JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
http://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Ahmad, A., Yusof, R., Zulkifli, N. S. A., & Ismail, M. N. (2021). An improved pheromone-based kohonen self-organizing map in clustering and visualizing the balanced and imbalanced datasets. *Journal of Information and Communication Technology, 20*(4), 651-676. https://doi.org/10.32890/jict2021.20.4.8

An Improved Pheromone-Based Kohonen Self-Organising Map in Clustering and Visualising Balanced and Imbalanced Datasets

[1]Azlin Ahmad, [2]Rubiyah Yusof, [3]Nor Saradatul Akmar Zulkifli & [4]Mohd Najib Ismail
[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia
[2]Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Malaysia
[3]Faculty of Computer System & Software Engineering, Universiti Malaysia Pahang, Malaysia
[4]School of Computing Asia Pacific University Technology Park Malaysia, Malaysia

azlin@fskm.uitm.edu.my
rubiyah.ic@utm.my
saradatulakmar@ump.edu.my
najib.ismail@staff.apu.edu.my

Received: 20/11/2020 Revised: 24/1/2021 Accepted: 1/9/2021 Published:27/9/2021

ABSTRACT

The data distribution issue remains an unsolved clustering problem in data mining, especially in dealing with imbalanced datasets. The Kohonen Self-Organising Map (KSOM) is one of the well-known clustering algorithms that can solve various problems without a pre-

defined number of clusters. However, similar to other clustering algorithms, this algorithm requires sufficient data for its unsupervised learning process. The inadequate amount of class label data in a dataset significantly affects the clustering learning process, leading to inefficient and unreliable results. Numerous research have been conducted by hybridising and optimising the KSOM algorithm with various optimisation techniques. Unfortunately, the problems are still unsolved, especially separation boundary and overlapping clusters. Therefore, this research proposed an improved pheromone-based PKSOM algorithm known as iPKSOM to solve the mentioned problem. Six different datasets, i.e. Iris, Seed, Glass, Titanic, WDBC, and Tropical Wood datasets were chosen to investigate the effectiveness of the iPKSOM algorithm. All datasets were observed and compared with the original KSOM results. This modification significantly impacted the clustering process by improving and refining the scatteredness of clustering data and reducing overlapping clusters. Therefore, this proposed algorithm can be implemented in clustering other complex datasets, such as high dimensional and streaming data.

Keywords: Clustering, imbalanced data, Kohonen self-organising map, optimisation, pheromone.

INTRODUCTION

As of today, data visualisation has become an essential component of data analytics, especially in this era of big data. How to interpret the data so that everyone can understand is the critical problem in big data. The processing of extensive data will trigger issues in the analysis and extraction of data insights. Data are typically categorised into three types: (1) structured, (2) unstructured, and (3) semi-structured. Big data would also have to be handled because of the significant increase in data. Therefore, in data mining, particularly in clustering, numerous research have been developed and performed.

In various data mining applications, mainly in big data analysis, numerous clustering techniques have been applied. Clustering is theoretically a mechanism by which similar objects are clustered into clusters consisting of similar features, and the different ones are grouped into separate groups or clusters. The clusters will

accurately represent the unseen patterns and insights of the dataset during the clustering without knowing the data's target classes (Seman et al., 2012). Unlike the classification process, this clustering technique does not involve target or labelled data; it searches the data set for data exhibiting similar behaviour or characteristics. Here, the measured distance is utilized as a standard measure that calculates the distance between and within groups. Then, cluster analysis is applied to study clusters' meaning and hidden insights based on data characteristics. Numerous cluster exploration techniques are available in which different types of clusters are generated by each of these techniques (Joseph et al., 2018; Seman & Sapawi, 2020).

One of it is the Kohonen Self-Organising Map (KSOM) and Artificial Intelligence (AI) clustering approach that had been introduced by Kohonen (1999). This KSOM algorithm executes vector quantisation based on feature similarities; thus, it solves complex problems in many areas. For example, Riese et al. (2019) proposed a Supervised Self-Organising Map (SuSi) framework to investigate this framework's effectiveness for high-dimensional data, i.e. the hyperspectral data. They performed two tasks: regression and classification using KSOM and the results were compared with Random Forest (RF). For this research, two types of datasets were used: (1) a small dataset for regression of soil moisture, and (2) a large dataset for classification of land cover. As a result, KSOM had better performance for regression tasks than RF. It achieved satisfying results for classification tasks even though the results were below RF's performance.

Ashokkumar et al. (2019) implemented the KSOM algorithm to classify epilepsy using adjustable analytic wavelet transform in Electroencephalogram (EEG) signals. KSOM was utilised to produced a better performance rate in classifying the EEG signals into seizure brain signal and normal brain signal. Using the wavelet transform, KSOM reduced the hidden information of subband without any prior knowledge. Meanwhile, Sakkari and Zaied (2019) proposed a new Unsupervised Deep Self-Organising Map (UDSOM) algorithm, similar to the existing KSOM architecture for feature extraction. KSOM was used to abstract the subregion using the MNIST dataset, and they found that the computation complexity for this proposed algorithm was lower than the Convolutional Neural Network (CNN). Furthermore, KSOM was utilised to define the driving manoeuvres

using time series data (Lakshminarayanan, 2020). Each of the KSOM nodes represented the possible set of driving manoeuvres based on the observed fleet. They also compared the fuel consumption of two different trips on the same track using this algorithm. The clusters could easily specify the manoeuvre's content without a controlled measurement environment. This KSOM algorithm was also applied in object detection by Skuratov et al. (2020). They used KSOM to distinguish the boundaries of objects and determine the zones of interest in recognising objects in satellite images. KSOM has been one of the fastest algorithms as compared to other algorithms such as Region-based Convolutional Neural Networks (RCNN), You Only Look Once (YOLO), and Single Shot Detector (SSD).

