Firefly Algorithm with Mini Batch K-Means Entropy Measure for Clustering Heterogeneous Categorical Timber Data

Nurshazwani Muhamad Mahfuz[1]
Faculty of Computer and Mathematical Science
Universiti Teknologi MARA
Shah Alam, Malaysia

Muhammad Shaiful Nordin[3]
Malaysian Timber Industry Board
Menara PGRM, Jalan Pudu Ulu, 56100 Cheras
Kuala Lumpur, Malaysia

Marina Yusoff [2]
Institute for Big Data Analytics and Artificial Intelligence
Faculty of Computer and Mathematical Science
Universiti Teknologi MARA, Shah Alam, Malaysia

Zakiah Ahmad[4]
College of Engineering
Universiti Teknologi MARA
Shah Alam, Malaysia

***Abstract*—Clustering analysis is the process of identifying similar patterns in various types of data. Heterogeneous categorical data consists of data on ordinal, nominal, binary, and Likert scales. The clustering solution for heterogeneous data clustering remains difficult due to partitioning complex and dissimilarity features. It is necessary to find a solution to high-quality clustering techniques to efficiently determine the significant features of the data. This paper emphasizes using the firefly algorithm to reduce the distance gap between features and improve clustering performance. To obtain an optimal global solution for clustering, we proposed a hybrid of mini-batch k-means (MBK) clustering-based entropy distance measures (EM) with a firefly optimization algorithm (FA). This study compares the performance of hybrid K-Means, Agglomerative, DBSCAN, and Affinity clustering models with EM and FA. The evaluation uses a variety of data from the timber perception survey dataset. In terms of performance, the proposed MBK+EM+FA has superior and most effective clustering. It achieves a higher accuracy of 96.3 percent, a 97 percent F-measure, a 98 percent precision, and a 97 percent recall. Other external assessments revealed that the Homogeneity (HOMO) is 79.14 percent, the Fowlkes-Mallows Index (FMI) is 93.07 percent, the Completeness (COMP) is 78.04 percent, and the V-Measure (VM) is 78.58 percent. Both proposed MBK+EM+FA and MBK+EM took about 0.45s and 0.35s to compute, respectively. The excellent quality of the clustering results does not justify such time constraints. Surprisingly, the proposed model reduced the distance measure of all heterogeneous features. The future model could put heterogeneous categorical data from a different domain to the test.**

***Keywords*—Clustering; mini batch k-means; entropy; heterogeneous categorical; firefly optimization algorithm**

I. INTRODUCTION

Categorical data or known as qualitative data is a type of data that can be stored and identified using the names or labels that have been assigned to it. Most statistical analysis approaches are incompatible with it, and only bar graphs and pie charts can visualize the data. Due to the rapid emergence and growth of information, it has become increasingly important to discover the group structure of objects within them. It also has difficulty extracting helpful information. One of the most effective methods that can extract such information is clustering. This technique allows organizing the data to access the information. Using clustering techniques, data analysts can easily extract valuable information from large datasets without supervision. The categorical clustering techniques were widely used in many real-world applications such as security analytics [1][2] and solving cold-start recommendation problems [3].

The traditional methods commonly used for data clustering problems are hierarchical and partitional. In some cases, the clustering process relies on distance or similarity measures. In a clustering algorithm, the data objects are typically represented in Euclidean distance in *k*-means. The clustering process's objective is to minimize the square distance from the cluster center to the cluster domain. *k*-means is widely used as a clustering algorithm and is effective when dealing with enormous volumes of data. However, *k*-means cannot be directly applied to data sets with categorical features. The transformation and parameter adjustment into numerical form is required since machines cannot interpret the categorical features directly. Label encoding, one hot encoding, and dummy variable encoding are methods for converting categorical data into numerical data.

Meanwhile, *k*-modes using simple matching dissimilarity measures can directly be applied for purely categorical data clustering. As categorical data cannot be estimated using mean or medians, the Euclidean distance metric was replaced with a simple matching dissimilarity measure, and the mean calculation for representing centroids was substituted with modes. However, these methods have some disadvantages, such as empty groups may appear in the first step of the solution and the final division of data is not optimal due to the appearance of extreme points. It is also trapped in local optima and local maxima and is sensitive to initial cluster centers.

Nature-inspired algorithms have gained much attention as global optimization tools to assist in solving various real-life complex optimization problems such as profit production [4], scheduling [5], queuing system [6], market segmentation, and opinion mining [7]. Nature-inspired algorithms have received a lot of attention as global optimization tools to help with real-life complex optimization problems like profit production [4], scheduling [5], queuing systems [6], market segmentation, and opinion mining [7].

Stochastic methods can be used to overcome clustering problems. Optimization methods refer to finding feasible solutions for problems to give an efficient, robust solution. Recent research has used Artificial Bee Colony (ABC) [8] and Cuckoo Search [9] to improve categorical data clustering quality. Optimizing a complex clustering problem is difficult due to the lack of a single measure that works best for heterogeneous categorical data.

This paper proposes the hybrid firefly algorithm and MBK to improve clustering performance using entropy distance metrics as similarity measures. The experimental results were also compared using the entropy distance method with Mini Batch *k*-Means (MBK), *k*-Means, Agglomerative Hierarchical, Density-based spatial clustering of applications with noise (DBSCAN), and Affinity Propagation clustering methods on a public survey data.

