

## Clustering heterogeneous categorical data using enhanced mini batch K-means with entropy distance measure

Nurshazwani Muhamad Mahfuz<sup>1</sup>, Marina Yusoff<sup>1,2</sup>, Zainura Idrus<sup>1</sup>

<sup>1</sup>Faculty of Computer and Mathematical Science, Universiti Teknologi MARA, Shah Alam, Malaysia

<sup>2</sup>Institute for Big Data Analytics and Artificial Intelligence, Universiti Teknologi MARA, Shah Alam, Malaysia

### Article Info

#### Article history:

Received Jan 22, 2022

Revised Jul 25, 2022

Accepted Aug 20, 2022

#### Keywords:

Categorical data

Clustering

Entropy distance measure

Heterogeneous

Mini batch k-means

Similarity measures

### ABSTRACT

Clustering methods in data mining aim to group a set of patterns based on their similarity. In a data survey, heterogeneous information is established with various types of data scales like nominal, ordinal, binary, and Likert scales. A lack of treatment of heterogeneous data and information leads to loss of information and scanty decision-making. Although many similarity measures have been established, solutions for heterogeneous data in clustering are still lacking. The recent entropy distance measure seems to provide good results for the heterogeneous categorical data. However, it requires many experiments and evaluations. This article presents a proposed framework for heterogeneous categorical data solution using a mini batch k-means with entropy measure (MBKEM) which is to investigate the effectiveness of similarity measure in clustering method using heterogeneous categorical data. Secondary data from a public survey was used. The findings demonstrate the proposed framework has improved the clustering's quality. MBKEM outperformed other clustering algorithms with the accuracy at 0.88, v-measure (VM) at 0.82, adjusted rand index (ARI) at 0.87, and Fowlkes-Mallow's index (FMI) at 0.94. It is observed that the average minimum elapsed time-varying for cluster generation,  $k$  at 0.26 s. In the future, the proposed solution would be beneficial for improving the quality of clustering for heterogeneous categorical data problems in many domains.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



### Corresponding Author:

Marina Yusoff

Institute for Big Data Analytics and Artificial Intelligence, Universiti Teknologi MARA

Shah Alam, Selangor, Malaysia

Email: marina998@uitm.edu.my

## 1. INTRODUCTION

The evolution of categorical data in data mining has been widely influenced by the need for more accurate and reliable techniques. Data mining is a procedure that seeks to understand data patterns through data exploration and extraction. One of the aspects of data mining is on clustering solution in which the unsupervised learning algorithm plays an important role. Unsupervised learning algorithms focus on no target variable or unlabeled data. The data in much actual categorical data is primarily obtained from questionnaires [1] and mainly in heterogeneous categorical data. The categories are nominal, ordinal, binary, and Likert. Since data mining must also support heterogeneous categorical data, clustering algorithms must be scalable. There are various clustering methods like hierarchical, partitioned, and density based. Usually, most of the clustering algorithms used are designed for numerical data only. Clustering utilizing data labeling techniques and distance computation can be directly applied to any numerical operation [2].

K-means is often used to obtain an optimal data partition by squaring the distance from a cluster center to a minimum data point. K-means algorithm is also well-known and widely utilized in data clustering. Its advantage is that it has a fast and straightforward convergence rate. However, the main drawback is that all dataset features are considered equal in determining the membership of the object cluster. Therefore, it is evident that algorithms can be easily influenced by outliers [3]. In this case, k-means utilized the Euclidean similarity measure for clustering. Traditionally, the Euclidean similarity measure converts the binary to a numerical value in applying the categorical clustering analysis. However, this method would omit similarities and loss of essential information values [4]. The similarity measures the strength of the relationship between two data items. It is often used to group similar data objects. A distance measure is a function that calculates a distance between two or more objects. It is a requirement for the calculation distance function to get the similarity between the objects.

The similarity measures that are mostly used in numerical and categorical data are Euclidean distance and Hamming distance, respectively [5]. The Hamming and Euclidean distance measures lack categorical data interpretation. For instance, k-modes with Hamming distance use binary variables to reduce variations of probability distributions of categorical data [6]. It is also associated with distance measures-based clustering, especially in capturing the characteristics of similarity measures within all features and affects the interpretation of information from original categorical data. Many computational experiments have been performed to determine a better clustering algorithm using heterogeneous categorical data from survey questionnaires. Another essential point to consider is that Likert scale data is normally treated as a numerical form for both distance measures, and the results gathered seem to be biased.

Recently, entropy distance measure has been introduced to assess its capability in improving categorical data interpretation [7]. The result has shown that entropy distance measures can handle heterogeneous categorical data like nominal, ordinal, binary, and Likert scales. Interestingly, Likert data is considered categorical to offer a meaningful feature. The entropy distance measure can also investigate the perception of harmony among various datasets. However, entropy distance measure still lacks evaluation on various heterogeneous categorical data, especially data obtained from questionnaires. In addition, most clustering algorithms still rarely use entropy distance measures. Therefore, this article presents a proposed framework for a heterogeneous categorical data solution that uses entropy as a distance measure with a clustering technique. A mini batch k-means with entropy measure (MBKEM) algorithm is then introduced. Generally, the result of clustering algorithms is selected based on the evaluation of the performance. For sake of this experiment, the elapsed time of clustering was computed and fixed at seven times iterations. The proposed clustering method will be investigated with evaluation performance factors like accuracy, v-measure (VM), adjusted rand index (ARI), completeness (COM), Fowlkes-Mallow's index (FMI), and Silhouette index (SI) and compared to k-means, agglomerative hierarchical, density-based spatial clustering of applications with noise (DBSCAN), affinity propagation, and mini batch k-means.

This article is organized into several sections. Section 2 presents the explanations of the related work. The preliminary study is presented in section 3. Section 4 provides the explanations and illustrations of the proposed algorithm and framework for clustering heterogeneous categorical data. Section 5 presents the discussion on the results of the comparisons that are made on the performance of the proposed solution and other clustering methods. Finally, section 6 provides the conclusion based on the findings and future work that can be done.

