Indonesian Journal of Electrical Engineering and Computer Science
Vol. 16, No. 2, November 2019, pp. 752~758
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v16.i2.pp752-758

A comparative study on dimensionality reduction between principal component analysis and k-means clustering

Norsyela Muhammad Noor Mathivanan[1], Nor Azura Md.Ghani[2], Roziah Mohd Janor[3]
[1,2,3]Center for Statistical and Decision Sciences Studies,
Faculty of Computer & Mathematical Sciences Universiti Teknologi MARA, Malaysia
[2]National Design Centre Universiti Teknologi MARA, Malaysia

Article Info

Article history:

Received Jan 22, 2019
Revised Apr 20, 2019
Accepted May 14, 2019

Keywords:

Clustering
Feature selection
Principal component analysis
Simulation

ABSTRACT

The curse of dimensionality and the empty space phenomenon emerged as a critical problem in text classification. One way of dealing with this problem is applying a feature selection technique before performing a classification model. This technique helps to reduce the time complexity and sometimes increase the classification accuracy. This study introduces a feature selection technique using K-Means clustering to overcome the weaknesses of traditional feature selection technique such as principal component analysis (PCA) that require a lot of time to transform all the inputs data. This proposed technique decides on features to retain based on the significance value of each feature in a cluster. This study found that k-means clustering helps to increase the efficiency of KNN model for a large data set while KNN model without feature selection technique is suitable for a small data set. A comparison between K-Means clustering and PCA as a feature selection technique shows that proposed technique is better than PCA especially in term of computation time. Hence, k-means clustering is found to be helpful in reducing the data dimensionality with less time complexity compared to PCA without affecting the accuracy of KNN model for a high frequency data.

Corresponding Author:

Nor Azura Md.Ghani,
Center for Statistical and Decision Sciences Studies
Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA
40450 Shah Alam, Selangor Malaysia
Email: azura@tmsk.uitm.edu.my

1. INTRODUCTION

A huge amount of information can be obtained from different form of data with the rapid growth of the Internet. These data are virtually provided through digital libraries, e-commerce websites, social networks, mobile applications and other sources [1]. Currently, one of the major form of data is unstructured text [2]. These data are complex and not well-organized unlike structured text. They normally face the curse of dimensionality. A vector of word counts in a vector-space model of text documents may consists dimensionality more than 10,000 and the given sample size need to estimate a function of several variables to provide a good accuracy of the model. However, most of the high dimensional data are inherently sparse data [3]. For instance, a word may appear 100 times in one document but may not appear in other documents. Hence, the data need to undergone a good data pre-processing process to obtain the best structure of data to be used as an input for prediction or classification models.

One possible approach to simplify a high dimensional data is to apply some form of dimensionality reduction [4]. This can be done in two different ways either by using feature extraction or feature selection. In feature extraction, the original vector space is transformed into new vector space according to special characteristics. On the other hand, feature selection is used to keep the most relevant variables from the

original data set. The utilization of both techniques accordingly provide a better data pre-processing process [5]. Many researchers claimed that principal component analysis (PCA) is the most popular feature extraction method [6-8]. PCA is a classical statistical technique to transform attributes of data set into new set of uncorrelated attributes known as principal components. This technique is used to reduce the dimensionality while maintaining as much of the variability of the data set as possible [9].

PCA can increase the efficiency given the classifiers taking place in a smaller dimension but when using this technique, the time requirement for pre-processing the data is increasing tremendously. PCA is an unsupervised technique, which makes no use of information related to the class variable. There is another form of unsupervised technique called as clustering technique. One of the well-known clustering technique is k-means clustering. The simplicity and efficiency of this clustering algorithm make it useful for discovering the structure of data. Hence, this study proposed alternative method to reduce the dimensionality of the data by using the feature selection technique with k-means clustering. The comparison between PCA and k-means clustering in selecting the features for high dimensional data are provided in the study.

2. DATA BACKGROUND

The study presents experimental results using two types of data sets which are real and synthetic data sets. There are two real data sets used in the study. Table 1 shows the description of the two corpora selected for this study. The first data set has been collected from one of the major chain market online stores in Malaysia using prototype web scrapers developed under STATSBDA project known as Price Intelligence (PI) by Department of Statistics Malaysia (DOSM). A few of leaf nodes are selected to represent categories from the browse tree of the website. The data corpus consists of products' description for four categories under baby products which are baby diapers and wipes, baby milk powder, baby food, and baby toiletries.