The rest of the paper follows the organization of the section as follows. Section II describes the related works. Section III explains the firefly algorithm. The mini-batch k-means algorithm is in Section IV. The proposed algorithm is introduced in Section V. The experimental results are discussed in Section VI. Section VII and Section VIII are the discussion and conclusion of the paper.

II. Related Work

The related work on categorical clustering, FA, and clustering models that use metaheuristic methods are reviewed in this section. Nominal, binary, ordinal, and Likert data types are categorical [10]. The combination of these data types is simultaneously considered to have heterogeneous information or data. The main goal of data clustering is to predict and find the groups for each data object from unlabeled data. On the other hand, selecting appropriate representations for data is one of the central problems in machine learning and data mining.

Clustering categorical data is very challenging. The challenge includes processing non-categorical variables and a required procedure for applying the similarity measures for the matching process. Regarding the grouping data from previous studies, the set-valued *k*-modes algorithm outperforms three existing categorical clustering techniques and is scalable to big datasets [11]. Algorithm new *k*-means like a method for categorical clustering data has shown that the proposed clustering method outperformed the k-means [12]. A Unified Entropy-Based Distance Metric for ordinal-and-nominal attributes on both real and benchmark data sets shows that the proposed metric surpasses the existing alternatives. A Unified Entropy-Based Distance Metric for ordinal-and-nominal attributes on both real and benchmark data sets shows that the proposed metric surpasses the existing alternatives [13]. A holo-entropy-based hierarchical clustering technique for categorical data outperforms other known algorithms in terms of efficiency, accuracy, and reproducibility [14].

The linear programming model outperformed the traditional and other enhanced *k*-modes algorithms on categorical datasets [8]. Learning-Based dissimilarity for categorical data clustering outperforms in terms of several performance indicators [16]. Compared to *k*-modes, the *k*-Approximate Modal Haplotype achieves an average performance increase of 0.51 percent in Precision and 0.40 percent in Normalized Discounted Cumulative Gain. However, because of their scalability, *k*-modes are more flexible to utilize [3]. The conventional categorical clustering performed well but provided local optimum solutions that affect data division.

The firefly optimization algorithm is one of the techniques that has been effectively used to solve issues in several areas due to its global optimum solution, resilience, efficiency, and capacity to handle problems in various sectors, including NP-hard, versatility, and other outstanding benefits. A comprehensive overview of FA that covers the various domains where the method is applied to a wide range of real-world applications with satisfying clustering results. Regarding clustering validity metrics, the FA shows that the new clustering methods outperform existing clustering and other hybrid metaheuristic methodologies [17]. In both distance and performance measurements for clustering tasks, the inward intensified exploration fa and compound intensified exploration fa models show statistically significant superiority [18]. Both algorithms are proposed to overcome the limitation of the FA model and *k*-means clustering. Compared to other algorithms, the firefly algorithm with *k*-means clustering produces better results, demonstrating the usefulness of the firefly algorithm with *k*-means clustering in offering a competitive solution to the traveling salesman problem [19]. *k*-means clustering with modifying the firefly algorithm is significantly more efficient than other algorithms [20]. From this, it can be seen that the combination of FA and conventional clustering algorithms outperform them in many real-world data sets, including numerical data.

Furthermore, in the previous literature, some clustering techniques for categorical datasets combined with other hybridization of optimization have been investigated and outperform the traditional *k*-modes algorithm in several aspects [21], as well as Fuzzy *k*-partition based on the ABC, outperforms the baseline algorithm in terms of its validity clustering for categorical data clustering [8]. The ABC algorithm was inspired by the foraging habits of bees and is one of the swarm-based metaheuristic optimization algorithms. Several studies in categorical clustering for hybridization of global optimization using a conventional clustering algorithm. However, few studies on hybridizing global optimization algorithms and conventional clustering algorithms use entropy as a distance measure.

Therefore, in this research, MBK+EM+FA has proposed to find the optimum global result for heterogeneous categorical data clustering using entropy distance as a similarity measure.

Because of its automatic subdivision and capacity to deal with multimodality, FA is preferred above other algorithms [22].

III. Firefly Algorithm (FA)

Recently flashing behavior of fireflies has been identified as unique to the species. FA is established by Xin She Yang [15], and it was based on the idealized behavior of the flashing characteristics of fireflies. From the literature, it is found that the FA algorithm can outperform when compared to many other algorithms. FA algorithm expands, and new variants emerge to solve all optimization problems. FA is chosen instead of other algorithms due to its simple, flexible, fast convergence, which efficiently solves many real-world problems.

FA is a swarm-intelligence-based algorithm, so it has similar advantages to the other swarm intelligence-based algorithms. FA has two significant advantages over other algorithms: automatic subdivision and the ability to deal with multimodality. FA offers simplicity, flexibility, and ease of implementation. Recently, FA is one of the bio-inspired algorithms used to solve clustering problems inspired by biochemical and social aspects of real fireflies [19]. Fireflies' behaviors, such as short and rhythmic flashes, can be considered operators of computational intelligence methods. The basic assumptions formulation of the FA algorithm is as follows.

- The intensity of a firefly decreases with the increase in distance. The firefly attracts the firefly that is closer to it. The light intensity I decrease as the distance r^2 increases, $I \propto \frac{1}{r^2}$.
- All fireflies are unisex and attracted to other fireflies regardless of their sex.
- The objective function defines the brightness of a firefly.