Recently, KSOM has been implemented to analyse the coronavirus (COVID-19) pandemic cases worldwide by Melin et al. (2020). They used COVID-19 data cases worldwide, which occurred from January 2020 to May 2020. This included three conditions: number of confirmed, recovered, and death cases. The analysis was on the spatial evolution of COVID-19. They grouped countries according to the severity of the cases (very high, high, medium, and low) and adopted similar strategies to tackle the coronavirus problem. Their findings found that the death tendency caused by COVID19 is related to chronic diseases such as hypertension and diabetes. Therefore, it can be concluded that several variables such as learning parameters and topology map sizes may influence the KSOM clustering process and the result (Ahmad et al., 2017; Ahmad & Yusof, 2013; Ahmad & Yusof, 2014). Then, different map sizes are used to train the dataset to find the most suitable map size to represent the group in the data set correctly. Moreover, the characteristics of the data set play an important role in the clustering stage of learning, especially its distribution. In any neural network algorithm, the dominant group and the majority group will significantly affect the training phase. This scenario will also lead to a skewed outcome because the clustering process cannot be done correctly if the amount of data in some classes is insufficient.

A good clustering algorithm should produce interpretable, understandable, and functional results and can be more applicable and useful in solving the problem in other areas (Bryant, 2014; Han et al., 2006; Roohi, 2013; Yusoff et al., 2007). Various studies have been

conducted to enhance the clustering learning process and improve its performance. Table 1 shows several attempts to use different types and sizes of datasets to evaluate the proposed modified algorithm. Most of these works could only work well with a small dataset, especially with the balance distribution. Not only that, those proposed algorithms could not solve the common clustering problems, which are overlapping clusters and separation boundary (Shivakumar & Rajashekararadhya, 2020). Therefore, the initial KSOM algorithm has been hybridised with many intelligent techniques, such as Particle Swarm Optimisation (PSO) (Yang & Chi, 1997), Ant Colony Optimisation (ACO) (Mora et al., 2008; Wu & Chow, 2007), and Simulated Annealing (Mishra & Behera, 2012).

Table 1

Related Works on Hybrid KSOM

Work by	Techniques	Proposed Solution	Weakness
Yin and Chi (1997)	Ant-based SOM (ABSOM)	Solve the local minima problem	Works well with ONLY small dataset (Iris and Wine)
Aupetit et al. (1999)	Continuous SOM (C-SOM)	Solve the discrete representation in SOM and preserve topology	For 1-D SOM
Yin (2002)	Visualisation-induced SOM (ViSOM)	Solve data cluster structures in visualisation (for distorted shape)	Overlapped clusters
Sasamura and Saito (2003)	Growing SOM (GSOM)	Improve and optimise the neighbourhood preservation	Incapable of dealing with high dimensional data
Wu and Chow (2005)	Probabilistic Regularised SOM (PRSOM)	Improve visualisation effect	Requires a large amount of neurons – cause heavy computation

(continued)

Work by	Techniques	Proposed Solution	Weakness
Mora et al. (2008)	Kohonen – Ants (KohonAnts)	Group into clusters based on similar features	Cannot handle imbalanced data
Mohebi and Sab (2009)	SOM + Simulated Annealing	Solve overlapping clusters	Works well with small data only (Iris)
Meshra and Behera (2012)	SOM + modified K-Mean	Reduce the dimension and determine the number of cluster for the dataset	Works well with small data only (Iris)
Abdullah et al. (2010)	SOM + rough set + GA	Predict clusters and detect uncertainty that comes from overlapped data	Works well with small data only (Iris)
Overbeek et al. (2013)	Growing SOM (GSOM)	Use GSOM to assemble DNA sequence data	Small dataset is used
Ahmad et al. (2017)	Pheromone-based KSOM	Modify the learning process (distance and neighbourhood updates) using the pheromone approach from Ant Clustering Algorithm	Works well with small, medium, and large datasets, and still has many overlapped data

Yang and Chi (1997) suggested the Ant-Based Self-Organising Function Map algorithm, also known as ABSOM. ABSOM hybridised the KSOM algorithm and ACO. They use the pheromone method of the ant colony system to memorize the most matching unit. Although this hybridization solved local minima problem, this ABSOM algorithm is only suitable for small datasets with a normal distribution, such as Iris and Wine datasets. Next, Fernandes et al. (2008) tried to solve the overlapping problem in clustering by proposing the KohonAnts algorithm (also a combination of KSOM and ACO algorithm). Next, Fernandes et al. (2008) tried to solve the overlapping clustering problem by implementing KohonAnts; which is also implemented ACO concept. Same as ABSOM, this KohonAnts can only group the balanced dataset into desired group, accurately.

K-Means and Genetic Algorithm are other strategies hybridised with KSOM (Bouyer et al., 2010; Mihsra & Behera, 2012; Sameer & Abu Bakar, 2017). To solve the mentioned problem, Mishra and Behera modified K-Means; the unknown number of k, where KSOM acted as the visualisation method and helped to visualise the number of clusters before completing the learning process. At the same time, Bouyer et al. (2010) introduced a two-level clustering algorithm that hybridised KSOM with Genetic Algorithm and rough set theory. This hybrid algorithm aimed to solve the uncertainty problem that came from overlapped data. Genetic Algorithm was implemented to find the original cluster for the overlapped data, and they found that this hybrid algorithm was capable of clustering the dataset precisely, with few errors.