## 2. RELATED WORKS

The increasing attention to the study of categorical data similarity has raised concerns about the quality of analysis performance. Nominal, binary, ordinal, and Likert data types are considered categorical [8]. It is a challenge to analyze and interpret these categories of data, especially the Likert scale. For instance, ordinal features for Likert data assume that Likert items have the same meaning regardless of whether they are neutral or undecided, affecting the outcome or performance of reality research issues and may also cause biases [9]. Several attempts have been established to rectify these. Among these attempts are applied k-modes that use simple matching distance metrics to partition the datasets into many groups. However, the results gained have had a low intra-cluster similarity, and the starting points may lead to non-repetitive clustering patterns [10]. For example, k-modes using Hamming distance measure inaccurately differentiate the species as they create only one cluster [11] since the algorithm becomes unstable due to selecting the highest frequency of the data. The categorical data approach typically transforms a set of data into numerical ones by considering the relative frequency of the aspect. A review of categorical clustering data implied that the initialization of the centroid method had demonstrated promising results [12]. However, each feature had a hard category value in a hard centroid. This feature reflected the misclassified region [13]. In general, the Euclidean distance metric as a similarity between the data objects is not fully considered [14] and reduces the precision of the decision of the result. Regarding that issue, the clustering algorithm with entropy similarity

measure to investigate the weight of features including nominal, binary, ordinal, and Likert scale. Distance metric using entropy is a method that can be used to determine the weight of feature, and the information of entropy consider the uncertainty of all possible occurrence.

The concept of entropy is applied in many areas of studies, for example in studying customers' satisfaction. Surveys are used to gather data from a population. Individual replies to two or more questions are utilized to determine how a survey's scale is determined. Survey responses are combined into a single score by using a scoring system. This scale and entropy values can be used to get the expected information. It is also known as the second law of thermodynamics and the measurement of uncertainty. Information entropy, which is often known as Shannon entropy, was initially proposed by Shannon in 1948 [15]. This measurement of information entropy is subject to error. The higher the information entropy is, the lower the usefulness value of information is. On the other hand, the lower the information entropy is, the lower the uncertainty and the higher the information's value are. Thus, this research aims to utilize entropy distance measure with the clustering algorithms as it has the potential to improve the clustering solution mainly in the heterogeneous categorical data.

### 3. PRELIMINARIES

This section introduces and provides the preliminary concepts of clustering required. The choice of distance metrics is crucial since it has a significant influence on the effectiveness of the clustering. The explanations of preliminary observations made on the mini batch k-means algorithm concepts and entropy distance measure capability in providing the solutions to heterogeneous data clustering are discussed.

#### 3.1. Mini batch k-means (MBK) algorithm

The approach of the MBK clustering algorithm was adapted from [16]. The algorithm of mini batch k-means is stated in algorithm 1. The data was incrementally stored and updated using a distributed random batching approach called mini batch k-means. The data was stored and updated in a series of short batches. The cluster was updated using the data and prototype values in each batch. The more iterations in a batch, the greater the learning rate will be. Clusters must go through many iterations before they reach a consensus. It may be seen in several cases as the impact of new data decreases with the increasing iteration number. The greater the number of clusters is, the less similar the micro-batch is to a larger batch. MBK has several advantages. They have faster computation time, the most straightforward unsupervised learning that solves clustering techniques, and greater accuracy when working with mixed and large datasets [17]. However, the previous solution MBK has not yet been tested in heterogeneous categorical data.

#### Algorithm 1. Mini-batch K-means algorithm

Input:  $X$  is the similarity matrix  $\{x_1, x_2, \dots, x_n\}$ , in which  $n$  is the number of input values, mini-batch size ( $b$ ), iterations ( $t$ ), cluster number ( $k$ ), the total number of features ( $d$ ),  $d_{nom}$  feature are nominal features,  $d_{ord}$  the feature is ordinal features.

Output: Number of Clusters,  $C$

```

1. Start
2. Establish the initialization of  $k$  cluster nodes,  $\mu = \{\mu_{c_1}, \mu_{c_2}, \dots, \mu_{c_k}\}$ 
3. Create each cluster,  $C_i = \theta$  ( $1 \leq i \leq k$ )
4. Initialize the number of clusters with data.,  $N_{c_i} = (1 \leq i \leq k)$ 
5. for  $p = 1$  to  $t$  do:
6.  $M = \{x_m | 1 \leq m \leq b\}$  #  $M$  is the batch dataset, and  $x_m$  is a random sample from  $X$ 
7. for  $m = 1, 2, \dots, b$  do
8.    $\mu_{c_1}(x_m) = \frac{1}{|c_1|} \sum_{x_m \in c_1} x_m (x_m \in M)$ 
9.   end for
10.  for  $m = 1, 2, \dots, b$  do
11.     $\mu_{c_1} = \mu_{c_1}(x_m)$ 
12.     $N_{c_1} = N_{c_1} + 1$ 
13.     $\rho = \frac{1}{N_{c_1}}$ 
14.     $\mu_{c_1} = (1 - \rho) \mu_{c_1} + \rho x_m$ 
15.  end for
end for

```

#### 3.2. Entropy distance measure

Performing clustering requires reasonable distances between the attributes to obtain a meaningful cluster. In a clustering method, by default, the distance measures like Euclidean distance and Hamming distance are used in clustering methods such as hierarchical clustering. They perform well in most of the homogenous categorical data [18]. In heterogeneous data, the capability of entropy distances is offered.

Euclidean and Hamming distance measures have a drawback. They can only identify only one cluster at each iteration and often result in a cluster with weak intra-similarity [19]. It is evident that entropy distance measure is possible to improve the result of clustering in heterogeneous categorical data. This statement is supported by [7].

#### 4. PROPOSED CLUSTERING SOLUTION FOR HETEROGENEOUS CATEGORICAL DATA

This section provides the highlight of a proposed algorithm and the process flow for clustering heterogeneous categorical data using MBKEM. In heterogeneous categorical data, the entropy distance measure is applied to handle inadequate heterogeneous information of categorical data. The clustering process of the heterogeneous categorical data leads to information loss and eventually results in insufficient information for decision-making.