The FA algorithm has two critical features: variation in the light intensity and formulation of attractiveness (Yang 2010). The flashes are used as a communication tool by fireflies. Light is absorbed in the media, so the attractiveness of two fireflies will vary with the degree of absorption. The light intensity, $I\,(r)$ varies to the inverse square law as stated in equation 1.

$$I\,(r) = \frac{I_s}{r^2} \tag{1}$$

The light intensity, I, and the absorption coefficient, γ, varies with the distance, r for a given as in equation 2.

$$I = I_0 e^{-\gamma r} \tag{2}$$

Where I_0 is the original of the light intensity. The combined effect of inverse square law and absorption can be assumed as Gaussian form and represented in equation 3.

$$I = I_0 e^{-\gamma r^2} \tag{3}$$

The attractiveness of fireflies is proportional to the light intensity of the adjacent fireflies. The attractiveness, β of firefly is in equation 4.

$$\beta(r) = \beta_0 e^{-\gamma r^2} \tag{4}$$

Where β_0 is attractiveness at $r = 0$, the Euclidean distance between the firefly, i at x_i, and firefly, j at x_j, is given by equation 5.

$$r_{ij} = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2} \tag{5}$$

The computation movement of firefly i is attracted to another more attractive, brighter firefly j, and the formula is in equation 6.

$$x_i = x_i + \beta_0 e^{-\gamma r_{ij}^2}(x_i - x_j) + \alpha\left(rand - \tfrac{1}{2}\right) \tag{6}$$

Where α is the randomization parameter and $rand$ is a random number generator uniformly distributed between 0 and 1. The second represents the attraction, and the third term is randomization.

Algorithm 1: The Pseudocode of FA

1. Start
2. Initialize algorithm parameters:
3. Define the objective function $f(x)$, where $x = (x_1, \ldots\ldots, x_d)$
4. Generate the initial population of fireflies x_i $(i = 1, 2, \ldots\ldots, n)$
5. Determine the light intensity of the firefly at x_i by using objective function $f(x_i)$
6. **While** (K<Maximum_Generation) // Where $k = 1$ to maximum
7. **for** $p = 1 : n$ // all n fireflies
8. **for** $q = 1$ to n //n fireflies
9. if ($I_p < I_q$)
   Move firefly p towards q by using
   $$x_i = x_i + \beta_0 e^{-\gamma r_{ij}^2}(x_i - x_j) + \alpha\left(rand - \frac{1}{2}\right)$$
10. **end** if
11. Calculate new solutions and update light intensity for the next iteration
12. **end** for q
13. **end** for p
14. Sort the fireflies based on the intensity value and find the current best solution.
15. end while

IV. Mini Batch K-Means (MBK) Clustering

A distributed random batching strategy known as Mini Batch K-means was used to store and update the data incrementally. Data and prototype values from each batch were used to update the cluster—the learning rate increases with the number of iterations in a batch. Before reaching a consensus, clusters must go through several iterations. There are several benefits to using MBK, such as its shorter computation time, the most simple unsupervised learning that solves clustering methodologies, and higher accuracy when dealing with mixed and huge datasets [23].

MBK+EM (Mini Batch K-Means with Entropy Measure) is an embedded entropy distance measure with Mini Batch K-Means that aims to determine the quality of the performance of clustering in heterogeneous categorical data. The previous experiment demonstrates that the Mini Batch K-mean with Entropy Measure at k =2 outperformed other clustering algorithms in clustering accuracy at 88%, V-Measure at 0.82, Adjusted Rand Index at 0.87, and Fowlkes-Mallow's Index at 0.94. The experiment was fixed seven times the average minimum elapsed time-varying for cluster generation, k at 0.26s.

V. Proposed Algorithm

The MBK algorithm has an advantage in reducing the amount of computation to converge to a local solution [24]. However, it has a drawback to finding the local optimum clustering results and the existing FA, which has a problem remembering the best solution for each firefly in the past. When there are no brighter-colored fireflies around, it also moves at random. The FA in this research performs the clustering procedure with the optimal centroid point. Therefore, the MBK is combined with the firefly optimization algorithm to obtain the global optimum solution and gives the firefly method a substantial improvement in clustering performance.

FA was chosen over other optimization algorithms because previous research shows that hybridizing FA with conventional clustering algorithms improves result quality [19]. The MBK+EM+FA algorithm combines the advantages of its distance measure, which would be possible to improve the clustering performance. One step of the MBK algorithm is utilized at the end of all the iterations of FA.

Fig. 1 depicts the process flow of the proposed hybridization FA with MBK+EM. The MBK+EM operator is incorporated into the FA to locate the centroid. The fireflies distribute randomly in the search space based on the objective function from the population at random from the given data objects. The distance between the firefly's position and the actual data in the dataset determines the intensity of each firefly. After analyzing the distance, all data shows the minimum distance value among the fireflies. The movement of the fireflies in the search space indicates the firefly's brightness. The iterative process of the swarm involves a comparison of the intensity of one firefly to another firefly, and the firefly's brightness defines its firefly movement. The attractiveness varies according to the distance between the two fireflies. Then, calculate the intensity for new fireflies based on firefly movement. To determine the new position, apply the MBK+EM operator to the entire population of fireflies. For each firefly, the MBK+EM operator is used to compute the mean value of all associated objects. Then, update the intensity values and evaluate the new firefly position for the entire firefly. Before proceeding to the next iteration, the fireflies' selection depends on their intensity value. The result is continuously updated during iterative processes until the stopping criteria are met. A post-process to select the final centroids for determining the best solution while ranking the position of the fireflies.