Table 1. Summary Description of Data Sets

Dataset	Category	Instance	Number of Feature
Baby	4	471	387
SMS Spam	2	5574	6981

The second data set is a well-known data collection that provides an ideal benchmark used to evaluate text classification model [10]. The SMS spam messages data set is originally collected from the Grumbletext Website where cell phone users make public claims about SMS spam messages [11]. This data set consists of two categories either ham or spam message. The study has also used a simulation data to compare the performance of selecting the features between K-Means clustering and PCA. Each class is composed of a number of normally distributed clusters. A Normal distribution with mean and standard deviation equal to zero and one accordingly is used to draw number of useful independent features for each cluster. The simulation data deal with two-class classification problem with sparse binary input features. The data is generated through a hypercube data program [12] which is appended in scikit-learn of python programming.

3. RESEARCH METHOD

There are several steps involve in performing a text classification. This study is composed of the basic steps which are data extraction, data preprocessing and feature extraction. There are several steps involve in pre-processing the data which are tokenization, word stop removal, and stemming process[13]. This study has used bag-of-word to extract the features before performing the feature selection to reduce the dimensionality of the data. All the procedures use in the study are implemented through R-Programming Software. The software has been widely used to solve a statistical problem in various field of studies include in study of population growth [14], age prediction [15], pattern recognition [16] etc.

3.1. Data Dimensionality Reduction Techniques

There are two different feature selection approaches use in the study which are PCA and feature selection with K-Means clustering.

3.1.1 Technique I: Principal Component Analysis (PCA)

PCA is a linear method uses to embed the data into a linear subspace of lower dimensional. The steps involve are shown in Figure 1. The method finds a linear basis which is possible orthogonal of reduced dimensionality for the data with containing the maximum number of variance in the data. Mathematically, let

P be a matrix of data with *N* observations and *F* features and let *p(n)* represent the n^{th} row vector. The data are transformed into the principal component space by $t_{j(n)} = w_j \cdot p_{(n)}$, where w_j is the *F*-dimension loading vector and $t_{j(n)}$ is the j^{th} component score. The weight of the first principal component w_1 is found by

$$\left\{ w_1 = \arg\max \frac{w^T P^T P w}{w^T w} \right\} \tag{1}$$

The next principal components can be obtained by subtracting the first *j* components from the data,

$$\hat{P}_j = P - \sum_{m=1}^{j-1} P w_m w^T{}_m \tag{2}$$

and the loadings is calculated by,

$$\left\{ w_j = \arg\max \frac{w^T \hat{P}^T \hat{P} w}{w^T w} \right\} \tag{3}$$

Normally, the first few principal components consist a majority of the variance. However, the number of principal components need to be included in the new transform data depends on the ability of the j^{th} principal components to provide full information about the actual data.

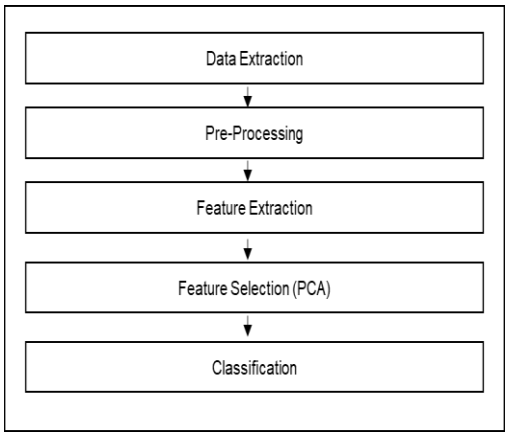

Figure 1. Text classification with feature selection using PCA

3.1.2 Technique II: K-Means Clustering

The k-means clustering is a well-known algorithm that follows a gradient descent procedure [13]. The features undergo the first level of feature selection with using one of feature selection techniques namely Correlation-based feature selection (CFS). It is used to filter the feature before using the clustering technique. The steps involve are shown in Figure 2. Given the data set size of n with data points of p_1, p_2, …, p_n where each data point is in the K^d. Then, the minimum variance clustering of the data set is separated into k clusters by finding the k points $\{m_c\}$ (c=1,2, …, k) in K^d such that,