##### 4.1. MBKEM algorithm

In this research, we propose the enhancement of MBK using an embedded entropy distance measure to ascertain the quality of the performance of clustering using heterogeneous categorical data as indicated in algorithm 2. The entropy distance measure is expected to assist heterogeneous categorical information clustering capability in handling information loss. The algorithm is stated in algorithm 2 which is the MBKEM algorithm. The algorithm starts with the initialization cluster node, then it creates the clusters and initializes the number of clusters, as shown in steps 2 to 4. Steps 5 to 7 are the steps to determine the reliability of the features. This technique includes the heterogeneous information provided by the questionnaire's nominal and ordinal qualities. The next stage is the computation of the probabilities associated with each feature. The identification of weight for each feature is in steps 8 to 10. All attributes' reliability and total reliability, including nominal and ordinal data, are determined to allocate weights to features. Next, steps 11 to 25 are the processes to clarify the distance between two individuals. Then, the distance between each feature category is calculated using weights and entropies. Entropy is constructed using a dissimilarity matrix. The generation of the distance matrix is from the entropy of the choices made by the respondents. Step 27 is the step in which the sample is randomly selected. Steps 28 to 30 determine the cluster center for each sample in a batch set. In step 29, the cluster center that is the closest in proximity to the data sample is stored. Steps 31 to 36 are to synchronize each batch set with the cluster center, in step 32 is to obtain the cached central for  $x_m$ , step 33 is to determine the rate of learning for each cluster center, and the gradient step is to update the cluster center.

##### Algorithm 2. Enhanced mini-batch K-means with entropy measure

Input:  $X$  is the similarity matrix  $\{x_1, x_2, \dots, x_n\}$ , in which  $n$  is the number of input values, mini-batch size ( $b$ ), iterations ( $t$ ), cluster number ( $k$ ), the total number of features ( $d$ ),  $d_{nom}$  feature are nominal features,  $d_{ord}$  the feature is ordinal features.

Output: Number of Clusters,  $C$

1. Start
2. Establish the initialization of  $k$  cluster nodes,  $\mu = \{\mu_{c_1}, \mu_{c_2}, \dots, \mu_{c_k}\}$
3. Create each cluster.,  $C_i = \theta$  ( $1 \leq i \leq k$ )
4. Initialize the number of clusters with data.,  $N_c = (1 \leq i \leq k)$
5. for  $r = 1$  to  $d$  do
6.      $r = \frac{E_{O_r(s)}}{S}$
7. end for
8. for  $r = 1$  to  $d$  do
9.      $w = \frac{R}{\sum R}$
10. end for
11. for  $r = 1$  to  $d_{ord}$  do #  $d_{ord}$  for ordinal data distance,
12.     if  $i_r \neq j_r$  then
13.          $dist(O_r(i_r), O_r(j_r))^2 = w \cdot \sum_{s=(i_r, j_r)}^{(i_r, j_r)} E_{O_r(s)}$
14.     else
15.          $dist(O_r(i_r), O_r(j_r))^2 = 0$
16.     end if
17. end for
18. for  $r = d_{ord} + 1$  to  $d$  do #  $d_{ord} + 1$  for nominal data distance
19.     if  $i_r \neq j_r$  then
20.          $\varphi(O_r(i_r), O_r(j_r))^2 = w \cdot \sum_{s=i_r, j_r} E_{O_r(s)}$
21.     else
22.          $dist(O_r(i_r), O_r(j_r))^2 = 0$
23.     end if
24. end for
25. distance between two categories,  $Dist(x_i, x_j) = \sqrt{\sum_{r=1}^d dist(O_r(i_r), O_r(j_r))^2}$

```

26. for  $p=1$  to  $t$  do:
27.  $M = \{x_m | 1 \leq m \leq b\}$  #  $M$  is the batch dataset, and  $x_m$  is a random sample from  $X$ 
28. for  $m = 1, 2, \dots, b$  do
29.    $\mu_{c_1}(x_m) = \frac{1}{|c_1|} \sum_{x_m \in c_1} x_m (x_m \in M)$ 
30.   end for
31.   for  $m = 1, 2, \dots, b$  do
32.      $\mu_{c_1} = \mu_{c_1}(x_m)$ 
33.      $N_{c_1} = N_{c_1} + 1$ 
34.      $\rho = \frac{1}{N_{c_1}}$ 
35.      $\mu_{c_1} = (1 - \rho) \mu_{c_1} + \rho x_m$ 
36.   end for
end for

```

#### 4.2. Process flow for clustering heterogeneous categorical data

The process flow provides a rough indication of the actions to take from the data pre-processing until the computational result. Figure 1 demonstrates the process flow for clustering heterogeneous categorical data using MBKEM. The process flow is divided into three phases: phase 1, phase 2, and phase 3. The detailed explanations of each phase are presented in the following subsections.