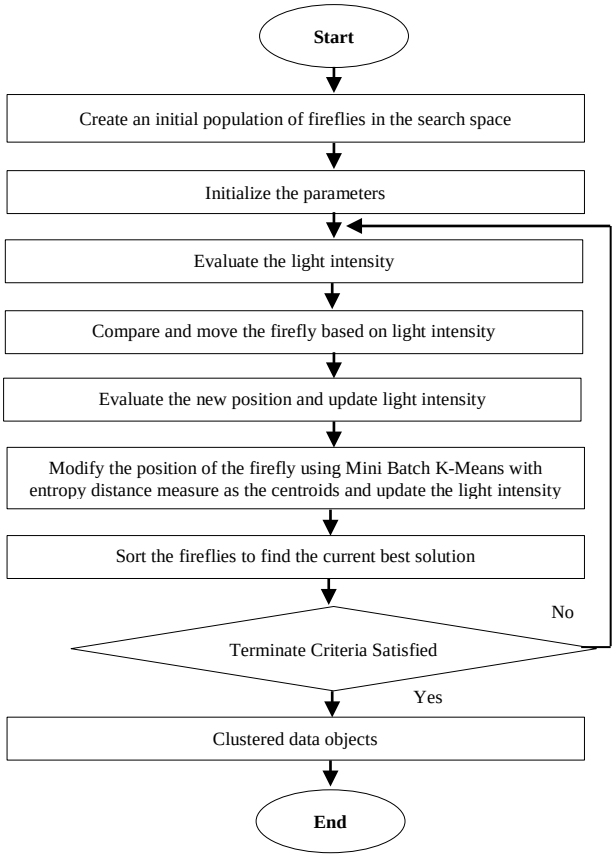

Fig. 1. Flowchart of Proposed MBK+EM+FA Algorithm.

The proposed Hybrid FA+MBK algorithm with EM phase is indicated in Algorithm 2. The following is a brief overview of all aspects of the Hybrid FA algorithm. The proposed algorithm starts by initializing the algorithm parameter in Step 2. The declaration of objective function performs in Step 3. The next step is randomly initiating fireflies' initialization. The position of the firefly represents the centroid of the clustering problem. Step 5 aims to determine the light intensity of the firefly to calculate the distance between the position of the fireflies. It is the initialization phase to estimate the light intensity of each firefly. Step 6-Step 10 demonstrates the movement of the fireflies in the search space, indicating the firefly's brightness. The intensity of one firefly compared with other fireflies during the iterative process in the swarm and the firefly's brightness defines its firefly movement.

The intensity value estimates the new position and place of the firefly obtained after the completion of movement calculation in steps 11 and 12. Global optima were used in the proposed firefly algorithm to control the movement of the firefly. This research will update a maximum global optimum in any iteration of the algorithm. Based on fireflies' brightness, fewer fireflies will move towards the brightest firefly. Then, the light intensity will be updated, and the current feasible or optimal solution will be found.

In steps 13 and 14, the MBK with EM is applied to the entire population of the fireflies to find a new position by updating light intensity. In the proposed FA, entropy distance was used to compute the distance of fireflies to global optima.

Step 15 focuses on the optimal values in each cluster that could discover after all the data are clustered by sorting the light intensity. Finally, the iteration executes until the maximum number of iterations.

Algorithm 2: The Pseudocode of MBK+EM+FA

1. **Input**
2. Initialize algorithm parameters
3. Define the objective function $f(x)$, where $x = (x_1, \ldots\ldots, x_d)$
4. Generate the initial population of fireflies x_i $(i = 1, 2, \ldots\ldots, n)$
5. Determine the light intensity of the firefly at x_i by using objective function $f(x_i)$
6. **While** (K <Maximum_Generation) // Where $k = 1$ to maximum
7. **for** $p = 1 : n$ // all n fireflies
8. **for** $q = 1$ to n //n fireflies
9. if ($I_p < I_q$)
   Move firefly p towards q by using
   $$x_i' = x_i + \beta_0 e^{-\gamma r_{ij}^2} + \propto \left(rand - \frac{1}{2}\right)$$
10. **end** if
11. Calculate new solutions and update light intensity
12. **end** for q
13. Apply MBK with entropy distance measure, then find new solutions and update light intensity
14. **end** for p
15. Sort the fireflies based on the intensity value and find the current best solution;
16. end while
17. **Output**: Clustered data objects

VI. Experimental Results

A. Dataset

The experiment was performed using the secondary data of a survey on public timber utilization. The data pre-processing and cleaning procedures included removing unwanted observations, fixing the data structure, and imputing the missing data. The number of instances is 2407 was obtained from the Malaysian Timber Industry Board (MTIB), Malaysia. The dataset consists of 111 categorical features such as race, type of housing, level of knowledge, etc. This type of data is considered heterogeneous categorical data [25].