$$\frac{1}{n} \sum_{i=1}^{n} [\min_c d^2(x_i, m_c)] \tag{4}$$

Is minimized, where d (x_i, m_c) denotes the Euclidean distance between x_i and m_c. The technique begins with randomly select the cluster centroids, and iteratively updates these centroids to decrease the objective function in (4). The algorithm will keep updating the cluster centroids until the local minimum is

found. After obtaining the desired clusters, the CFS technique is re-apply to the clustered data for reducing the feature in each cluster. The features in each cluster are gathered back together as the input data for the classification model.

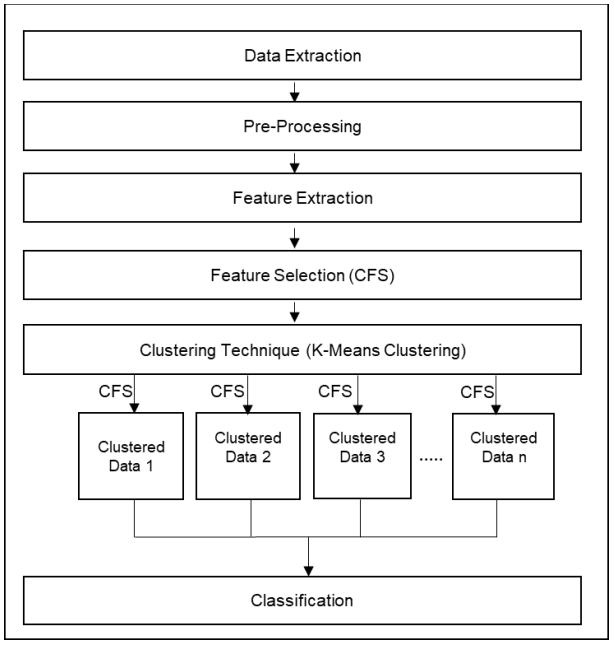

Figure 2. Text Classification with feature selection using k-means clustering

3.2. Classification Model

A supervised machine learning model is widely used by researches to solve classification problem [17]. The classification model that will be used in the study is K-Nearest Neighbor (KNN) model. The model is claimed to be one of the most effective classification models in text mining [18-19]. KNN is an instance-based learning where the function is only approximated locally and all computation is done during the classification. During the learning process, each item is assigned to a class represented by the majority label of its k-nearest neighbors in the training data set [20]. This study used the default nearest neighbor rule with the K value equal to one. The generalized pseudocode for KNN algorithm [21] is represented in Figure 3. The performance measures used to evaluate the trained data are accuracy, precision, recall and F1-measure. This study also measures the execution time of each classification model because it is also one of the important result can be measured from a study [22].

```
for all the unknown samples (p)
    for all the known samples (q)
        Compute the distance between each of p and q
    end for
        Find the k smallest distances locate the
        corresponding samples, p1 ,.., pk
        Assign each of q to the class which appears
        more frequently
end for
```

Figure 3. KNN Algorithm

4. RESULTS AND DISCUSSION

A feature selection technique is introduced to reduce a data complexity before performing classification model. This study found interesting outcomes related to usefulness of k-means clustering to reduce the dimensionality of high frequency data set. The performance evaluation for two real data sets used in the study is shown in Table 2 and 3 accordingly. Both data sets are partitioned into 70% of training text

data and 30% of testing text data. There are three KNN models involve which are no feature selection (KNN), feature selection using K-Means Clustering (KM-KNN) and feature selection using PCA (PCA-KNN).