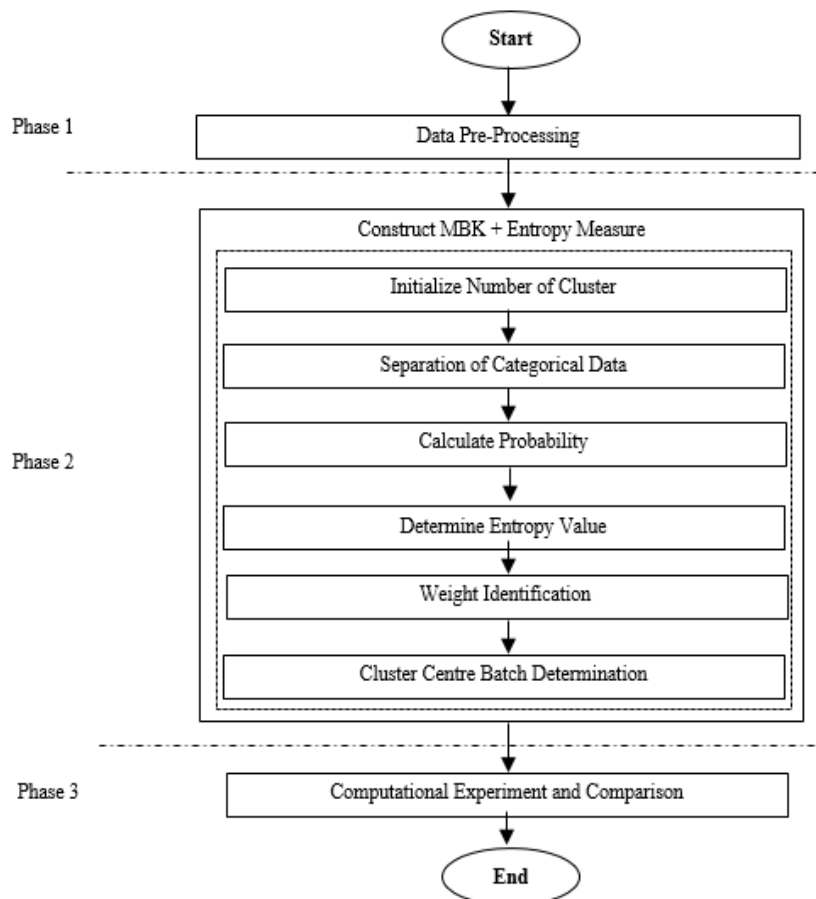


Figure 1. Process flow for clustering heterogeneous categorical data using MBKEM

##### 4.2.1. Phase 1

Phase 1 involves the pre-processing of the existing survey data. The data pre-processing step is one of the preliminary steps that can be performed during the cluster analysis process. It involves analyzing the data to transform it into an appropriate format for analysis. In this phase, the steps include imputing missing values, fixing data structure entry, and removing unwanted observations. The data collected from a questionnaire is usually stored as strings.

#### 4.2.2. Phase 2

Phase 2 involves the construction of MBKEM, and there are six stages involved in this. It starts with initializing the number of clusters, separating categorical data, calculating probability, determining entropy values, identifying weight, and determining cluster center batch. The categorical features are separated and located in the same categories. The measure of occurrence probability value of  $O_{r(s)}$  in feature,  $F_r$  is defined in (1).

$$p_{O_{r(s)}} = \frac{\sigma O_{r(s)}}{N} \quad (1)$$

In which,  $\sigma O_{r(s)}$  is the number of data objects in the dataset with the  $r$ th values equal to  $O_{r(s)}$ . Shannon's entropy is used to evaluate the information of entropy. The entropy is to identify the starting point of the mini batch k-means clustering to determine the weight and decide from a collection of options. Shannon entropy is a straightforward measure of uncertainty in a dataset, as stated in (2). The entropy values of categories  $O_{r(s)}$  in features,  $F_r$  is written as (2).

$$E_{O_{r(s)}} = -p_{O_{r(s)}} \log p_{O_{r(s)}} \quad (2)$$

$p_{O_{r(s)}}$  is the occurrence probability of value  $O_{r(s)}$ . The entropy value indicates the smaller value of entropy, the more typical behavior, and uncertainty. Weighing is a crucial component of entropy information. According to the information theory, the greater the use-value of an individual feature as quantified by its entropy value is, the more relevant or information-rich the judgment will be [20]. The ability to comprehend decisions between two alternatives (e.g., Likert scales) is based on more important information and provides a more convincing result. Consequently, the significance of weight features is proportional to the amount of information provided. The weighting factor used to determine the relative relevance is depicted in (3).

$$w_{F_r}^I = \frac{E_{O_{r(s)}}}{\sum_{s=1}^d E_{F_s}} \quad (3)$$

As the number of categories increases, the feature may generate longer distances and contribute more than necessary; hence, it must be appropriately weighed. As shown in (4) is the formula of the weighting scale for features.

$$w_{F_r}^S = \frac{\frac{1}{S_{F_r}}}{\sum_{s=1}^d \frac{1}{S_{F_s}}} \quad (4)$$

$S_{F_r}$  is the maximum entropy of a feature in which each category is likely to occur equally. As mentioned earlier, the combined weighting of the two weights is denoted by (3). The element of  $S_{F_r}$  is defined in (5).

$$S_{F_r} = -\log \frac{1}{v_r} \quad (5)$$

The weight for magnitude and scale of features is written in (6). As shown in (7) is the formula to find the reliability of features,  $R_{A_r}$

$$w_{F_r} = \frac{R_{F_r}}{\sum_{s=1}^d R_{F_s}} \quad (6)$$

$$R_{F_r} = \frac{E_{F_r}}{S_{A_r}} \quad (7)$$

In which  $R_{F_r}$  is the reliability value that indicates the proportion of the maximum volume of information stored in features,  $F_r$ . The value of  $R_{F_r}$  indicates that the greater the value of reliability is, the more convincing the distance is.

#### 4.2.3. Phase 3

In phase 3, a series of experiments on MBKEM is performed using performance measures of clustering validation. Clustering validation refers to finding the optimal clusters to match the partition of clusters naturally without the need for class information. There are six parameters to measure clustering quality. They are clustering accuracy (CA), external validation (COM), VM, ARI, and FMI, internal

validation using SI and elapsed time. CA quantifies the proportion of clustered data objects that are successfully clustered. It conveys the precision with which the results are obtained. CA values are more significant than one implying improved clustering capability and precision [21]. The cluster  $C$  is partitioned into a set of clusters  $\{c_1, c_2, \dots, c_k\}$  on dataset,  $O$  with  $n$  number of objects and the formula is estimated in (8):

$$CA = \frac{\sum_{l=1}^k C_l}{|O|} \quad (8)$$

where  $k$  is the number of clusters desired,  $C_l$  is the number of objects and  $|O|$  is equal to the number of objects in the dataset. External validation is the process of evaluating the performance of clustering using prior knowledge like class labels such as COM, VM, and ARI [22]. Completeness is considered comprehensive if it incorporates all data points that belong to a given class. A score between 0.0 and 1.0 is obtained. A labeling score of 1.0 indicates perfect labeling. V-measure can be used to ascertain the degree of agreement between two clustered datasets that have been clustered independently. The formula of completeness in (9):

$$COM = 1 - \sum_{c,k} \frac{n_{ck}}{N} \log\left(\frac{n_{ck}}{n_k}\right) \quad (9)$$

In which  $n_{ck}$  is the ratio of the number of samples labeled in a cluster that has the same and the total number of samples. Meanwhile, the ARI is a more sophisticated form of the rand index (RI) that assesses the agreement between genuine and acquired labels in terms of their projected agreement. RI determines the degree of similarity between two clustering by evaluating all pairs of samples and counting those assigned to the same or different cluster in the predicted and actual clustering. There are no duplicate clusters ( $ARI=1$ ), and random labeling occurs regardless of the number of clusters ( $ARI=0$ ). The larger the ARI value is, the more influential the grouping will be. The formula of ARI in (10).