B. Parameter Settings

This subsection explains the parameter, notation, and value associated with FA evaluation. We adapt the same parameter values as suggested by the originator of FA in the late year between 2007 and 2008 [26]. The population size is 2407, the total number of respondents involved. As a stopping criterion, we also set the maximum number of iterations equal to 25. Rosenbrock is used as a benchmark objective function since previous research has shown that Rosenbrock is one of the objective functions that has shown successful performance [27]. This research implemented a static parameter for the generation and number of fireflies. The maximum value for Beta (β), Alpha (α), and Gamma (γ) is 1, 0.2, and, respectively. The algorithm parameters are summarized in Table I.

TABLE I. Parameter Setting

Parameter	Notation of Parameter	Parameter Value
Brightness	Objective Function	Rosenbrock
Beta (β)	Attractiveness	1
Alpha (α)	Randomization of Parameter	0.2
Gamma (γ)	Light Absorption Coefficient	1
Number of Fireflies (n)	Population	2407
Number of Generations (g)	Iteration	25

C. Performance Measures

An external validation measure and a confusion matrix and used for the performance measure. The external validation measure is Homogeneity (HOMO), Fowlkes-Mallows Index (FMI), Completeness (COMP), and V-Measure (VM). The validation assures how good the clustering solutions are by their different ways of computations. This measure aims to measure and validate the clustering quality [28].

In the confusion matrix [16], there were four performance metrics such as true positive (TP), false positive (FP), true negative (TN), and false-negative (FN). TP is the metric that could accurately predict the optimized feature from the features in feature space collection. In contrast, the TN metric predicts the weaker feature or incorrect feature relevant to the diabetes classification process in feature space, and the FP measure can predict the incorrect diabetes feature in feature space. FN rate is an outcome of the model incorrectly predicting the negative feature effectively.

F-measure (F) is a combination of precision and recall that measures the cluster that contains only objects of a particular class and is used to balance false negatives by weighting recall parameter $\eta \geq 0$. The formula of the F-measure is in Equation 7.

$$F = \frac{(\eta^2+1)\, P \times R}{\eta^2 \times P + R} \tag{7}$$

Precision (P) estimates the ratio of the true positives among the cluster. The formula of Precision is in Equation 8.

$$P = \frac{TP}{TP+FP} \tag{8}$$

Recall (R) is a combination of all objects grouped into a specific class. The formula of recall is in Equation 9.

$$R = \frac{TP}{TP+FN} \tag{9}$$

Fowlkes-Mallows Index (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 in Equation 10.

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \tag{10}$$

Accuracy (ACC) is defined only as the proportion of the actual results. The accuracy measure can be referred to in Equation 11.

$$ACC = \frac{1}{n}\sum_{i=1}^{k} a_i \quad (11)$$

Where a_i is the number of data objects in both clusters, i and k are the numbers of clusters, and n is the total number of objects in the dataset. HOMO covers all clusters that contain only data points members of a single class. A score between 0.0 and 1.0 is obtained. A score of 1.0 stands for perfectly homogeneous labeling.

$$HOMO = -\sum_{c,k}^{1} \frac{n_{ck}}{N} \; log\; log\; (\frac{n_{ck}}{n_k}) \quad (12)$$

Completeness (COMP) 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 defined in Equation 13.

$$COMP = 1 - \sum_{c,k}^{1} \frac{n_{ck}}{N} \; log\; log\; (\frac{n_{ck}}{n_k}) \quad (13)$$

V-Measure (VM) is the harmonic average between homogeneity and completeness. It can be used to determine the degree of agreement between two clustered datasets that have been clustered independently. Furthermore, if any of the two VM failed to meet the criteria, the clustering number remains zero.

$$V - Measure\ (VM) = 2 \times \frac{(HOMO \times COMP)}{(HOMO + COMP)} \quad (14)$$

D. Experimental Results

It has been explained earlier than due to local optimum issue which aimed to improve performance of clustering external validity using heterogeneous categorical data. Several clustering algorithms were accelerated and tested with the coincident clustering problem to produce good clustering results. Moreover, the categorical nature of data creates additional complexity in clustering. MBK + EM + FA has been proposed to address such limitations of existing clustering algorithms. The technique's efficiency can be evaluated by measuring the quality of clustering results of various parameters. In order to judge the performance of the proposed technique over state-of-the-art algorithms of the other four hybrid clustering algorithms, such as a hybrid of *k*-means, Agglomerative, DBSCAN, and Affinity with EM distance metric, we conduct several experiments. Thus, this section highlights the experimental results for all five clustering algorithms: MBK, *k*-means, Agglomerative, DBSCAN, and Affinity embedded with FA and without FA using the performance measures and the computational time.

Table II demonstrates the performance of the algorithms mainly on f-measure, Precision, and recall. To add, it is interesting to note that the proposed MBK+EM+FA outperforms other algorithms with the highest in terms of F-measure, Precision, and recall, with 97%, 98%, 97%, and 96.30%, respectively. It has increased more than 0.15 of F-measure, 12% of Precision, and 16% of recall compared to MBK+EM, which is no FA is embedded as an optimization strategy. The performance of the proposed clustering algorithm was also compared with other clustering algorithms.