Furthermore, the present researchers have previously proposed a pheromone-based KSOM (PKSOM) capable of clustering the datasets with different distribution and sizes (Ahmad et al., 2017; Ahmad & Yusof, 2016). The PKSOM algorithm is inspired by the Ant Clustering Algorithm (ACA) pheromone method introduced by Handl et al. (2007). PKSOM solved the separation boundary problem amongst clusters, and partly solved the overlapping cluster problem. Several overlapped clusters were formed, where some of the related data had been incorrectly clustered. Therefore, in this work, an improved PKSOM (iPKSOM) is proposed by adapting the probability procedures for pick up and drop down by Gao (2016). A detailed discussion on the improved algorithm, experimental works, and analysis of results will be discussed in the next section.

PROPOSED METHODOLOGY

Overlapped clusters, scattered data, and difficulty in identifying the cluster's separation boundary are common problems faced by KSOM and other modified KSOM algorithms. These problems arise due to the similarity of the features between objects and the proportion of data imbalance. Therefore, to solve these issues, adjustment and refinement must be done. Generally, the KSOM learning process has two main processes; (1) distance calculation and (2) weight update. These two main procedures are important in identifying the validity of the clustering results. The Euclidean Distance (ED) is used in the

distance measurement stage to calculate the distance of objects (Kohonen, 1990). However, a single distance between two points is the distance measured. Usually, the Gaussian function is used as a neighbourhood function. Therefore, some refinement has been done to the previously modified algorithm (named PKSOM) by adapting the improved Pheromone Density Measure by Gao (2016).

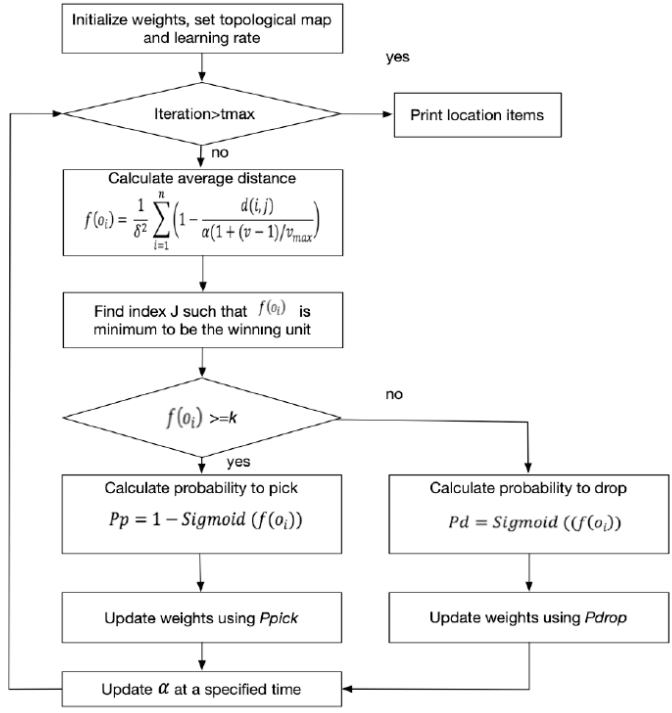
For the calculation of the average distance in PKSOM, the calculation of the pheromone density measurement (PDM) is adapted from ACA. The PDM function observes the distance and can measure the average similarity of objects, o_i, and other objects that located in neighbourhood δ. By adding a few parameters, PDM had been improved. v and v_{max}, as Gao (2016) suggested, are shown in Equation 1.

$$f(o_i) = \frac{1}{\delta^2} \sum_{i=1}^{n} \left(1 - \frac{d(i,j)}{\alpha(1 + (v-1)/v_{max}}\right) \quad (1)$$

$$d(i,j) = \sqrt{\sum_{i=1}^{n} (w_{ij} - x_i)^2} \quad (2)$$

Figure 1

The Flow of iPKSOM



The parameter v defines the ants' speed, while vmax represents the maximum speed, and for this, *v* is distributed randomly within the range of [*1, vmax*]. n denotes the number of cluster nodes in the space. The *d(i,j)* is the Euclidean Distance in Equation 2 to measure the distance between object *i* and *j*, and m represents the attribute number. The overall process of iPKSOM is shown in Figure 1.

Also, the modified method implements the exploration and exploitation concept in ACA, just like PKSOM. In iPKSOM, the probabilities of being picked up in Equation 3 and the object's probability to be dropped in Equation 4 were used. According to Deneubourg et al. (1990), the possibilities were based on the corpse clustering and the larvae sorting in ants. The distinct object should be picked up and dropped at some other location, where more objects should be more of that type were present. The choices

to drop and pick the current object are random and affected by other nearby objects. If the neighbouring objects are similar, the probability of discarding objects can be increased. Conversely, if the surrounding neighbourhoods are different, the probability of selecting an item will increase. The sigmoid function was used in both procedures; drop and pick up, as shown in Equation 5. *c* is a slope constant that can speed up the algorithm convergence if increased.

$$Pp = 1 - Sigmoid\ (f(o_i)) \tag{3}$$

$$Pd = Sigmoid\ ((f(o_i)) \tag{4}$$

$$sigmoid(x) = \frac{1 - e^{-cx}}{1 + e^{-cx}} \tag{5}$$

The decision of which procedure to perform will be based on the value of *f(oi)*. The picked-up process is executed if *f(oi)* equals or is greater than the threshold, *k*. This indicates that the current object is different from neighbouring objects. Instead, the dropped-down process is performed. Therefore, the current object to be moved to a different position away from the object. If the value of $f(o_i)$ is less than the threshold *k*, it indicates that the current object is similar to the neighbouring object, and the dropped-down will be executed. For this research, experiments were conducted using various numbers of *k* and discriminant factor (α) values to improve the exploration and exploitation of this iPKSOM algorithm.