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_j \binom{c_i}{2} \sum_j \binom{d_j}{2}] / \binom{n}{2}}{\frac{1}{2} [\sum_j \binom{c_i}{2} + \sum_j \binom{d_j}{2}] - [\sum_j \binom{c_i}{2} \sum_j \binom{d_j}{2}] / \binom{n}{2}} \quad (10)$$

In which given a dataset with  $n$  objects, suppose  $U = \{u_1, u_2 \dots u_s\}$  and  $V = \{v_1, v_2 \dots v_s\}$  represent the original classes and the clustering results respectively.  $n_{ij}$  denotes the number of objects in a cluster  $u_i$  and cluster  $v_j$  respectively.  $c_i$  and  $d_j$  is the number of objects in class  $u_i$  and cluster  $v_j$ . FMI quantifies the performance of a clustering technique by comparing it to other clusters. A score close to zero indicates largely independent labeling, whereas a value close to one reflects clustering agreement. The formula for FMI is defined in (11).

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \quad (11)$$

In which  $TP$  are pairs of observations of the same cluster,  $FP$  is pairs of observations of the same cluster but different in the predicted cluster, and  $FN$  is pairs of observations that are not part of the same cluster but same in a predicted cluster. Internal validation examines clusters generated by the clustering algorithm only by comparing the data. The Silhouette method is a well-known internal measure that estimates cluster-related parameter consistency. This method quantifies the similarity of items to their cluster (cohesion) concerning other clusters (separation). The optimum value is one. Near-zero values indicate the presence of overlapping clusters. Negative values frequently suggest that a sample is incorrectly assigned to a cluster because another cluster is more similar to the sample [23]. The formula of SI is defined in (12).

$$SI = \frac{(b_i - a_i)}{\max(a_i, b_i)} \quad (12)$$

In which  $a_i$  is the distance mean in the cluster and  $b_i$  is the minimum average distance to points in another cluster.

## 5. RESULTS AND DISCUSSION

This research aims to evaluate the impact of the proposed techniques compared to the existing conventional clustering techniques. An entropy measure as a distance metric is proposed to minimize the distance value within a cluster. As a result, it offers a way to improve the quality of clustering tasks. The

primary aim of this research is to enhance the quality of clustering through conventional clustering techniques with entropy measures. This section presents the results and discussions on the computational experiments to measure the quality of clustering over the existing methods and for setting the ideal number of clusters.

### 5.1. Dataset

In conducting the experiments, heterogeneous datasets were taken from the secondary data of a survey on public timber utilization. The public was the end-users who used timber. The total of an unlabeled public dataset was 2,407 respondents. The variable observations used to analyze the public perception of timber usage were 74 qualitative features that included five nominal features, 30 binary features, four ordinal features, and 35 Likert scales. Binary was the nominal feature, and Likert scales were the ordinal feature. The data pre-processing and cleaning procedures such as removing unwanted observations, fixing the data structure, and imputing the missing data were applied.

### 5.2. Computational result

The method was developed using the Python programming language. A computational experiment comparison is tested in this section to validate the impact of proposed solutions on the existing method. In this case, CA, external validation COM, VM, ARI, and FMI, internal validation using SI, and elapsed time are considered for evaluation. The following subsections provide concise results on the quality of the clustering and proposed clustering methods.

#### 5.2.1. Clustering accuracy

Figure 2 shows the findings of the comparison made on the accuracy of the algorithm clustering in MBKEM. It shows that the clustering accuracy is 88.1%. The accuracy result indicates that the proposed algorithm is more accurate and capable of convergence.

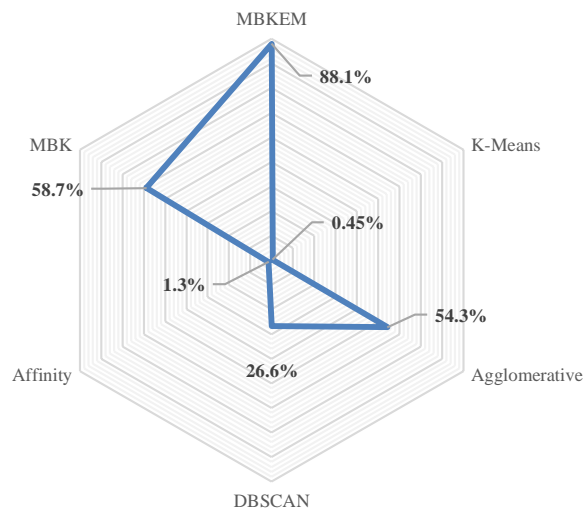


Figure 2. Comparison of accuracy score of clustering algorithm

#### 5.2.2. External evaluation

This section explains performance measures on external evaluation. Figure 3 shows the comparative performance of external evaluation with varying  $k$  values using MBKEM algorithm in Figure 3(a), k-means algorithm in Figure 3(b), agglomerative algorithm in Figure 3(c), DBSCAN algorithm in Figure 3(d), affinity propagation in Figure 3(e) and MBK in Figure 3(f). From the figure, it demonstrated that the proposed MBKEM has shown the highest performance in the VM, COM, ARI, and FMI. VM is at 0.82, C is at 0.81, ARI is at 0.87, and FMI is 0.94 at  $k = 2$ , indicating that the two partitions are nearly aligned, more similar, and flourishing a clustering algorithm.

Usually, data categorization is influenced by the choice of unsupervised clustering. Unlike other clustering algorithms, DBSCAN and affinity propagation do not require the number of clusters as a parameter. The external performance evaluation of DBSCAN and affinity propagation influence the preference parameter and damping factor. However, DBSCAN and affinity propagation have a significant



benefit which that they do not require information on cluster data when clustering. Most of the external performance evaluation scores from MBKEM, k-means, agglomerative, and MBK algorithms have shown the best performance score of VM, COM, ARI, and FMI are at  $k = 2$ .