Table 2. Performance Measure for Baby Data Set

Model	Accuracy (%)	Precision	Recall	F1-Measure	Execution Time (Second)
KNN	97.18	0.9717	0.9767	0.9793	0.81
PCA-KNN	97.18	0.9717	0.9767	0.9793	1.56
KM-KNN	97.18	0.9717	0.9767	0.9793	1.06

Table 3. Performance Measure for SMS Spam Data Set

Model	Accuracy (%)	Precision	Recall	F1-Measure	Execution Time (Second)
KNN	95.16	0.9701	0.8026	0.7268	510.59
PCA-KNN	95.34	0.9748	0.8079	0.7322	1587.03
KM-KNN	95.52	0.9561	0.8280	0.7686	490.81

The performance of the three models are similar for baby data set. However, the KNN works faster than the other two models. Meanwhile, there is an improvement for the performance of KM-KNN compared to other models for sms spam data set. In addition, the model also consumes less computation time. From the comparison, it is shown that an accuracy of a small data set may not be affected by a model without any feature selection techniques. However, these techniques seem to help in increasing the model accuracy and efficiency for a large data set. This study has also found that PCA is able to reduce data dimensionality but it requires a certain amount of time to transform the data before performing the classification model. This study has also observed the performance of both feature selection techniques through simulation data. The comparison between the three models are visualized in Figure 4.

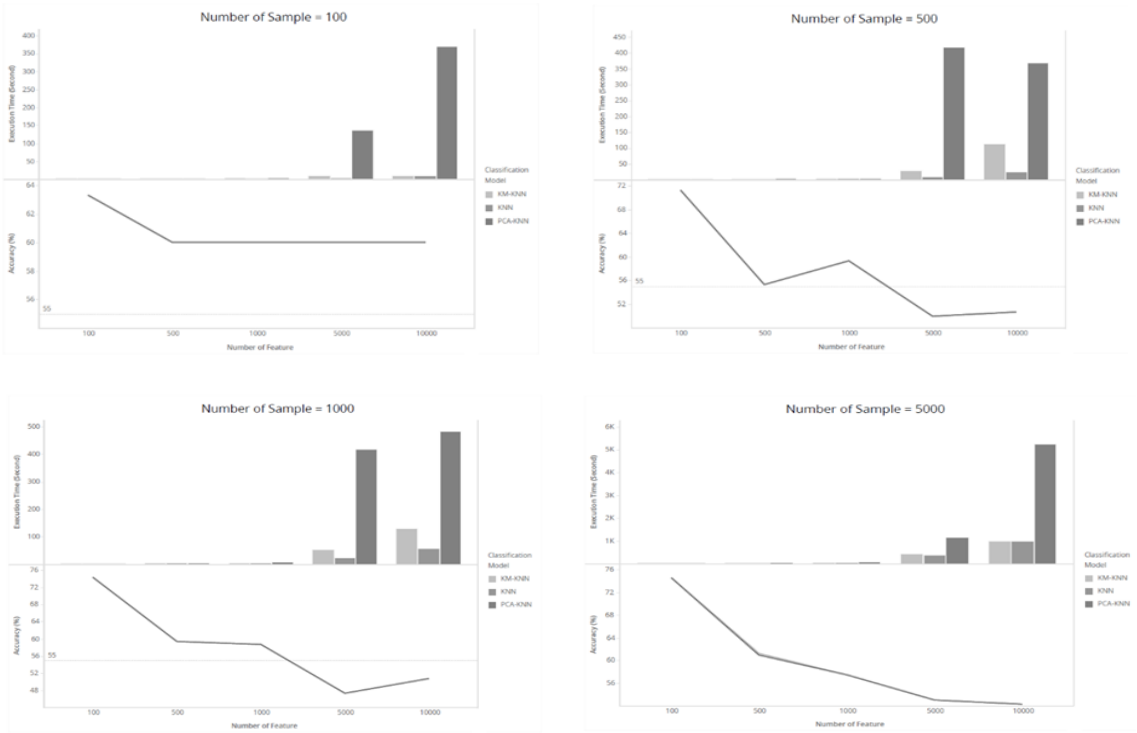

Figure 4. Comparison of Accuracy and Execution Time between Three KNN Models for Simulation Data

It is shown that the accuracy of the data is remained the same after applying feature selection techniques such as PCA and K-Means Clustering. The possible explanation is a feature selection technique

may act as a way to reduce the dimensionality and ease the computation of KNN model but it does not influence the performance of the model. This result is supported by previous studies where they claimed there is a high tendency that the complexity of the computation is being reduced without affecting the performance of a classification model [23-24]. Hence, this study found that the accuracy of KNN model remains the same with the application of feature selection towards normally distributed data set.

