IMPROVED K-MEANS CLUSTERING USING PRINCIPAL COMPONENT ANALYSIS AND IMPUTATION METHODS FOR BREAST CANCER DATASET

ROSLAN ARMINA

A thesis is submitted in fulfillment of the
requirements for the award of the degree of
Master of Philosophy

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

AUGUST 2018

This thesis is special dedicated to my lovely family for their endless love, support and encouragement.

ACKNOWLEDGEMENT

Alhamdulillah, all praise to ALLAH S.W.T, the Almighty, most Gracious and most Merciful for the blessing and guidance.

Here, I would like to express a heartfelt gratitude to my supervisors, **Associate Prof Dr. Azlan Mohd Zain** for their guidance, generous support, endless advice and enormous patience throughout my research work.

Finally, I would like to thank to all those who have contributed directly and indirectly in process to finish my research work. My sincere appreciation also extends to all my colleagues for their help and support.

ABSTRACT

Data mining techniques have been used to analyse pattern from data sets in order to derive useful information. Classification of data sets into clusters is one of the essential process for data manipulation. One of the most popular and efficient clustering methods is K-means method. However, the K-means clustering method has some difficulties in the analysis of high dimension data sets with the presence of missing values. Moreover, previous studies showed that high dimensionality of the feature in data set presented poses different problems for K-means clustering. For missing value problem, imputation method is needed to minimise the effect of incomplete high dimensional data sets in K-means clustering process. This research studies the effect of imputation algorithm and dimensionality reduction techniques on the performance of K-means clustering. Three imputation methods are implemented for the missing value estimation which are K-nearest neighbours (KNN), Least Local Square (LLS), and Bayesian Principle Component Analysis (BPCA). Principal Component Analysis (PCA) is a dimension reduction method that has a dimensional reduction capability by removing the unnecessary attribute of high dimensional data sets. Hence, PCA hybrid with K-means (PCA K-means) is proposed to give a better clustering result. The experimental process was performed by using Wisconsin Breast Cancer. By using LLS imputation method, the proposed hybrid PCA K-means outperformed the standard K-means clustering based on the results for breast cancer data set; in terms of clustering accuracy (0.29%) and computing time (95.76%).

ABSTRAK

Teknik perlombongan data digunakan untuk menganalisis corak dari set data untuk mendapatkan maklumat yang berguna. Pengelasan set data menjadi kelompok ialah satu daripada proses penting dalam manipulasi data. Salah satu kaedah pengelompokan yang paling popular dan cekap ialah purata-K. Walau bagaimanapun, terdapat kesukaran menganalisis set data berdimensi tinggi bagi kaedah pengelompokan purata-K dengan wujudnya nilai data yang hilang. Tambahan pula, kajian terdahulu menunjukkan bahawa ciri set data berdimensi tinggi yang dipersembahkan mempunyai masalah berbeza bagi pengelompokan purata-K. Bagi masalah nilai yang hilang, kaedah imputasi diperlukan untuk meminimumkan kesan set data berdimensi tinggi yang tidak lengkap dalam proses pengelompokan purata-K. Kajian ini mengkaji kesan algoritma imputasi dan teknik pengurangan dimensi terhadap prestasi pengelompokan purata-K. Tiga kaedah imputasi dilaksanakan untuk anggaran nilai yang hilang iaitu K-Jiran Terdekat (KNN), Kuasa Dua Terkecil Setempat (LLS) dan Analisis Komponen Utama Bayesian (BPCA). Analisis Komponen Utama (PCA) ialah kaedah pengurangan dimensi yang mempunyai keupayaan pengurangan dimensi dengan mengeluarkan atribut yang tidak perlu bagi set data berdimensi tinggi. Oleh itu, hibrid PCA dengan purata-K (PCA purata-K) dicadangkan untuk memberikan hasil pengelompokan yang lebih baik. Proses eksperimen dilakukan dengan menggunakan set data Kanser Payudara Wisconsin. Dengan menggunakan kaedah imputasi LLS, PCA purata-K hibrid yang dicadangkan telah menghasilkan pengelompokan purata-K yang lebih baik berbanding dengan purata-K piawai berdasarkan hasil bagi set data kanser payudara, dari segi ketepatan pengelompokan (0.29%) dan masa pengkomputeran (95.76%).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC	Area under ROC curve
BPCA	Bayesian Principal Component Analysis
KNN	K-nearest Neighbour
LLS	Local Least Square
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristics

LIST OF SYMBOLS

FN	False Negative
FP	False Positive
TN	True Negative
TP	True Positive

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

This chapter discusses about the introduction of this research. The contents include information about missing value, and Principal Component Analysis for dimension reduction and K-means clustering. Then, the problem background and problem statement are stated, three research objectives are presented from the aim research, and research scope is identified.