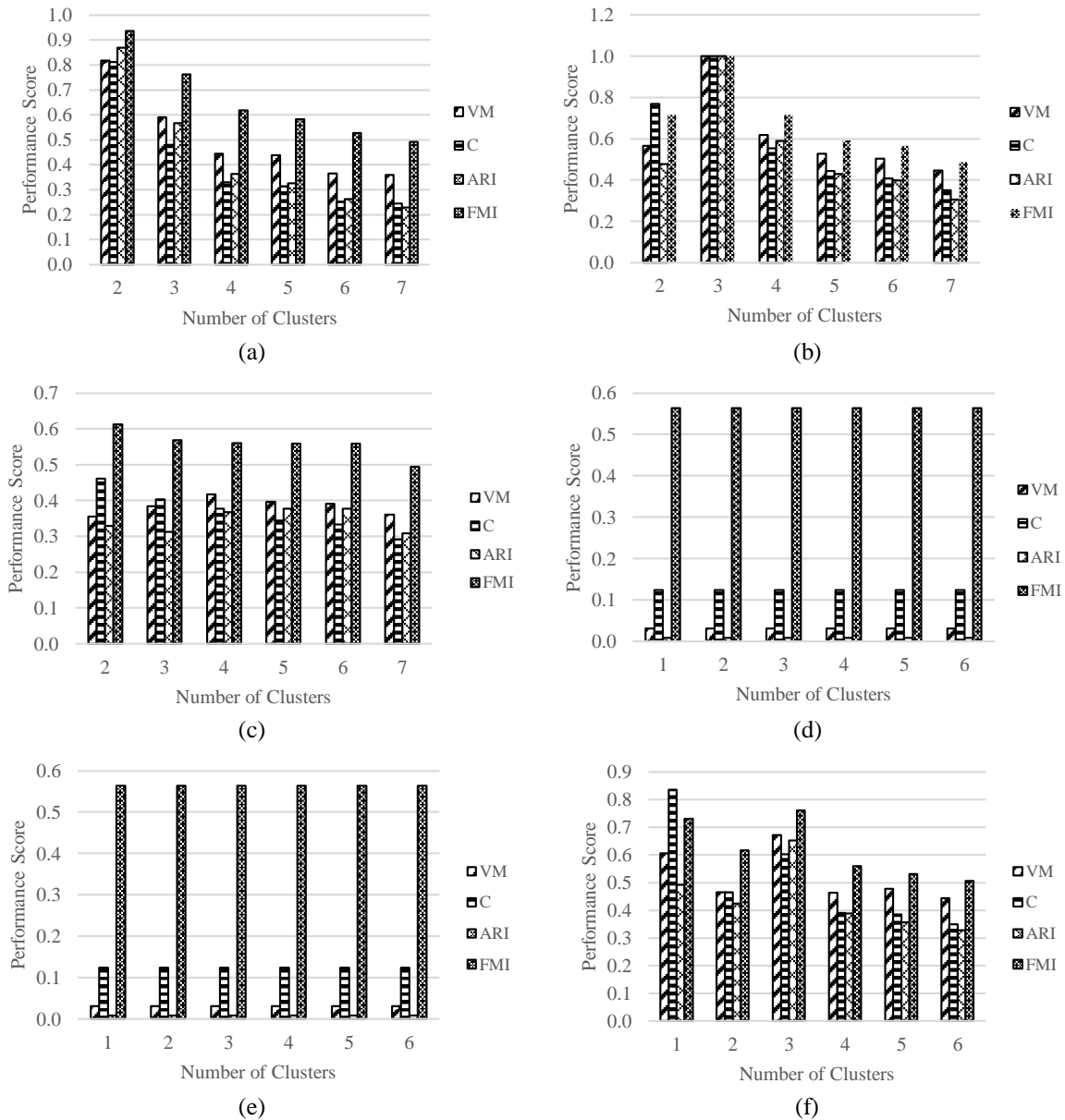


Figure 3. Comparing external evaluation performance using different clustering techniques such as (a) MBKEM algorithm, (b) k-means algorithm, (c) agglomerative algorithm, (d) DBSCAN algorithm, (e) affinity propagation, and (f) MBK algorithm

**5.2.3. Internal evaluation**

Figure 4 shows the comparison of internal evaluation using the SI at different numbers of clusters ( $k$ ). Based on Figure 2, the proposed MBKEM shows the highest performance for SI compared to other clustering algorithms. SI is one of the best indicators for estimating the formation of clusters. The result shows that all clusters are in the right cluster since all values of SI are positive. Figure 3 indicates that the proposed MBKEM gains the best result at  $k=2$ . The result of SI is due to the silhouette value being at the highest. As the number of clusters increases, the importance of SI decreases. The value of SI near 1 reveals a good value.

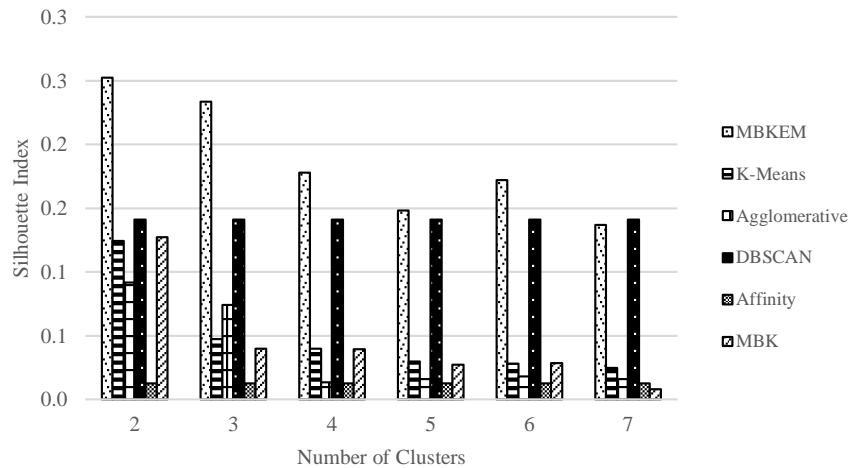


Figure 4. Comparison of silhouette index

**5.3. Computational elapsed time**

The elapsed time is defined as the time required to cluster the data. The quality of the cluster is identified as the lesser amount of time is taken, the better quality the cluster will be. For different numbers of clusters ( $k$ ), elapsed time of the MBKEM, k-means, agglomerative, DBSCAN, affinity, and MBK is visualized in Table 1. The elapsed time of the MBK algorithm is less than others at seven times iteration. Hence, MBK has revealed the best clustering algorithm with minimum computational time. However, the proposed MBKEM still shows the minimum computational time compared to K-means, agglomerative, DBSCAN, and affinity propagation.