It is also apparent from Figure 4. that PCA requires a lot of time to transform the data with the increases number of samples and features. The result in line with previous studies that found the disadvantage of PCA when applied to large datasets where a huge amount of time is required in performing an eigenvalue decomposition to find the principal components [9, 25]. Meanwhile, it is noticeable that the execution time for KM-KNN is getting closer to KNN model as the number of feature increases from 100 to 10000. This shows that K-Means clustering is useful in reducing the data dimensionality with less amount of time for high frequency data set.

5. CONCLUSION

This study is mainly focused on evaluating the efficiency of KNN model using feature selection techniques. The most obvious finding to emerge from this study is that k-means clustering helps in increasing the efficiency of KNN model for a large data set. This study has also identified that KNN model without feature selection technique is suitable for a small data set. The proposed feature selection technique with using K-Means clustering performs better than the existing well-known feature selection technique which is PCA. This technique is helpful because researchers often deal with large number of features especially in text mining.

ACKNOWLEDGEMENTS

The research is financially supported by the University Teknologi MARA and Ministry of Education Malaysia under the Grant Scheme (600-IRMI/FRGS 5/3 (120/ 2019)). The authors would like to express their sincere appreciation to the Department of Statistics Malaysia for providing knowledge and data supports.

REFERENCES

[1] Mohammad Fikri and Riyanarto Sarno, "A comparative study of sentiment analysis using SVM and SentiWordNet," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, (3), pp. 902-909, Mar 2019.
[2] S. Havre, E. Hetzler, P. Whitney, & L. Nowell. ThemeRiver., "Visualizing thematic changes in large document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8(1), pp. 9-20, Jan 2002.
[3] S. Huang, M. O. Ward & E. A. Rundensteiner., "*Exploration of dimensionality reduction for text visualization*," in Proc. IEEE Third Intl. Conf. on Coordinated and Multiple Views in Exploratory Visualization, pp. 63-74, Jul 2005.
[4] J. Verbeek., *Supervised feature extraction for text categorization*, in Tenth Belgian-Dutch Conference on Machine Learning (Benelearn'00), 2000.
[5] Muhammad 'Arif Mohamad, Haswadi Hassan, Dewi Nasien & Habibollah Haron., "A Review on Feature Extraction and Feature Selection for Handwritten Character Recognition," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 6(2), 2015.
[6] A.I. Taloba, D.A. Eisa & S.S.A. Ismail., "Comparative Study on using Principle Component Analysis with Different Text Classifiers". CoRR, abs/1807.03283. 2018.
[7] S. L. Lam and D. L. Lee, "*Feature reduction for neural network based text categorization*," in Database Systems for Advanced Applications, 1999. Proceedings., 6th International Conference on, IEEE, pp. 195-202, 1999.
[8] A. Selamat and S. Omatu, "Web page feature selection and classification using neural networks," *Information Sciences*, vol. 158, pp. 69-88, 2004.
[9] Thendral Tharmalingam and Vijaya Vijayakumar, "A Hybrid Linear Kernel with PCA in SVM Prediction Model of Tamil Writing Pattern," *International Journal of Simulation Systems, Science & Technology (IJSST)*, 19.04.21, Aug 2018.
[10] S. J. Delany, M. Buckley & D. Greene., "SMS spam filtering: Methods and data," *Expert Systems with Applications*, vol. 39(10), pp. 9899-9908, 2012.
[11] T.A. Almeida, J.M. GÃ³mez Hidalgo, & A. Yamakami., *Contributions to the Study of SMS Spam Filtering: New Collection and Results*, Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11), Mountain View, CA, USA, 2011.
[12] S. Perkins, K. Lacker & J. Theiler, "Grafting: Fast, Incremental Feature Selection by Gradient Descent in Function Space," *JMLR*, vol. 3, pp. 1333-1356, Mar 2003.
[13] Norsyela Muhammad Noor Mathivanan, Nor Azura Md Ghani, & Roziah Mohd Janor., "Improving Classification Accuracy Using Clustering Technique," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 7(3), pp. 465-470, Sep 2018.