1.2 Background of Problem

In recent time, data analyzing methods are important for massive quantity of high dimensional data set. In many field area, such as image processing, computational biology, information retrieval and global climate research, high dimensional datasets are frequently encountered. Classification or bunching of these data into set of categories or clusters is one of the essential in manipulating these data (Abbas et al, 2008). To analyze such high dimensional data set, one of the data manipulating methods is clustering.

Clustering is a delineative task that partition data points into disjoint group such that data point belonging to same cluster are similar while data point that belong to different clusters is dissimilar. One of the most popular and efficient clustering methods is the K-means method which uses prototypes (centroids) to represent clusters by optimizing the squared error function. However, clustering of high dimensional data set poses demanding task that should satisfy both the requirement of the computation efficiency and result quality. K-means clustering algorithm often does not work well for high dimension data set (Mann et al., 2013). Furthermore, the presence of missing value in data sets has been major problem for precise prediction. Real life database contains large data sets with lot of presence of multiple missing value. Unfortunately, many algorithms for data mining analysis require a complete matrix of values as input. For example, analysis method such as hierarchical clustering and k-means clustering are not robust to missing data (Napoleon et al., 2011), and reduce precision of analysis even with few missing value (Moorthy et al., 2014). In order to overcome presence of multiple missing values problem, methods for imputing missing value are needed to minimize the effect of incomplete data sets for prediction model. Three imputation methods are applied for missing value estimation such as K-nearest neighbors (KNN), Least Local Square (LLS) and Bayesian Principle Component Analysis (BPCA) which is often used for Wisconsin Breast Cancer data set for missing value estimation.

Moreover, to mine high dimensional data set, an efficient reduction technique is very important (Napoleon et al, 2011). In dimension reduction technique, data features are reduced by transforming the original high dimension data set to a lower dimension one through Eigen value decomposition (Falahi et al., 2014). Numerous methods have been conducted and many experimental analyses have been done to find out an efficient reduction technique so as to reduce the dimension of a high dimensional data set without affecting the original data’s. One of the widely used dimension reduction techniques is by using Principal Component Analysis (PCA) technique (Aydilek et al., 2014).

1.3 Problem Statements

Problems related to the K-mean clustering on incomplete dataset are stated in the questions and the highlight research question as below:

i. How to solve the problem of missing values occurred in high dimension data set and determine the best method?

ii. How to overcome the problem of cluster high dimensional dataset for breast cancer?

iii. How does PCA can potentially reduce the high dimensional dataset for clustering breast cancer process?

iv. How does hybrid PCA-K-mean can potentially improve the performance of clustering result in term of accuracy and computing time?

In this research, the improvement aspect is focused on using the new approach by implementing PCA as dimension reduction technique in clustering the high dimensional dataset. The focuses are indicated as below:

i. Imputation methods, KNN, LLS and BPCA, are used to estimate missing value of incomplete breast cancer dataset and compared by using normalized root mean square error.

ii. PCA is used to assist standard K-mean clustering to reduce data features by transforming complete high data set into low dimension data set.

iii. PCA dimension reduction technique is used for clustering complete dimensional dataset is used to determine a better performance in terms of accuracy and computing time.

1.4 Objectives of the Research

The aim of this research is to propose a better hybrid clustering algorithm by using various imputed techniques data to obtain a better performance of clustering result.

The objectives of this research are:

i. To develop a complete high dimensional dataset by using imputation method for clustering breast cancer dataset and compare the performance using NMRSE.

ii. To develop a new hybrid PCA-K-mean algorithm to solve the problem of high dimensionality of data sets and improve the performances of K-mean clustering of breast cancer dataset.

iii. To determine the performances of the new hybrid PCA-K-mean algorithm in terms of accuracy and time complexity

1.5 Scopes of the Study

The scopes of this research are:

i. The study focusses on K-mean clustering technique.

ii. Wisconsin breast cancer is used as experimental dataset.

iii. PCA is used as dimension reduction technique for clustering breast cancer dataset.

iv. Percentage difference of accuracy and computing time between standard clustering and PCA-K-mean is used to evaluate performance of PCA-K-mean clustering technique.