EXPERIMENTAL WORKS

For this research, all selected datasets were trained and tested using the proposed algorithm and the original KSOM clustering algorithms. Table 2 displays a summary of the datasets utilised in this research. There are six datasets selected for this research: Iris, Seed, Glass, Titanic, Wisconsin Diagnostic Breast Cancer (WDBC), and Tropical Wood datasets. These datasets represented different data categories: small, medium, and large, with balanced and imbalanced data distributions (Wegman, 1995). The Iris, Seed, Glass, Titanic, and WDBC datasets were acquired from the UCI Machine Learning

Repository (https://archive.ics.uci.edu/ml/) and Kaggle (www.kaggle.com) as benchmark datasets. The Tropical Wood dataset was obtained from the Centre of Artificial Intelligence and Robotics (CAIRO), Universiti Teknologi Malaysia (UTM), Malaysia.

Table 2

The Description of the Datasets

Dataset	No. of Attributes	No. of Instances	Class Distribution	Size of Dataset	Data Distribution
Iris	4	150	Setosa (33%) Versicolor (33%) Virginica (33%)	Small (14,000 bytes)	Balanced
Glass	9	214	Kama (33%) Rosa (33%) Canadian (33%)	Small (24,000 bytes)	Imbalanced
Seed	7	210	Window Glass (76%) Non-Window Glass (34%)	Small (24,000 bytes)	Balanced
Titanic	6	1,309	Survived (38%) Not Survived (62%)	Medium (107,000 bytes)	Imbalanced
WDBC	30	569	Benign (62%) Malignant (38%)	Medium (164,000 bytes)	Imbalanced
Tropical Wood	157	5,040	Small Pores (5%) Medium Pores (77%) Large Pores (18%)	Large (5,467,000 bytes)	Balanced

Dataset Selection

Iris Dataset

The Iris dataset consists of four main attributes: sepal length, sepal width, petal length, petal width and one class label. The total data for this dataset is 150 data, where each class is well-distributed among all

three categories of iris; Setosa, Versicolor and Virginica. Thus, this dataset can be categorized as a small and balanced dataset.

Seed Dataset

The Seed dataset is also a small dataset consisting of seven attributes representing three categories: Kama, Rose, and Canadian. These attributes are area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient, and kernel groove length. Each category has a balanced distribution of data, i.e. 70 each.

Glass Dataset

The Glass dataset is a small dataset that comprises 215 samples with nine attributes: Refractive Index (RI), the amount of Sodium (Na), Magnesium (Mg), Aluminum (Al), Silicon (Si), Potassium (K), Calcium (Ca), Barium (Ba), and Iron (Fe). There are seven subcategories of glass: building windows float processed, building windows non-float processed, vehicle windows float processed, vehicle windows non-float processed (not including the database), containers, tableware, and headlamps. Therefore, this paper only considered two main classes: (1) window glass and (2) non-window glass. The distribution of both categories is imbalanced, where 76 percent of the data go to window-glass, while 24 percent are non-window glass.

Titanic Dataset

The Titanic dataset contains 1,309 data samples with 14 attributes. However, only eight significant attributes were selected, and the other insignificant attributes were removed from the training set. The data set has two main categories: (1) survived and (2) not survived. The distribution of these two categories was imbalanced, where 809 samples fall under the 'not survived' category, while the remaining data were under 'survived'. Similar to the Glass dataset, the distribution of both classes, survived and not survived, was imbalanced.

WDBC Dataset

Wisconsin Diagnostic Breast Cancer (WDBC) has 569 cases, with both benign and malignant diagnostic categories as a medium data set. This data set is unbalanced because there are 62% more benign data than malignant data. The data set consists of 30 real-valued input features for each cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concave point, symmetry, and fractal dimension. This data set is categorized as a medium-sized data set with an imbalanced distribution.

Tropical Wood Dataset

For this dataset, the CAIRO research team from UTM received the wood samples from the Forest Research Institute of Malaysia (FRIM), including 52 tropical wood species in cubic form. Every species has five samples cubes. The two feature extractors are Basic Grey Level Aura Matrix (BGLAM), and Statistical Properties of Pores Distribution (SPPD), which are used to extract the wood features (Khairuddin et al., 2011). Both methods generate 157 features from the dataset, where 136 features are produced by BGLAM, while the other 21 features from SPPD. This dataset is categorised under a large dataset, and it has 5,040 samples of data representing three sizes of wood pores: small, medium, and large.

Performance Measures

In clustering the datasets, multiple performance metrics were chosen to measure the performance of the algorithms. Quantisation and topological errors were selected, just like KSOM and PKSOM, since both errors were used in determining the consistency of iPKSOM clustering. The quantisation error is the similarity of the input to its output, where iPKSOM is more reliable with a lower quantisation error than the one with a higher quantisation error (Yang & Chi, 1997). However, the topographic error determines the preservation of the topology, which will measure the distance between the first and second-best matching units. If the error is small, it is indicated that the

first and second matching units are close, and vice versa. Therefore, it can be assumed that the topology is not preserved (Mishra & Behera, 2010; Mora et al., 2008). Furthermore, the clustering results of all datasets for all algorithms were compared based on accuracy.

ANALYSIS OF RESULTS

Thorough experiments were performed to examine the effectiveness of the proposed algorithms. The iPKSOM results were then compared with the original KSOM based on the number of clusters produced, the number of misclustered data, and the percentage of accuracy delivered by each algorithm. Table 3 displays the full results for all six datasets.