[14] Norsyela Muhammad Noor Mathivanan, Puzziawati Ab Ghani, & Nor Azura Md Ghani., "Tracing Mathematical Function of Age Specific Fertility Rate in Peninsular Malaysia," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol 9(3), pp. 637-642, 2018.
[15] I. M. Umesh, G. N. Srinivasan & Matheus Torquato, Software Aging Forecasting Using Time Series Model", *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, Vol 8(3), pp. 589-596.2017.
[16] Nur Azimah Abdul Rahim, Nor Azura Md Ghani, Norazan Mohamed, Hishamuddin Hashim & Ismail Musirin. "The Application of Modified Least Trimmed Squares with Genetic Algorithms Method in Face Recognition," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 8(1), pp. 154-158. 2017.
[17] Nor Azura Md Ghani, Saadi bin Ahmad Kamaruddin, Norazan Mohamed Ramli, Ismail Musirin & Hishamuddin Hashim., "Modified BPNN via Iterated Least Median Squares, Particle Swarm Optimization and Firefly Algorithm," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 8(3), pp. 779-786, 2017.
[18] Norsyela Muhammad Noor Mathivanan, Nor Azura Md Ghani & Roziah Mohd Janor., "E-Commerce Product Classification Using Supervised Learning Models," *International Journal of Engineering & Technology*, vol. 8(1.7), pp. 214-218, 2019.
[19] D. Hand, et al., "Principles of data mining," *International journal of medical toxicology and drug experience*, vol. 30, 2001.
[20] S. Kanj, F. Abdallah, T. Denoeux & K. Tout., "Editing training data for multi-label classification with the k-nearest neighbor rule," *Pattern Anal. Appl.*, vol. 19(1), pp. 145-161, 2015.
[21] S.B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background," *International Journal of Engineering Research and Applications*, vol. 3(5), pp. 605-610, 2013.
[22] Nor Azura Md Ghani, Saadi bin Ahmad Kamaruddin, Norazan Mohamed Ramli, Ismail Musirin & Hishamuddin Hashim., "Enhanced BFGS Quasi-Newton Backpropagation Models on MCCI Data," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 8(1), pp. 101-106, 2017.
[23] R. K. Bania., "Survey on Feature Selection for Data Reduction," *International Journal of Computer Applications*, vol. 94, pp. 1-7, 2014.
[24] Z. M. Hira and D. F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data," *Advances in Bioinformatics*, vol. 2015, Article ID 198363, 13 pages, 2015.
[25] E. Martel, R. Lazcano, J. López,, D. Madroñal, R. Salvador, S. López, E. Juarez, R. Guerra, C. Sanz, R. Sarmiento., "Implementation of the Principal Component Analysis onto High-Performance Computer Facilities for Hyperspectral Dimensionality Reduction: Results and Comparisons," *Remote Sens.*, vol. 10(6), pp. 864. 2018.

BIOGRAPHIES OF AUTHORS

Norsyela Muhammad Noor Mathivanan is now a doctorate student in the Center for Statistical Studies and Decision Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia under the supervision of Nor Azura Md. Ghani and Roziah Mohd Janor. Her research interest related to big data, text mining and machine learning.
E-mail: syelamohdnoor@gmail.com

Nor Azura Md.Ghani is an Associate Professor in Center for Statistical Studies and Decision Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia. She is also Head of Data Research Unit, Research Management Center, Institute Research Management & Innovation, Universiti Teknologi MARA, Malaysia and Vice Chair IEEE Computer Society Malaysia Chapter. Her expertise is big data, statistical pattern recognition and forensic statistics.
E-mail: azura@tmsk.uitm.edu.my

Roziah Mohd Janor is a Professor of Statistics at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Malaysia. Currently she is serving as the Assistant Vice Chancellor at the Institute Quality & Knowledge Advancement, UiTM and she is now overseeing all the quality initiatives of the university, including institutional accreditation, programme accreditation, quality excellence model, quality management systems, Innovation @ Work and the University Ranking Project. Since 2018, she serves as the President of the MyQAN, a quality assurance network for all Malaysian higher education institutions.
E-mail: roziahmj@uitm.edu.my