1.6 Research Significant and Contribution

This study used KNN, LLS and BPCA imputation techniques to estimate missing value of incomplete breast cancer dataset. Consequently, PCA is used as a dimension reduction technique for clustering the breast cancer dataset. The PCA-k-mean has potential to improve the accuracy and computing time for better performance for clustering result. Finally, this research significantly indicates that proposed hybrid enables to improve result of clustering by generating complete lower dimensional breast cancer data set.

At the end of this research, the contributions are will be carry out on the new hybrid algorithm and the comparative analysis of different imputation techniques. This research will have the following contributions:

i. This research proposes PCA in clustering as dimension reduction technique to improve clustering result.

ii. This research determines the best imputation method used for missing value estimation using hybrid algorithm.

iii. This research proposes PCA in clustering to produce clustering result in shorter time and better accuracy.

iv. Data mining process for clustering analysis is expected to give more accurate and faster computing time.

1.7 Summary

In this chapter, overall overview of the title, principal component analysis and clustering technique are explained. Besides that, problem background and problem statements are stated. Then, the research aims, and objectives are included. The scopes of the research are also presented are discussed in this chapter.

REFERENCES

Abbas, O. A. (2008). Comparisons Between Data Clustering Algorithms. International Arab Journal of Information Technology (IAJIT), 5(3).

Afef, Brahim & Limam, M. (2013). Robust ensemble feature selection for high dimensional data sets. In High Performance Computing and Simulation (HPCS), 2013 International Conference on (pp. 151-157). IEEE.

Archana, Purwar, & Singh, S. K. (2015). Hybrid prediction model with missing value imputation for medical data. Expert Systems with Applications, 42(13), 5621-5631.

Aittokallio, T. (2009). Dealing with missing values in large-scale studies: microarray data imputation and beyond. Briefings in bioinformatics, 11(2), 253-264.

Asadi, Srinivasulu D. (2010). Ch. DV Subba Rao, V. Saikrishna “A Comparative study of Face Recognition with Principal Component Analysis and Cross-Correlation Technique”. International Journal of Computer Applications (0975–8887), 10(8), 17-21.

Aydilek, I. B., & Arslan, A. (2012). A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks. International Journal of Innovative Computing, Information and Control, 7(8), 4705-4717.

Aydilek, I. B., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Information Sciences, 233, 25-35.

Barnard, J., & Meng, X. L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. Statistical methods in medical research, 8(1), 17-36.

Batista, G. E., & Monard, M. C. (2002). A Study of K-Nearest Neighbour as an Imputation Method. HIS, 87(251-260), 48.

Blum, M. G., Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation.

Statistical Science, 28(2), 189-208.

Brás, L. P., & Menezes, J. C. (2007). Improving cluster-based missing value estimation of DNA microarray data. Biomolecular engineering, 24(2), 273-282.

Ceren, Güzel, M. Mahmut Kaya, and Oktay Yıldız. "Breast Cancer Diagnosis Based on Naïve Bayes Machine Learning Classifier with K-Nearest Neighbour Missing Data Imputation." 3rd World conference on innovation and Computer Sciences. 0.9 0.92 0.94 0.96 0.98. 2013.

Chandar, K. P., Chandra, M. M., Kumar, M. R., & Latha, B. S. (2011). Multi scale feature extraction and enhancement using SVD towards secure face recognition system. In Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), 2011 International Conference on(pp. 64-69). IEEE.

Chuang, L. Y., Tsai, J. H., & Yang, C. H. (2010). Binary particle swarm optimization for operon prediction. Nucleic acids research, 38(12), e128-e128.

Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient kNN classification algorithm for big data. Neurocomputing, 195, 143-148.

Dubey, S. R., Dixit, P., Singh, N., & Gupta, J. P. (2013). Infected fruit part detection using k-means clustering segmentation technique. Ijimai, 2(2), 65-72.

Duma, M., Marwala, T., Twala, B., & Nelwamondo, F. (2013). Partial imputation of unseen records to improve classification using a hybrid multi-layered artificial immune system and genetic algorithm. Applied Soft Computing, 13(12), 4461-4480.