Table 1. Elapsed time (seconds)

Executions	MBKEM	K-Means	Agglomerative	DBSCAN	Affinity	MBK
1	0.26	0.28	0.78	2.07	14.78	0.12
2	0.27	0.29	0.79	2.06	14.62	0.12
3	0.26	0.30	0.77	2.07	14.54	0.12
4	0.26	0.28	0.77	2.07	15.04	0.12
5	0.26	0.28	0.77	2.07	14.88	0.12
6	0.27	0.28	0.78	2.08	15.00	0.16
7	0.26	0.29	0.77	2.07	14.87	0.13

**5.4. Discussion**

Most of the experiments conducted for MBKEM have provided better results than k-means, agglomerative, DBSCAN, affinity propagation, and MBK. MBKEM utilizing the entropy distance measure has mainly brought a significant accuracy improvement. The nature of the entropy computation method can examine the degree of harmony or degree of consistency in a data group. Each feature in each category of the entropy measure is treated differently. The concept of entropy itself indicates the information stored on the entropy values. The entropy is associated with the weight of thinking cost and decision making. The weight of entropy analysis for each feature indicates the probability of a choice and can decide on a set of alternatives. The higher the weight is, the higher the value of variations is. Previous studies also supported that the entropy distance measure provides better performance and has higher accuracy for categorical datasets due to producing a weighted class of each data in a dataset [24]. Previous studies indicate that using the entropy weight technique to evaluate decision-making has been effective [25]. The choice of thinking cost or decision-making implies the distance between two categorical data. The smaller the entropy value is, the lesser the information stored in the choice and the higher probability of selection will be.

The distance measure for other clustering algorithms employed in this study is the Euclidean distance measure. This measure is the standard distance measure and is mostly used in clustering. Focusing exclusively on the scale of similarity-based samples and some essential features may ignore the data. The structures of ordinal features cannot represent the distance using the Euclidean distance measure. Therefore, the dynamic core structure of information in data cannot be represented, leading to non-optimal results. The non-optimal consequence will affect the performance. The clustering process must be accurate with low complexity to guarantee efficiency.

The limitation of MBKEM is not on the speediest computational time than in the traditional MBK. This occurrence could be due to each feature that is deeply treated and handled batch by batch. There are pros and cons to this. Overall, MBKEM still consumes less computational cost than k-means, agglomerative, DBSCAN, and affinity propagation. Interestingly, the proposed MBKEM that embeds entropy distance measure could be a new clustering method variant, especially in heterogeneous categorical data. Thus, enhancing the clustering algorithm is still required to improve the solution's effectiveness and be tested in many problem domains.

## 6. CONCLUSION

Clustering categorical data is complex since the values have no inherent order. A good metric to analyze a questionnaire is the entropy distance metric. This research has demonstrated a novel framework based on heterogeneous categorical data of unsupervised clustering methodologies using entropy distance measure as a similarity. Using the similarity measure of entropy enables the information for each data feature to be considered. The MBKEM framework has been proposed and evaluated using heterogeneous categorical data. The performance evaluation metric and time complexity have been investigated, and comparisons with other algorithms have been made to prove the effectiveness of the proposed method. MBKEM has mostly outperformed the other unsupervised clustering methods. This new idea of clustering heterogeneous categorical data solution has demonstrated that the evaluation outperformed in terms of CA, VM, COM, ARI, FMI, and SI at  $k=2$ . The execution time of the MBKEM is a bit higher than conventional K-means since each feature is comprehensively treated. Hence, MBKEM's can be improved with the swarm intelligence method in increasing the grouping performance of the heterogeneous categorical data. We believe this framework can be a valuable basis for another relevant research.

## ACKNOWLEDGEMENTS

The authors would like to acknowledge the Ministry of Higher Education (Fundamental Research Grant Scheme (FRGS) Grant: 600-IRMI/FRGS 5/3 (213/2019)) and Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA (UiTM) for the financial support provided for this research project.