Table 3

Summary Results for All Datasets using KSOM and iPKSOM

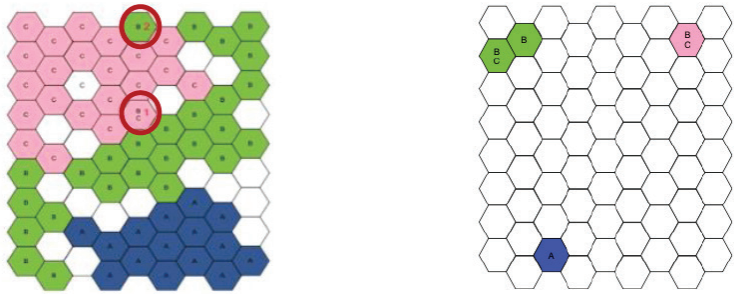
Datasets	KSOM			iPKSOM		
	No. of clusters produced	No. of incorrect data	Accuracy %	No. of clusters produced	No. of incorrect data	Accuracy %
Iris	3	3	98.00	3	3	98.00
Seed	3	14	93.33	3	11	95.1
Glass	2	10	90.95	2	4	91.57
Titanic	2	48	91.33	2	36	92.1
WDBC	2	63	88.93	2	52	89.34
Tropical Wood	4	190	96.23	3	110	97.8

Based on the results, both algorithms (KSOM and iPKSOM) produced the same results for the Iris dataset, whereby the percentage accuracy was 98 percent. Meanwhile, for the other five datasets, iPKSOM produced better accuracy with an average accuracy of 93.18 percent as compared to KSOM's average accuracy of 92.15 percent. The proposed algorithm also produced a lower number of misclustered data, showing that iPKSOM could handle medium and large datasets with balanced or imbalanced distribution. It was then proven that the accuracy for iPKSOM was consistent and is the proposed algorithm

was more scalable throughout all datasets as compared to KSOM. Notwithstanding these three measurements, iPKSOM was then compared with the original KSOM based on the visualisation quality.

Figure 2

Clustering Results for Iris Dataset

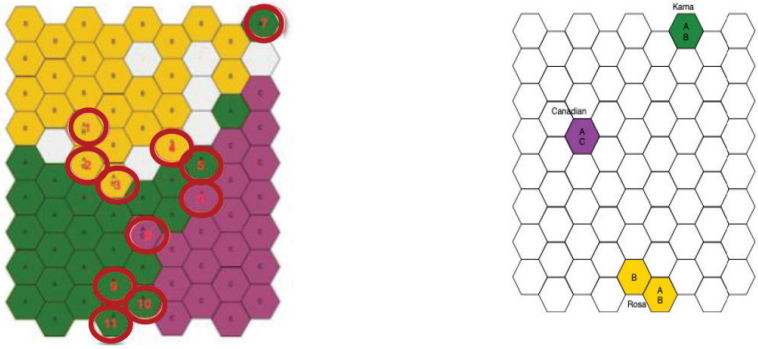


(A) KSOM (B) iPKSOM

Figure 2 shows the clustering result for the Iris dataset. KSOM clustered the Iris dataset into three desired classes of clusters, whereby each of the groups comprised several clusters of each species. However, there are two overlapped clusters (as circled in Figure 2(a)). IPKSOM also clustered the dataset into three distinct clusters, but with each group having only one or two clusters, namely Setosa, Virginica and Versicolor. All of the data from Setosa were grouped into one cluster. As for the other two species of Virginica and Versicolor, there were some overlapped data. There were three misclustered data of which two were from Virginica, and the other one was from Versicolor.

Figure 3

Clustering Results for Seed Dataset



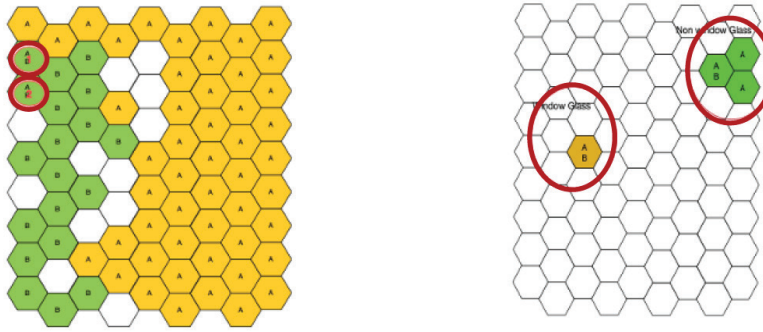
(A) KSOM (B) iPKSOM

In addition, the KSOM algorithm groups the Seed data set according to three groups: Kama (A), Rosa (B), and Canadian (C). The clustered data is scattered, and there are 11 overlapping clusters appear in the existing clusters (as presented in Figure 3(a)). However, the technique proposed by iPKSOM arranges all the clusters into three different groups, and these groups are precisely located separately according to their categories, as shown in Figure 3(b).

Next, the clustering results performed by KSOM and iPKSOM in the Glass dataset is presented in Figure 4. KSOM partitions the dataset into two groups: (1) window glass (on the right side of the map) and (2) non-window glass, with two overlapping groups. This is shown in Figure 4(a). In contrast, iPKSOM divides the data set into two separated groups with the least number of clusters.

Figure 4

Clustering Results for Glass Dataset

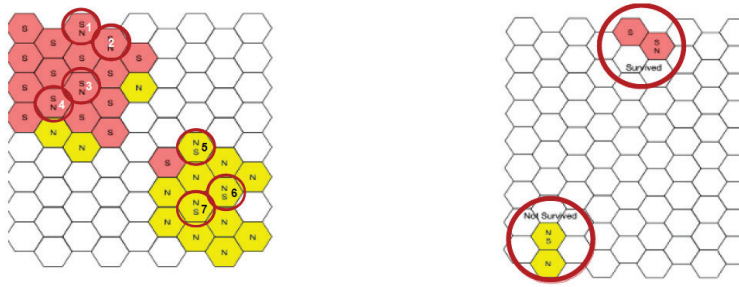


(A) KSOM (B) iPKSOM

Furthermore, the iPKSOM also reduced the number of clusters used for the Titanic dataset. Based on Figure 5(a), it can be seen that KSOM clustered this dataset into two separate clusters. There were seven overlapped clusters with many overlapped data as compared to iPKSOM; only two overlapped clusters with a lesser number of overlapped data. For this dataset, the number of overlapped clusters also decreased to 7 percent.