Engchuan, W., Meechai, A., Tongsima, S., & Chan, J. H. (2016). Handling batch effects on cross-platform classification of microarray data. International Journal of Advanced Intelligence Paradigms, 8(1), 59-76.

Falahi A, Kanna & Atif, Y & Abraham, Ajith. (2014). Models of Influence in Online Social Networks. International Journal of Intelligent Systems. 29. . 10.1002/int.21631.

Gan, X., Liew, A. W. C., & Yan, H. (2006). Microarray missing data imputation based on a set theoretic framework and biological knowledge. Nucleic Acids Research, 34(5), 1608-1619.

García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. Neural Computing and Applications, 19(2), 263-282.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1), 389-422.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning. In The elements of statistical learning (pp. 9-41). Springer New York.

Hourani, M. A., & El, E. I. M. (2009). Microarray missing values imputation methods: Critical analysis review. Computer Science and Information Systems, 6(2), 165-190.

Hsu, H. H., Hsieh, C. W., & Lu, M. D. (2011). Hybrid feature selection by combining filters and wrappers. Expert Systems with Applications, 38(7), 8144-8150.

James, G. M., Hastie, T. J., & Sugar, C. A. (2000). Principal component models for sparse functional data. Biometrika, 87(3), 587-602.

Jain, A., & Satish, B. (2009). Clustering based short term load forecasting using support vector machines. In PowerTech, 2009 IEEE Bucharest (pp. 1-8). IEEE.

Jerez, J. M., Molina, I., Subirats, J. L., & Franco, L. (2006). Missing data imputation in breast cancer prognosis. BioMed, 6, 323-328.

Jonsdottir, T., Hvannberg, E. T., Sigurdsson, H., & Sigurdsson, S. (2008). The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining. Expert Systems with Applications, 34(1), 108-118.

Kastrin, A., & Peterlin, B. (2010). Rasch-based high-dimensionality data reduction and class prediction with applications to microarray gene expression data. Expert Systems with Applications, 37(7), 5178-5185.

Keerin, P., Kurutach, W., & Boongoen, T. (2013). An improvement of missing value imputation in DNA microarray data using cluster-based LLS method. In Communications and Information Technologies (ISCIT), 2013 13th International Symposium on (pp. 559-564). IEEE.

Kim, H., Golub, G. H., & Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics, 21(2), 187-198.

Lian En, C. H. A. I., Chow Kuan, L. A. W., Mohd Saberi Mohamad, C. K. C., & Yee Wen Choon, S. D. (2014). Investigating the effects of imputation methods for modelling gene networks using a dynamic bayesian network from gene expression data. The Malaysian journal of medical sciences: MJMS, 21(2), 20.

Li, H., Zhao, C., Shao, F., Li, G. Z., & Wang, X. (2015). A hybrid imputation approach for microarray missing value estimation. BMC genomics, 16(9), S1.

Li, C., Diao, Y., Ma, H., & Li, Y. (2008). A statistical PCA method for face recognition. In Intelligent Information Technology Application, 2008. IITA'08. Second International Symposium on (Vol. 3, pp. 376-380). IEEE.

Li, S., Wu, X., & Hu, X. (2008). Gene selection using genetic algorithm and support vectors machines. Soft Computing-A Fusion of Foundations, Methodologies and Applications, 12(7), 693-698.

Liew, A. W. C., Law, N. F., & Yan, H. (2010). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in bioinformatics, 12(5), 498-513.

Lilien, R. H., Farid, H., & Donald, B. R. (2003). Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. Journal of computational biology, 10(6), 925-946.

Maillo, J., Ramírez, S., Triguero, I., & Herrera, F. (2017). kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data. Knowledge-Based Systems, 117, 3-15.

Mann, A. K., & Kaur, N. (2013). Review paper on clustering techniques. Global Journal of Computer Science and Technology.

Moorthy, K., Saberi Mohamad, M., & Deris, S. (2014). A review on missing value imputation algorithms for microarray gene expression data. Current Bioinformatics, 9(1), 18-22.

Muhammad, A., Mazliham, M. S., Boursier, P., & Shahrulniza, M. (2011). K-nearest neighbor algorithm for improving accuracy in clutter based location estimation of wireless nodes. Malaysian Journal of Computer Science, 24(3).