TABLE II. CLUSTERING PERFORMANCE FOR F-MEASURE, PRECISION AND RECALL

Clustering Algorithms	F	P	R
MBK+EM+FA	0.97	0.98	0.97
MBK+EM	0.82	0.86	0.81
K-Means+EM+FA	0.44	0.36	0.60
K-Means+EM	0.16	0.19	0.14
Agglomerative+EM+FA	0.44	0.36	0.60
Agglomerative+EM	0.16	0.19	0.14
DBSCAN+EM +FA	0.45	0.36	0.59
DBSCAN+EM	0.17	0.12	0.32
Affinity+EM+ FA	0.02	1.00	0.01
Affinity+EM	0.01	0.38	0.00

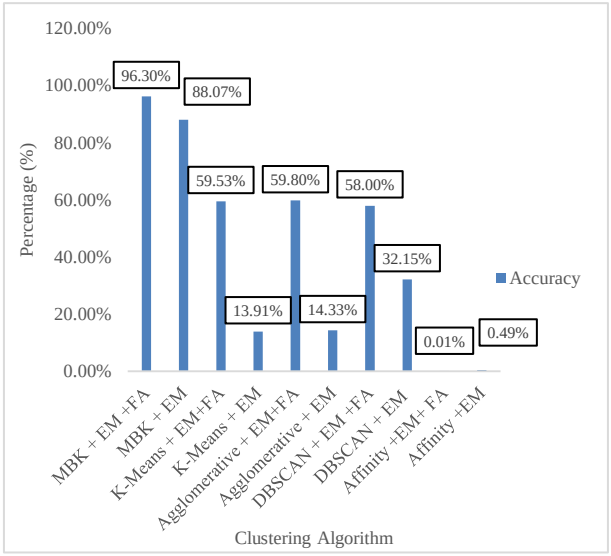

Fig. 2. Comparison of Accuracy Performance.

Fig. 2 shows the accuracy score of all algorithms. Overall, most algorithms have shown an increase in accuracy when a hybrid with FA. Only Affinity+EM is excluded. The proposed MBK+EM+FA has increased by 2.9%. It is revealed that the proposed algorithm is more accurate and capable of converging compared to other algorithms. It is observed that the f-measure, Precision, recall, and accuracy of the proposed clustering algorithm are the highest values. These values indicate that the performance of the proposed clustering algorithm was satisfactory based on high-performance parameters achieved.

Table III shows the comparative performance of the proposed clustering algorithm compared with other hybrid clustering algorithms based on external validation in terms of homogeneity, FMI, completeness, and V-Measure. Interestingly, all five clustering algorithms mostly performed better with the hybrid FA. All measurements have demonstrated much increment. It shows that clustering validation on homogeneity agreement, the perfectness of

labeling, and independent ability in clustering and clustering agreement is acceptable. However, the most efficient clustering algorithm is the proposed MBK+EM +FA. As seen in the table, it is evident that the increase of about 0.3 compared with MBK+EM for all external validation measurements. Most of the values are between 0.78 and 0.94. The highest FMI value is 0.9307. The obtained FMI indicates a good clustering agreement offered by the proposed MB+EM+FA due to the value being almost zero.

The time consumption of the algorithm is defined as the time required to assess all the data to generate clusters. The elapsed time of the algorithms is calculated and expressed in seconds, as shown in Table IV. Table IV indicates that the proposed MBK+EM+FA that uses entropy measure and FA consumes less time than K-Means+EM+FA, Agglomerative+ EM+FA, DBSCAN+EM+FA, and Affinity+EM+FA. It reveals that the hybrid proposed algorithm is much more effective than others. However, the use of MBK+EM+FA is a bit higher in computational time compared to MBK+EM. However, as mentioned in the previous section, it offers a good accuracy performance. A slight increase in execution time does not cause an issue in the practical use of the algorithm because a better performance is obtained. Overall, the computational time is less than a minute. Thus, it can be said to be sufficient time for the execution of cluster data since the proposed algorithm involves a searching mechanism compared to the others [29][30].

TABLE III. CLUSTERING PERFORMANCE COMPARISON FOR HOMO, FMI, COMP AND VM

Clustering Algorithms	HOMO	FMI	COMP	VM
MBK+EM+FA	0.7914	0.9307	0.7804	0.7858
MBK+EM	0.4777	0.804	0.4576	0.4673
K-Means+ EM+FA	0.0006	0.7192	0.0639	0.0013
K-Means+EM	0.4479	0.7742	0.4218	0.4343
Agglomerative+EM+FA	0.0006	0.7192	0.0639	0.0013
Agglomerative+EM	0.4496	0.7679	0.4214	0.4349
DBSCAN+EM+FA	0.0090	0.6938	0.0324	0.0141
DBSCAN+EM	0.0005	0.7033	0.0392	0.0002
Affinity+EM+FA	0.8493	0.1285	0.1232	0.2152
Affinity+EM	0.729	0.1178	0.101	0.1699

TABLE IV. EXECUTION TIME (PER SECONDS)

Clustering Algorithms	Time (s)
MBK+EM+FA	0.4589
MBK+EM	0.3550
K-Means+ EM+FA	0.5839
K-Means+EM	0.7850
Agglomerative+EM+FA	14.1792
Agglomerative+EM	13.9627
DBSCAN+EM +FA	87.7652
DBSCAN+EM	8.3744
Affinity+EM+FA	23.03922
Affinity+EM	11.7965

VII. DISCUSSION

The capability and effectiveness of a clustering approach to reduce the clustering error and improve the accuracy are the most crucial qualities in clustering. There is no inherent distance between the feature of categorical data analysis remains challenging. This research aims to evaluate the clustering efficiency of heterogeneous categorical data. Entropy is a measure of information. The entropy distance generates within the clustering algorithm approach allows us to measure the distance of features systematically quantified. The entropy aided the clustering algorithms in selecting the center of the centroid. Due to the nature of the entropy, it can investigate the compatibility of data to produce weighted values that can represent each class in the dataset. The initial weight value of entropy can study the data well and produces a higher accuracy.