Figure 5

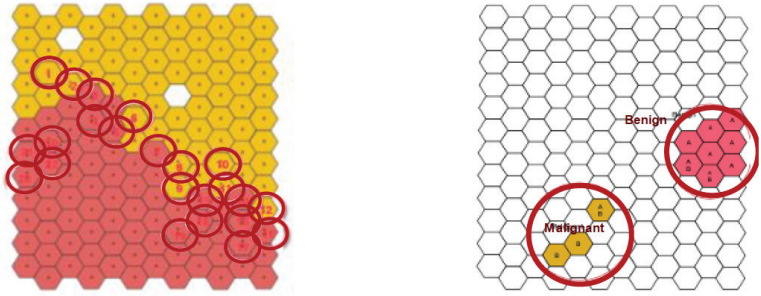
Clustering Results for Titanic Dataset



(A) KSOM (B) IPKSOM

Figure 6

Clustering Results for WDBC Dataset



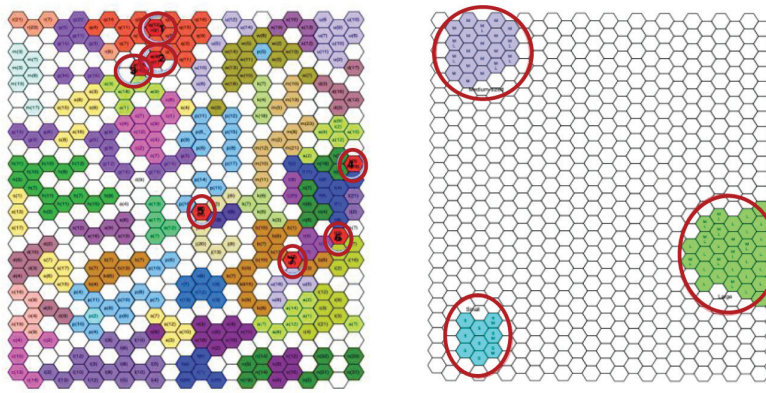
(A) KSOM (B) iPKSOM

Two cancer stages were grouped into the WDBC dataset: Benign (A) and Malignant (B). Similarly, Figure 6 (a) shows that KSOM mapped the WDBC dataset scattered in the topological map, like the other datasets. The number of overlapped clusters were 19 clusters (as shown in Figure 6(a)). Meanwhile, iPKSOM correctly categorised and clustered the datasets into two distinct groups, whereby both clusters were divided. The number of clusters and overlapped data was also minimal for each group.

Finally, as shown in Figure 7, iPKSOM precisely clustered the Tropical Wood dataset into three desired clusters. Compared to the KSOM algorithm, the volume of overlapped data was also reduced significantly. There were many overlapped clusters, and the Wood dataset was mapped scatteredly all over the topology map. This caused difficulty in analysing and interpreting the KSOM results. Based on these two algorithms' results, it can be inferred that the proposed algorithms improved the clustering method by solving the cluster separation boundary problem and reducing overlapped clusters, even though the overlapping problem was not fully solved.

Figure 7

Clustering Results for Tropical Wood Dataset



(A) KSOM (B) iPKSOM

The number of outliers or errors generated by these two algorithms are also observed. In most cases, the standard solution for obtaining robustness metrics is to calculate the standard deviation based on the percentage of data points (1-α), where α is usually 5% (BenGal, 2005; Liao et al., 2004; Tukey, 1977).

Figure 8

The Outliers Percentage for All Datasets

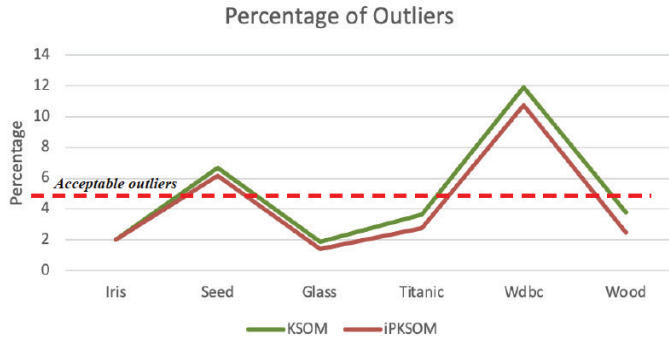


Table 4

The Percentage of Outliers Generated by Both Algorithms

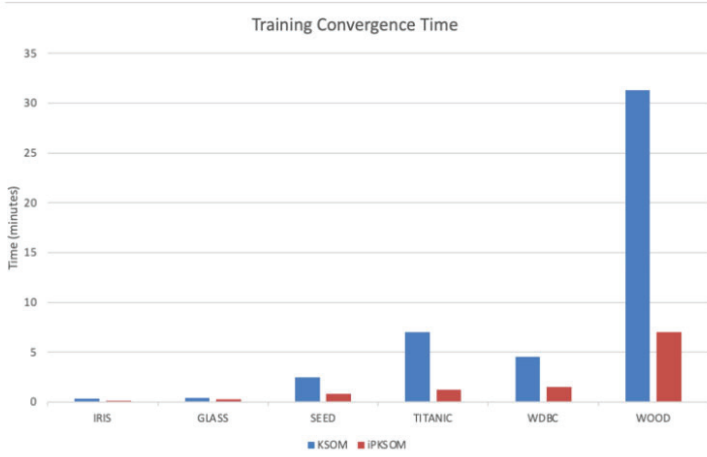
Dataset	Acceptable Outliers (5%)	Exact Number of Outliers		Percentage of Outliers (%)	
		KSOM	iPKSOM	KSOM	iPKSOM
Iris	8	3	3	2.00	2.00
Seed	11	14	11	6.67	6.19
Glass	11	4	3	1.90	1.43
Titanic	66	48	36	3.66	2.75
WDBC	28	63	52	11.9	10.77
Tropical Wood	252	190	110	3.77	2.48

Figure 8 and Table 4 show the acceptable errors for all datasets and the outliers produced by the KSOM and PKSOM algorithms for every dataset. Both algorithms produced a small number of outliers for small and large datasets. Nevertheless, both algorithms produced 6 percent – 11 percent of outliers for medium datasets. From the results, the IPKSOM algorithm generated a low percentage of outliers compared to KSOM for Seed, WDBC and tropical wood datasets. On the other hand, the original KSOM produces some outliers for small and medium data sets, such as Iris and Glass.