Napoleon, D., & Lakshmi, P. G. (2010). An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points. In Trendz in Information Sciences & Computing (TISC), 2010(pp. 42-45). IEEE.

Napoleon, D., & 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.

Nguyen, D. V., & Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. Bioinformatics, 18(1), 39-50.

Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K. I., & Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. Bioinformatics, 19(16), 2088-2096.

Ouyang, M., Welsh, W. J., & Georgopoulos, P. (2004). Gaussian mixture clustering and imputation of microarray data. Bioinformatics, 20(6), 917-923.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., ... & Worek, W. (2005, June). Overview of the face recognition grand challenge. In Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on (Vol. 1, pp. 947-954). IEEE.

Porter, P. L. (2009). Global trends in breast cancer incidence and mortality. salud pública de méxico, 51, s141-s146.

Purwar, A., & Singh, S. K. (2015). Hybrid prediction model with missing value imputation for medical data. Expert Systems with Applications, 42(13), 5621-5631.

Prat Aparicio, A. (2012). Comprehensive molecular portraits of human breast tumours. Nature, 2012, vol. 490, num. 7418, p. 61-70.

Robbins, K. R., Zhang, W., Bertrand, J. K., & Rekaya, R. (2007). The ant colony algorithm for feature selection in high-dimension gene expression data for disease classification. Mathematical medicine and biology: a journal of the IMA, 24(4), 413-426.

Rubin, D. B. (2004). Multiple imputation for nonresponse in surveys (Vol. 81). John Wiley & Sons.

Schmitt, P., Mandel, J., & Guedj, M. (2015). A comparison of six methods for missing data imputation. Journal of Biometrics & Biostatistics, 6(1), 1.

Sehgal, M. S. B., Gondal, I., & Dooley, L. S. (2005). Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. Bioinformatics, 21(10), 2417-2423.

Sehgal, M. S. B., Gondal, I., Dooley, L. S., & Coppel, R. (2008). Ameliorative missing value imputation for robust biological knowledge inference. Journal of biomedical informatics, 41(4), 499-514.

Shin, H. R., Curado, M. P., Ferlay, J., Heanue, M., Edwards, B., & Storm, H. (2007). Comparability and quality of data. Cancer incidence in five continents, 9.

De Souto, M. C., Jaskowiak, P. A., & Costa, I. G. (2015). Impact of missing data imputation methods on gene expression clustering and classification. BMC bioinformatics, 16(1), 64.

Tan, F., Fu, X., Zhang, Y., & Bourgeois, A. G. (2006). Improving feature subset selection using a genetic algorithm for microarray gene expression data. In Evolutionary Computation, 2006. CEC 2006. IEEE Congress on (pp. 2529-2534).

IEEE.

Tang, J., Zhang, G., Wang, Y., Wang, H., & Liu, F. (2015). A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. Transportation Research Part C: Emerging Technologies, 51, 29-40.

Thomas, Somasundaram, R. S., & Nedunchezhian, R. (2011). Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. International Journal of Computer Applications, Vol21, 10.

Tian, J., Yu, B., Yu, D., & Ma, S. (2014). Missing data analyses: a hybrid multiple imputation algorithm using gray system theory and entropy based on clustering. Applied intelligence, 40(2), 376-388.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520-525.

Tuikkala, J., Elo, L., Nevalainen, O. S., & Aittokallio, T. (2005). Improving missing value estimation in microarray data with gene ontology. Bioinformatics, 22(5), 566-572.

Xiang, Q., Dai, X., Deng, Y., He, C., Wang, J., Feng, J., & Dai, Z. (2008). Missing value imputation for microarray gene expression data using histone acetylation information. BMC bioinformatics, 9(1), 252.

Zhang, X., Song, X., Wang, H., & Zhang, H. (2008). Sequential local least squares imputation estimating missing value of microarray data. Computers in biology and medicine, 38(10), 1112-1120.

Zhang, Y., Ding, C., & Li, T. (2007). A two-stage gene selection algorithm by combining reliefF and mRMR. In Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on (pp. 164-171). IEEE.

Zhou, Y., Cao, S., Wen, D., Zhang, H., & Zhao, L. (2011). The study of face recognition based on hybrid principal components analysis and independent component analysis. In Electronics, Communications and Control (ICECC), 2011 International Conference on (pp. 2964-2966). IEEE.