. J., Pedrycz, W., Swiniarski, R. W. & Kurgan, L. A. (2007). *Data Mining: A Knowledge Discovery Approach*. New York, NY: Springer.

Clatworthy, J., Buick, D., Hankins, M., Weinman, J., and Horne, R. (2005). The use and reporting of cluster analysis in health psychology: A review. *British Journal of Health Psychology*, 10(3):329–358.

Davidson, S. (2006). *Principle and Practice of Medicine, 20th Edition.* Edinburg, New York.

Davy, M. and Sarturnino, L. (2007). Dimensionality Reduction for Active Learning with Nearest Neighbour Classifier in Text Categorization Problems, *Sixth International Conference on Machine Learning and Applications*, pp. 292-297.

Dhaliwal, P., Bhatia, M. P. S and Bansal, P. (2010). A Cluster-Based Approach for Outlier Detection in Dynamic Data Streams (KORM: *K*-median Outlier Miner). *Journal of Computing*, 2(2):74-80.

Ding, C. and He, X. (2004). *K*-means clustering via Principal Component Analysis. In *Twenty-first international conference on machine learning*. New York: ACM Press.

Doreswamy and Hemanth, K. S. (2012). A Novel Design Specification Distance (DSD) Based K-Mean Clustering Performance Evaluation on Engineering Materials' Database. *International Journal of Computer Applications*, 55(15): 26-33.

Durak, B. (2011). A Classification Algorithm Using Mahalanobis Distance Clustering of Data with Applications on Biomedical Data Sets. Master of Science, Middle East Technical University.

Ertoz, L., Steinbach, M., Kumar, V. (2003). Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data, *Proceedings of the Third SIAM International Conference on Data Mining*, San Francisco, 3:47-58.

Eyob, E. (2009). Social Implications of Data Mining and Information Privacy: *Interdisciplinary Frameworks and Solutions*, Pennsylvania: Idea Group Inc.

Fahim, A. M., Salem, A. M., Torkey, F. A., Saake, G. and Ramadan, M. A., (2009). An Efficient K-means with Good Initial Starting Points, *Georgian Electronic Scientific Journal: Computer Science and Telecommunications*. 2(19):47-57.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. Eds. (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI Press.

Fayyad, U. M., Reina, C., and Bradley, J. (1998). Initialization of Iterative Refinement Clustering Algorithms. *In proceedings of Fourth Int. Conf. On Knowledge Discovery and Data Mining*, AAAI, pp.194-198.

Gervini, D. and. Rousson, V. (2004). Criteria for Evaluating Dimension-Reducing Components for Multivariate Data. *American Statistician*, 58(1):72–76.

Guha, S., Rastoga, R. and Shim, K. (1998). An Efficient Clustering Algorithm for Large Databases. *In Proceedings of the ACM SIGMOND International Conference on Management Data*. ACM New York. pp 73-84.

Guojun, G., Chaoqun, M. and Jianhong, W. (2007). *Data Clustering: Theory, Algorithms and Applications*. ASA-SIAM Series on Statistics and Applied Probability.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17(2):107-145.

Hans-Joachim Mucha, Hans-Georg Bartel and Jens Dolata (2008). Effects of Data Transformation on Cluster Analysis of Archaeometric Data. *Proc. of the 31st Annual Conference of the Gesellschaft für Klassifikation*., Albert-Ludwigs-Universität Freiburg, pp 681-688.

Hartigan, J. A. (1972). Direct Clustering of a Data Matrix. *Journal of the American Statistical Association*, 67(337):123-129.

Hill, T., Lewicki, P. (2006). *Statistics Methods and Applications, A Comprehensive Reference for Science*, Tulsa: StatSoft Inc.

Informatics lecture note (2005). Accessed on 13th July 2014, *http://www.informatics.indiana.edu/predrag/classes/2005springi400/lecture_notes_4_1.pdf*

Jain, A. and Dubes, R. (1988). *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice–Hall.

Jain A. R., Murthy, M. N. and Flynn P. J. (1999). Data clustering: A Review. *ACM Computing Surveys*, 31(3):265-323.

Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis.* Prentice Hall.

Jolliffe, I. (2002). *Principal Component Analysis*, 2nd edition. Springer Series in Statistics. New York: Springer-Verlag.

Kanth, K., Agrawal, D. and Singh, A. (1998). Dimensionality reduction for similarity searching in dynamic databases. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pp. 166–176. New York: ACM Press.

Karthikeyani, V. N. and Thangavel, K. (2009a). Distributed Data Clustering: A Comparative Analysis. *Foundations of Computational Intelligence*. 6:371-397.

Karthikeyani, V. N. and Thangavel, K. (2009b). Impact of Normalization in Distributed K-means Clustering. *International Journal of Soft Computing* 4(4):168-172.

Kazeem, O. O., Rachid, O. and Marcus, N. (2013). Optimal Control Strategies and Cost-Effectiveness Analysis of a Malaria Model. *Biosystems*, 111: 83-101.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*, New Jersey: John Wiley and Sons.

Liu, C., Xie, J., Ge, Y. and Xiong, H. (2012). Stochastic Unsupervised Learning on Unlabeled Data. *JMLR: Workshop and Conference Proceedings on Unsupervised and Transfer Learning*, 27:111-122.

Lu, J. F., Tang, J. B., Tang, Z. M., and Yang, J. Y. (2008). Hierarchical Initialization Approach for *K*-means Clustering. *Pattern Recognition Letters*. 29(6):787‑795.

Maaten, L. J. P., Postma, E. O. and Herik, H. J. (2007). Dimensionality Reduction: A Comparative Review", *Tech. Rep. University of Maastricht*.