In addition, we also observed the training convergence times to measure the efficiency of the proposed algorithms. All training processes were performed using MacBook Pro using the same computer device, 2.53 GHz Intel Core 2 Duo. Figure 9 shows the convergence time of all data sets for both algorithms.

Figure 9

Training Convergence Performance for both algorithms



From the Figure 9, we can see that the PKSOM computation time for Iris, Glass, Seed, and WDBC datasets, were about 50%– 60% faster than KSOM. Based on the ratio of convergence time produced by both algorithms during the training process, there was a significant difference for the Wood dataset. The PKSOM computation time was 78% less than KSOM.

In conclusion, the proposed iPKSOM technique improves grouping results by significantly reducing the training time. The exploration and exploitation of similar objects in the topological graph are enhanced through the PDM pheromone method and the two mobile procedures of the drop-down and pick-up procedures. Hence, it accelerates the iPKSOM learning process. The use of the smallest number of clusters in the displayed results also shows that PDM produces a small similarity value, improving the efficiency of similar grouping objects close to each other. Different from the existing distance calculation, it only considers the distance between objects. It generates distance values of various ranges, which may cause data to be sparsely clustered on the map.

CONCLUSION

Although the KSOM algorithm has confirmed that a successful clustering algorithm, without knowing the target groups, can cluster various types of datasets, some drawbacks and problems need to be addressed. Issues such as cluster overlap, cluster creation, and dispersion of data may lead to difficulty in defining cluster separation boundary and managing large and imbalanced datasets. This KSOM algorithm was then modified to solve specific problems by proposing a modified KSOM algorithm, called iPKSOM. The theory of discovery and exploitation in this modified algorithm is based on the pheromone method, inspired by the ant clustering algorithm. Six distinct datasets were used to examine the proposed algorithm's efficiency in the clustering of balanced and imbalanced datasets.

In conclusion, it can be seen from the results that iPKSOM is able to partly solve the clustering problems, especially in reducing the overlapped clusters and improving the separation boundary problem. The scattering of clustering data has been improved and optimised by increasing the data density in clusters. Data for all the datasets in the same cluster were clustered and gathered close to each other. It also increases the accuracy for all datasets remarkably. This modified algorithm can be tested for complex datasets in the future to evaluate its ability to handle big data, mostly unstructured data.

ACKNOWLEDGMENT

The authors would like to thank the Ministry of Education Malaysia for financing this research project by a Research University Grant, i.e. Bestari Perdana 2018 Grant, with the project titled "Modified Clustering Algorithm for Analysing and Visualising the Structured and Unstructured Data" (600-RMI/PERDANA 513 BESTARI(059/2018)). Additionally, appreciation goes to the Faculty of Computer and Mathematical Sciences of Universiti Teknologi MARA (UiTM) for providing an excellent research environment in performing this research work.

REFERENCES

Abdullah, A. H. (2010). An optimized clustering algorithm using genetic algorithm and rough set theory based on Kohonen self organizing map. *International Journal of Computer Science and Information Security*, *8*(4), 39–44.

Ahmad, A., & Yusof, R. (2014, June). Refining the scatteredness of classes using pheromone-based Kohonen self-organizing map (PKSOM). In *INTELLI 2014: The Third International Conference on Intelligent Systems and Applications* (pp. 107–113).

Ahmad, A., & Yusof, R. (2013, July). Clustering the tropical wood species using Kohonen self-organizing map (KSOM). In *Proceedings of the 2nd International Conference on Advances in Computer Science and Engineering (CSE 2013)* (pp. 16–19). https://doi.org/10.2991/cse.2013.5

Ahmad, A., Yusof, R., & Mitsukura, Y. (2015, May). Pheromone-based Kohonen self-organizing map (PKSOM) in clustering of tropical wood species: Performance and scalability. In *2015 10th Asian Control Conference (ASCC)* (pp. 1–5). IEEE.

Ahmad, A., Yusoff, R., Ismail, M. N., & Rosli, N. R. (2017, July). Clustering the imbalanced datasets using modified Kohonen self-organizing map (KSOM). In *Proceedings of Computing Conference 2017* (pp. 751–755). https://doi.org/10.1109/SAI.2017.8252180

Ashokkumar, S. R., MohanBabu, G., & Anupallavi, S. (2020). A KSOM based neural network model for classifying the epilepsy using adjustable analytic wavelet transform. *Multimedia Tools and Applications*, *79*(15–16), 10077–10098. https://doi.org/10.1007/s11042-019-7359-0

Ben-gal, I. (2005). Outlier detection. In *Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers*. Kluwer Academic Publishers.

Bouyer, A., Hatamlou, A., & Abdullah, A. H. (2010). An optimized clustering algorithm using genetic algorithm and rough set theory based on Kohonen self organizing map. *International Journal of Computer Science and Information Security (IJCSIS)*, *8*(4), 39–44.

Bryant, T. (2014). Noise signal identification by modified self-organizing maps. *International Journal of Computer and Information Technology*, *02*(06), 48–53.