## REFERENCES

- [1] R. Yang, K. Honda, S. Ubukata, and A. Notsu, "A comparative study on questionnaire design for categorization based on fuzzy co-clustering concept and multi-view possibilistic partition," in *2019 International Conference on Fuzzy Theory and Its Applications (iFUZZY)*, Nov. 2019, pp. 1–4, doi: 10.1109/iFUZZY46984.2019.9066247.
- [2] T.-H. T. Nguyen, D.-T. Dinh, S. Sriboonchitta, and V.-N. Huynh, "A method for k-means-like clustering of categorical data," *Journal of Ambient Intelligence and Humanized Computing*, Sep. 2019, doi: 10.1007/s12652-019-01445-5.
- [3] G. Yamini and B. R. Devi, "A new hybrid clustering technique based on mini-batch K-means and k-means++ for analysing big data.," *International Journal of Recent Research Aspects*, pp. 203–208, 2018.
- [4] D. Cuesta-Frau, A. Molina-Picó, B. Vargas, and P. González, "Permutation entropy: enhancing discriminating power by using relative frequencies vector of ordinal patterns instead of their Shannon entropy," *Entropy*, vol. 21, no. 10, Oct. 2019, doi: 10.3390/e21101013.
- [5] H. Jia and Y. M. Cheung, "subspace clustering of categorical and numerical data with an unknown number of clusters," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3308–3325, Aug. 2018, doi: 10.1109/TNNLS.2017.2728138.
- [6] D.-W. Kim, K. H. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern Recognition Letters*, vol. 25, no. 11, pp. 1263–1271, Aug. 2004, doi: 10.1016/j.patrec.2004.04.004.
- [7] Y. Zhang, Y.-M. Cheung, and K. C. Tan, "A unified entropy-based distance metric for ordinal-and-nominal-attribute data clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 1, pp. 39–52, Jan. 2020, doi: 10.1109/TNNLS.2019.2899381.
- [8] H. Wu and S.-O. Leung, "Can likert scales be treated as interval scales?—A simulation study," *Journal of Social Service Research*, vol. 43, no. 4, pp. 527–532, Aug. 2017, doi: 10.1080/01488376.2017.1329775.
- [9] B. P. Subedi, "Using likert type data in social science research: confusion, issues and challenges," *International Journal of Contemporary Applied Sciences*, vol. 3, no. 2, pp. 1365–2308, 2016.
- [10] N. Sowmiya and B. Valarmathi, "A review of categorical data clustering methodologies based on recent studies," *IIOAB Journal*, vol. 8, no. 2, pp. 353–365, 2017.
- [11] N. A. Hamzah, S. L. Kek, and S. Saharan, "The performance of K-means and K-modes clustering to identify cluster in numerical data," *Journal of Science and Technology*, vol. 9, no. 3, pp. 25–32, 2017.
- [12] R. K. Brouwer, "A method for fuzzy clustering with ordinal attributes," *International Journal of Intelligent Systems*, vol. 22, no. 6, pp. 599–620, Jun. 2007, doi: 10.1002/int.20216.
- [13] W. Jiakai and G. Ruijun, "An extended fuzzy k-means algorithm for clustering categorical valued data," in *2010 International Conference on Artificial Intelligence and Computational Intelligence*, Oct. 2010, vol. 2, pp. 504–507, doi: 10.1109/AICI.2010.225.
- [14] L. Wang, Y. Zhang, and J. Feng, "On the Euclidean distance of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1334–1339, Aug. 2005, doi: 10.1109/TPAMI.2005.165.




- [15] S. Verdu, "Fifty years of Shannon theory," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2057–2078, 1998, doi: 10.1109/18.720531.
- [16] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th international conference on World wide web - WWW '10*, 2010, pp. 1177–1178, doi: 10.1145/1772690.1772862.
- [17] M. M. Chavan, A. Patil, L. Dalvi, and A. Patil, "Mini batch K-means clustering on large dataset," *International Journal of Scientific Engineering and Technology Research*, vol. 04, no. 07, pp. 1356–1358, 2015.
- [18] J. Cibulková, Z. Šulc, S. Sirota, and H. Režanková, "The effect of binary data transformation in categorical data clustering," *Statistics in Transition New Series*, vol. 20, no. 2, pp. 33–47, Jul. 2019, doi: 10.21307/stattrans-2019-013.
- [19] P. Zhang, X. Wang, and P. X. K. Song, "Clustering categorical data based on distance vectors," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 355–367, Mar. 2006, doi: 10.1198/016214505000000312.
- [20] R. Lu, H. Shen, Z. Feng, H. Li, W. Zhao, and X. Li, "HTDet: A clustering method using information entropy for hardware Trojan detection," *Tsinghua Science and Technology*, vol. 26, no. 1, pp. 48–61, Feb. 2021, doi: 10.26599/TST.2019.9010047.
- [21] P. Manivannan and P. I. Devi, "Dengue fever prediction using K-means clustering algorithm," in *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, Mar. 2017, pp. 1–5, doi: 10.1109/ITCOSP.2017.8303126.
- [22] U. Kokate, A. Deshpande, P. Mahalle, and P. Patil, "Data stream clustering techniques, applications, and models: comparative analysis and discussion," *Big Data and Cognitive Computing*, vol. 2, no. 4, Oct. 2018, doi: 10.3390/bdcc2040032.
- [23] A. M. Coroiu, R. D. Găceanu, and H. F. Pop, "Discovering patterns in data using ordinal data analysis," *Informatica*, vol. LXI, no. 1, pp. 78–91, 2016.
- [24] S. Sharma and S. Pemo, "Performance analysis of various entropy measures in categorical data clustering," in *2020 International Conference on Computational Performance Evaluation (ComPE)*, 2020, pp. 592–595, doi: 10.1109/ComPE49325.2020.9200074.
- [25] Y. Zhu, D. Tian, and F. Yan, "Effectiveness of entropy weight method in decision-making," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–5, Mar. 2020, doi: 10.1155/2020/3564835.

## BIOGRAPHIES OF AUTHORS






**Nurshazwani Muhamad Mahfuz**    is currently a Ph.D. student in Information Technology at Universiti Teknologi MARA (UiTM) Shah Alam. Her educational background consists of a Bachelor of Science (Statistics) (Hons) at Universiti Teknologi MARA (UiTM), Malaysia. She continued her education in Master of Applied Statistics at Universiti Teknologi MARA (UiTM), Malaysia. Her current research interests include data mining, clustering, statistics, multi-view learning, artificial intelligence, optimization, and their application in timber. She can be contacted at 2017373123@student.uitm.edu.my.



**Marina Yusoff**    is currently a senior fellow researcher at the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI) and Associate Professor of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Malaysia. She has a Ph.D. in Information Technology and Quantitative Sciences (Intelligent Systems). She previously worked as a Senior Executive of Information Technology in SIRIM Berhad, Malaysia. She is most interested in multidisciplinary research, artificial intelligence, nature-inspired computing optimization, and data analytics. She applied and modified AI methods in many research and projects, including deep learning, neural network, particle swarm optimization, genetic algorithm, ant colony, and cuckoo search for many real-world problems and industrial projects. Her recent projects are data analytic optimizer, audio, and image pattern recognition. She has many impact journal publications and contributes as an examiner and reviewer to many conferences, journals, and universities' academic activities. She can be contacted at marina998@uitm.edu.my.



**Zainura Idrus**    obtained her Ph.D. in Computer Science from the Faculty of Computer and Mathematical Sciences, UiTM Malaysia. Her main research interests are data visualization, machine learning and computer support collaborative work (CSCW). She has served as an invited reviewer. She has also published research papers in journals, proceeding as well as chapters of books. She can be contacted at zainura@fskm.uitm.edu.my.