MacQueen, J. (1967). Some Methods For Classification and Analysis of Multivariate Observations. *In proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, pp. 281-297.

Manpreet, K. and Usvir, K. (2013). A Survey on Clustering Principles with *K*-means Clustering Algorithm Using Different Methods in Detail. *International Journal of Computer Science and Mobile Computing (IJCSMC)*, 2(5):327-331.

Marghny, M. H., Rasha M. A. and Ahmed, I. T. (2011). An Effective Evolutionary Clustering Algorithm: Hepatitis C Case Study. *International Journal of Computer Applications.* 34(6):01-06

Mason, R. L., Chaou, Y. M. and Young, J. C. (2009). Monitoring Variation in a Multivariate Process when the Dimension is Large Relative to the Sample Size, *Communication in Statistics. Theory and Methods*, 36:(6), 939-951.

McLachlan, G. J., (1999). Mahalanobis Distance. *Resonance*, 4(6): 20-26, Article accessed on 18th February 2013, http://link.springer.com/article/10.1007%2FBF02834632

Milligan, G. and Cooper, M. (1988). A Study of Standardization of Variables in Cluster Analysis. *Journal of Classification*. 5:181–204.

Mohammed, E. and Wesam, M. A. (2012). Efficient and Fast Initialization Algorithm for *K*-means Clustering. *International Journal of Intelligent Systems and Applications.* 1:21-31.

Nazeer, K. A. and Sebastian, M. P. (2009). Improving the Accuracy and Efficiency of the *K*-means Clustering Algorithm, *Proceedings of the World Congress on Engineering*, 1:308-312.

Oracle ® Database. (2004). Singular Value Decomposition. *Online Documentation Library Master*. Index: 12c Release 1 (12.1).

Patel, V. R. and Mehta, R. G. (2011). Impact of Outlier Removal and Normalization Approach in Modified *K*-means Clustering Algorithm. *International Journal of Computer Science* (IJCSI), 8(5):331-336.

Rana, S., Jasola, S. and Kumar, R. (2010). A hybrid Sequential Approach for Data Clustering Using K-means and Particle Swarm Optimization Algorithm. *International Journal of Engineering, Science and Technology*. 2(6):167-176.

Rencher, A. C. (2002). Method of Multivariate Analysis Second Edition. Wiley Series in Probability and Statistics. John Willey & Sons, Inc. New York.

Rezaee, M. R., Lelieveldt, B. P. F. and Reiber, J. H. C. (1998). A New Cluster Validity Index for the Fuzzy C-mean. *PRL*, 19(3-4): 237-246.

Sairam, N., Manikandan, G. and Sowndarya, S. (2011). Performance Analysis of Clustering Algorithms in Detecting Outliers. *International Journal of Computer Science and Information Technologies*. 2:486-488.

Salleh, R. M. (2013). Robust Estimation Method of Location and Scale with Application in Monitoring Process Variability. Doctor Philosophy, Universiti Teknologi Malaysia, Skudai. Johor Bahru, Malaysia.

Su, T., and Dy, J. (2004). A Deterministic Method for Initializing *K*-means Clustering. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference*. pp. 784-786.

Sujatha, S. and Shanthi, A. S. (2012). Novel Initialization Technique for *K*-means Clustering Using Spectral Constraint Prototype. *Journal of Global Research in Computer Science.* 3(6):46-50.

Tajunisha, N. and Saravanan, V. (2010). Performance analysis of *K*-means with Different Initialization Methods. *International Journal of Artificial Intelligence & Applications (IJAIA),* 1(4):44-52.

Telgarsky, M. and Vattani, A. (2010). Hartigan's Method: *K*-means Clustering Without Voronoi. *Journal of Machine Learning Research Proceedings Track,* 9:820 827.

Theodoridis, S. and Koutroumbas, K. (1998). *Pattern Recognition*. Academic Press, San Diego.

Tsai, C. Y., and Chiu, C. C. (2008). Developing a Feature Weight Self-Adjustment Mechanism for a *K*-means Clustering Algorithm. *Computational Statistics and Data Analysis*, 52:4658-4672.

Valarmathie, P., Srinath, M. and Dinakaran, K. (2009). An Increased Performance of Clustering High Dimensional Data Through Dimensionality Reduction Technique, *Journal of Theoretical and Applied Information Technology*. 13:271-273.

Werner, M. (2003). Identification of Multivariate Outliers in Large Data Sets. Doctor Philosophy, University of Colorado, Denver.

World Health Organization Fact Sheet (2009). Updated March, 2014 Fact Sheet No 94. Accessed on 11th April 2014, http://www.who.int/mediacentre/factsheets/fs094/en

Yan, J., Zhang, B., Liu, N., Yan, S., Cheng, Q., Fan, W., Yang, Q., Xi, W. and Chen, Z. (2006). Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing, *IEEE Transactions on Knowledge and Data Engineering*, 18(3):320-333.

Zhao, Y., Wang, E., Liu, H., Rotunno, M., Koshiol, J., Marincola, F. M., Teresa, M. L. and McShane, M. L. (2010) Evaluation of Normalization Methods for two Channel MicroRNA Microarrays. *Journal of Translational Medicine* 8:62-69.

Zhao, Z. and Liu, H. (2007). Spectral Feature Selection for Supervised and Unsupervised Learning. *In Proceedings of the 24th International Conference on Machine Learning*, Corvallis, pp 1151–1157.

Zhu, Y., Yu, J. and Jia, C. Initializing. (2009). *K*-means Clustering Using Affinity Propagation. *Ninth International Conference on Hybrid Intelligent Systems.* 1:338-343.