Deneubourg, J., Goss, S., Franks, N., A., S.-F., Detrain, C., & Chretien, L. (1991, February). The dynamics of collective sorting: Robot-like ants and ant-like robots. In *Proceedings of the First International Conference on Simulation of Adaptive Behaviour* (pp. 356–365).

Fernandes, C., Mora, A. M., Merelo, J. J., Ramos, V., & Laredo, J. L. J. (2008). KohonAnts: A self-organizing ant algorithm for clustering and pattern classification. *Artificial Life*, 428–435. http://arxiv.org/abs/0803.2695

Gao, W. (2016). Improved ant colony clustering algorithm and its performance study. *Computational Intelligence and Neuroscience*, *2016*. https://doi.org/10.1155/2016/4835932

Han, J., Kamber, M., & Pei, J. (2006). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

Handl, J., & Meyer, B. (2007). Ant-based and swarm-based clustering. *Swarm Intelligence*, *1*(2), 95–113. https://doi.org/10.1007/s11721-007-0008-7

Joseph, F. O. M., Kumar, P., & Behera, L. (2018, July). Redundancy resolution of an index finger exoskeleton using self organizing map. *IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)* (pp. 863–868). https://doi.org/10.1109/AIM.2018.8452424

Khairuddin, U., Yusof, R., Khalid, M., & Cordova, F. (2011). Optimized feature selection for improved tropical wood species recognition system. *ICIC Express Letters*, *Part B: Applications*, 2, 441–446.

Kohonen, T. (1990). The self-organizing map. *Proceeding of the IEEE, 78*(9), 1464–1480.

Liao, W., Liu, Y., & Choudhary, A. (2004). A grid-based clustering algorithm using adaptive mesh refinement. In *Proceeding of the 7th Workshop on Mining Scientific and Engi-neering Data Sets* (Vol. 22, pp. 61–69). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.7811&rep=rep1&type=pdf

Lakshminarayanan, S. (2020). Application of self-organizing maps on time series data for identifying interpretable driving manoeuvres. *European Transport Research Review*, *12*(1). https://doi.org/10.1186/s12544-020-00421-x

Melin, P., Monica, J. C., Sanchez, D., & Castillo, O. (2020). Analysis of spatial spread relationships of coronavirus (COVID-19) pandemic in the world using self organizing maps. *Chaos, Solitons and Fractals Nonlinear Science, and Nonequilibrium and Complex Phenomena*, *138*(January), 1–7.

Mishra, M., & Behera, H. S. (2012). Kohonen self organizing map with modified K-means clustering for high dimensional dataset. *International Journal of Applied Information Systems*, *2*(3), 34–39.

Mora, A. M., Fernandes, C. M., Merelo, J. J., Ramos, V., Laredo, J. L. J., & Rosa, A. C. (2008). KohonAnts : A self-organizing ant algorithm for clustering and pattern classification. 428–435.

Riese, F. M., Keller, S., & Hinz, S. (2020). Supervised and semi-supervised self-organizing maps for regression and classification focusing on hyperspectral data. *Remote Sensing*, *12*(1). https://doi.org/10.3390/RS12010007

Roohi, F. (2013). Artificial neural network approach to clustering 1. *The International Journal of Engineering and Science (IJES)*, *2*(3), 33–38.

Sakkari, M., & Zaied, M. (2020). A convolutional deep self-organizing map feature extraction for machine learning. *Multimedia Tools and Applications*, *79*(27–28), 19451–19470. https://doi.org/10.1007/s11042-020-08822-9

Sameer, F., & Abu Bakar, M. R. (2017). Modified Kohonen network algorithm for selection of the initial centres of Gustafson-Kessel algorithm in credit scoring. *Pertanika Journal of Science and Technology*, *25*(1), 77–90.

Seman, A., Bakar, Z. A., & Isa, M. N. (2012). An efficient clustering algorithm for partitioning Y-short tandem repeats data. *BMC Research Notes*, *5*(1), 1. https://doi.org/10.1186/1756-0500-5-557

Seman A., Sapawi A.M. (2020) A complementary optimization procedure for final cluster analysis of clustering categorical data. In Vasant P., Zelinka I., Weber GW. (Eds.), Intelligent Computing and Optimization. *Advances in Intelligent*

Systems and Computing, *1072*. Springer, Cham. https://doi.org/10.1007/978-3-030-33585-4_30

Shivakumar, B. R., & Rajashekararadhya, S. V. (2020). Classification of landsat 8 imagery using Kohonen's self organizing maps and learning vector quantization. *Lecture Notes in Electrical Engineering*, *614*(January), 445–462. https://doi.org/10.1007/978-981-15-0626-0_35

Skuratov, V., Kuzmin, K., Nelin, I., & Sedankin, M. (2020). Application of Kohonen self-organizing map to search for region of interest in the detection of objects. *EUREKA, Physics and Engineering*, *2020*(1), 62–69. https://doi.org/10.21303/2461-4262.2020.001133

Wegman, E. J. (1995). Huge datasets and the frontiers of computational feasibility. *Journal of Computational and Graphical Statistics*, *4*(4), 281–295. https://doi.org/10.1080/10618600.1995.10474685

Wu, S., & Chow, T. W. S. (2005). PRSOM: A new visualization method by hybridizing multidimensional scaling and self-organizing map. *IEEE Transactions on Neural Networks*, *16*(6), 1362–1380. https://doi.org/10.1109/TNN.2005.853574

Yang, C. C., & Chi, S. (1997). An ant-based self-orgazining feature maps algorithm. In *5th Workshop On Self-Organizing Maps* (pp. 65–74).

Yusoff, M., Rahman, S. A., Mutalib, S., & Mohamed, A. (2007). *Kohonen neural network performance in license plate number identification*. 512–515.