Introducing the FA approach in the proposed clustering algorithms establishes a new search strategy for finding a globally optimal solution. The aim is to find the nearest features based on the entropy distance measure. In this case, the total number of categorical features is 111 features. Overall, achieving a clustering algorithm embedded with FA provides a good performance. It supports the previous work FA can contribute to an efficient solution due to its flexibility, simplicity, and fast convergence [31]. FA also offers a global search strategy [17] that can explore more search spaces for all features in finding the nearest neighbor. The generated value for the entropy measure seems to improve when using the FA. It could reflect in the clustering similarity measure based on a data point in both the intra-distance and inter-distance among the features. As illustrated in the above results, the best performance is by the proposed MBK+EM+FA. However, the high quality of the clustering results in more than constitutes for such restrictions of elapsed time. In addition, more work can be done such as embedding the *k*-interpolation model based on Kriging Method [25] and other computational optimization methods.

VIII. CONCLUSION

This research compares clustering algorithms that utilize the fundamental behavior of the firefly algorithm in order to improve clustering problem-solving. The proposed firefly algorithm can locate the cluster centers by comparing several clustering techniques to refine the centers as input. The experimental results have proved the effectiveness, capability, and efficiency of the FA and MBK with EM in improving the clustering performance in heterogeneous categorical data of public perception of timber utilization as compared to MBK+EM, K-Means+EM+FA, K-Means+EM, Agglome rative+EM+FA, Agglomerative+EM, DBSCAN+EM +FA, DBSCAN+EM, Affinity+EM+ FA, and Affinity+EM. In improving the solution, other objective functions considering inter and intra-cluster measurements could be used in future research to improve the proposed models. Another aspect is the evaluation of heterogeneous categorical data from a different domain.

ACKNOWLEDGMENT

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

REFERENCES

[1] Sapegin and C. Meinel, "K-metamodes: Frequency-and ensemble-based distributed k-modes clustering for security analytics," Proc. - 19th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2020, pp. 344–351, 2020, doi: 10.1109/ICMLA51294.2020.00062.

[2] S. V Ambadkar and S. P. Akarte, "Clustering Categorical Data for Internet Security Applications," Int. J. Sci. Tech. Adv., vol. 2, no. 1, pp. 115–118, 2016.

[3] N. Ifada, M. E. Ariyanto, M. K. Sophan, and M. Nikmat, "A Comparative Study of Centroid and Medoid based Categorical Data Clustering Methods for Solving Cold-start Recommendation Problem," CENIM 2020 - Proceeding Int. Conf. Comput. Eng. Network, Intell. Multimed. 2020, pp. 418–422, 2020, doi: 10.1109/CENIM51130.2020.9297960.

[4] Y. V. Via et al., "Optimization of Production Profits Using The Firefly Algorithm," pp. 291–296, 2020.

[5] A. Hassan and T. M. Tawfeeg, "Greedy Firefly Algorithm for Optimizing Job Scheduling in IoT Grid Computing," pp. 1–18, 2022.

[6] B. Filipowicz, "Firefly algorithm in optimization of queueing systems," vol. 60, no. 2, pp. 363–368, 2012, doi: 10.2478/v10175-012-0049-y.

[7] V. S. Rajput, "A New Approach of Firefly Algorithm for Optimizing Reviews of Opinion Mining," pp. 18–23, 2016.

[8] I. T. R. Yanto, Y. Saadi, D. Hartama, D. P. Ismi, and A. Pranolo, "A framework of fuzzy partition based on Artificial Bee Colony for categorical data clustering," 2nd Int. Conf. Sci. Inf. Technol., pp. 260–263, 2017, doi: 10.1109/ICSITech.2016.7852644.

[9] K. Lakshmi, N. Karthikeyani Visalakshi, S. Shanthi, and S. Parvathavarthini, "CLUSTERING CATEGORICAL DATA USING k-MODES BASED ON CUCKOO SEARCH OPTIMIZATION ALGORITHM," ICTACT J. Soft Comput., vol. 8, no. 1, pp. 1561–1566, 2017, doi: 10.21917/ijsc.2017.0218.

[10] H. Wu and S. O. Leung, "Can Likert Scales be Treated as Interval Scales?—A Simulation Study," J. Soc. Serv. Res., vol. 43, no. 4, pp. 527–532, Aug. 2017, doi: 10.1080/01488376.2017.1329775.

[11] F. Cao et al., "An Algorithm for Clustering Categorical Data with Set-Valued Features," IEEE Trans. Neural Networks Learn. Syst., vol. 29, no. 10, pp. 4593–4606, 2018, doi: 10.1109/TNNLS.2017.2770167.

[12] T. Hien, T. Nguyen, D. Tai, D. Songsak, and S. Van Nam, "A method for k-means-like clustering of categorical data," J. Ambient Intell. Humaniz. Comput., no. Berkhin 2002, 2019, doi: 10.1007/s12652-019-01445-5.

[13] Y. Zhang, Y. M. Cheung, and K. C. Tan, "A Unified Entropy-Based Distance Metric for Ordinal-and-Nominal-Attribute Data Clustering," IEEE Trans. Neural Networks Learn. Syst., vol. 31, no. 1, pp. 39–52, Jan. 2020, doi: 10.1109/TNNLS.2019.2899381.

[14] H. Sun, R. Chen, S. Jin, and Y. Qin, "A hierarchical clustering for categorical data based on holo-entropy," Proc. - 2015 12th Web Inf. Syst. Appl. Conf. WISA 2015, pp. 269–274, 2016, doi: 10.1109/WISA.2015.18.

[15] Y. Xiao, C. Huang, J. Huang, I. Kaku, and Y. Xu, "Optimal mathematical programming and variable neighborhood search for k-modes categorical data clustering," Pattern Recognit., vol. 90, no. Huang 1997, pp. 183–195, 2019, doi: 10.1016/j.patcog.2019.01.042.

[16] E. J. Rivera Rios, M. A. Medina-Pérez, M. S. Lazo-Cortés, and R. Monroy, "Learning-based dissimilarity for clustering categorical data," Appl. Sci., vol. 11, no. 8, pp. 1–17, 2021, doi: 10.3390/app11083509.

[17] A. E. S. Ezugwu, M. B. Agbaje, N. Aljojo, R. Els, H. Chiroma, and M. A. Elaziz, "A Comparative Performance Study of Hybrid Firefly Algorithms for Automatic Data Clustering," IEEE Access, vol. 8, pp. 121089–121118, 2020, doi: 10.1109/ACCESS.2020.3006173.

[18] H. Xie et al., "Improving K-means Clustering with Enhanced Firefly Algorithms," no. September, 2019, doi: 10.1016/j.asoc.2019.105763.

[19] A. Jaradat, B. Matalkeh, and W. Diabat, "Solving Traveling Salesman Problem using Firefly algorithm and K-means Clustering," 2019 IEEE Jordan Int. Jt. Conf. Electr. Eng. Inf. Technol. JEEIT 2019 - Proc., pp. 586–589, 2019, doi: 10.1109/JEEIT.2019.8717463.

[20] M. Takeuchi, T. Ott, H. Matsushita, Y. Uwate, and Y. Nishio, "K-Means Clustering with Modifying Firefly Algorithm," Int. Symp. Nonlinear Theory Its Appl., no. 1, pp. 576–579, 2017.

[21] A. Saha and S. Das, "Categorical fuzzy k-modes clustering with automated feature weight learning," Neurocomputing, vol. 166, pp. 422–435, 2015, doi: 10.1016/j.neucom.2015.03.037.

[22] X. S. Yang and X. He, "Firefly algorithm: recent advances and applications," Int. J. Swarm Intell., vol. 1, no. 1, p. 36, 2013, doi: 10.1504/ijsi.2013.055801.

[23] Meghana M Chavan, Asawari Patil, Lata Dalvi, and Ajinkya Patil, "Mini Batch K-Means Clustering On Large Dataset," Int. J. Sci. Eng. Technol. Res., vol. 04, no. 07, pp. 1356–1358, 2015.

[24] A. Feizollah, N. B. Anuar, R. Salleh, and F. Amalina, "Comparative study of k-means and mini batch k-means clustering algorithms in android malware detection using network traffic analysis," Proc. - 2014 Int. Symp. Biometrics Secur. Technol. ISBAST 2014, pp. 193–197, 2015, doi: 10.1109/ISBAST.2014.7013120.

[25] S. Bonnini, "Testing for heterogeneity with categorical data: Permutation solution vs. bootstrap method," Commun. Stat. - Theory Methods, vol. 43, no. 4, pp. 906–917, 2014, doi: 10.1080/03610926.2013.799376.

[26] X. Yang, Nature-Inspired Metaheuristic Algorithms. 2010.

[27] E. M. Mashhour, E. M. F. El Houby, K. T. Wassif, and A. I. Salah, "Feature selection approach based on firefly algorithm and chi-square," Int. J. Electr. Comput. Eng., vol. 8, no. 4, pp. 2338–2350, 2018, doi: 10.11591/ijece.v8i4.pp2338-2350.

[28] A. Karim, S. Azam, B. Shanmugam, and K. Kannoorpatti, "An Unsupervised Approach for Content-Based Clustering of Emails into Spam and Ham through Multiangular Feature Formulation," IEEE Access, vol. 9, pp. 135186–135209, 2021, doi: 10.1109/ACCESS.2021.3116128.

[29] A. Rauf, Sheeba, S. Mahfooz, S. Khusro, and H. Javed, "Enhanced K-mean clustering algorithm to reduce number of iterations and time complexity," Middle East J. Sci. Res., vol. 12, no. 7, pp. 959–963, 2012, doi: 10.5829/idosi.mejsr.2012.12.7.1845.

[30] M. P. Behera, A. Sarangi, and D. Mishra, "K-medoids crazy firefly algorithm for unsupervised data clustering," 1st Odisha Int. Conf. Electr. Power Eng. Commun. Comput. Technol. ODICON 2021, 2021, doi: 10.1109/ODICON50556.2021.9428980.

[31] N. Dey, Springer Tracts in Nature-Inspired Computing Applications of Firefly Algorithm and its Variants Case Studies and New Developments. Kolkata, West Bengal, India, 2020.