Indonesian Journal of Electrical Engineering and Computer Science
Vol. 24, No. 1, October 2021, pp. 600∼610
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v24.i1.pp600-610

An improved Kohonen self-organizing map clustering algorithm for high-dimensional data sets

Momotaz Begum[1], Bimal Chandra Das[2], Md. Zakir Hossain[3], Antu Saha[4], Khaleda Akther Papry[5]
[1,3,4,5]Department of Computer Science and Engineering, Dhaka University of Engineering and Technology, Gazipur, Bangladesh
[2]Department of General Educational Development, Daffodil International University, Dhaka, Bangladesh

Article Info

Article history:

Received Dec 2, 2020
Revised Jun 25, 2021
Accepted Jul 2, 2021

Keywords:

Clustering
EISEN Cosine correlation
High-dimensional data sets
Kohonen self-organizing map
Overlapping problem

ABSTRACT

Manipulating high-dimensional data is a major research challenge in the field of computer science in recent years. To classify this data, a lot of clustering algorithms have already been proposed. Kohonen self-organizing map (KSOM) is one of them. However, this algorithm has some drawbacks like overlapping clusters and non-linear separability problems. Therefore, in this paper, we propose an improved KSOM (I-KSOM) to reduce the problems that measures distances among objects using EISEN Cosine correlation formula. So far as we know, no previous work has used EISEN Cosine correlation distance measurements to classify high-dimensional data sets. To the robustness of the proposed KSOM, we carry out the experiments on several popular datasets like Iris, Seeds, Glass, Vertebral column, and Wisconsin breast cancer data sets. Our proposed algorithm shows better result compared to the existing original KSOM and another modified KSOM in terms of predictive performance with topographic and quantization error.

Corresponding Author:

Momotaz Begum
Department of Computer Science and Engineering
Dhaka University of Engineering and Technology, Gazipur, Bangladesh
Email: drmomotaz@duet.ac.bd

1. INTRODUCTION

Data are collection of information that can be observed, analyzed, and represented using images or other graphical tools. It can be of different types such as numerical, temporal, categorical, and multimedia that are used in data science and communication engineering. In tremendous application domain, the need to explore high dimensional data is increasing rapidly. High-dimensional data is more complex data having multiple attributes. Nowadays, various sectors, like e-commerce and micro biology, the researchers need to manipulate huge quantity of data for worthwhile decision making [1]. However, manipulating this data is a major concern, because high dimensional data may contain irrelevant attributes which may lead to lose clustering tendency. Moreover, distances among points in high-dimensional space decrease with the increase of dimensionality. Consequently, measuring these distances becomes meaningless which is referred to as the curse of dimensionality [2], is another crucial issue in this data. Accordingly, with the increase of dimensions or attributes many existing algorithms become inapplicable for identifying the actual clusters. Therefore, the study and analysis of this type of data are becoming more attractive to researchers.

Clustering is a method of grouping similar objects in a multivariate data sets. Cluster analysis is used in many applications including data classification, pattern recognition, and image processing [3]. The elementary

flows of clustering development process presents in Figure 1. Data objects with different features are collected before performing clustering algorithm. Using appropriate algorithm, effective features are selected and objects similarity are measured. Results of the algorithm are supported using suitable techniques and criteria. To figure out accurate knowledge, clustering results are integrated with other experimental results and analysis.

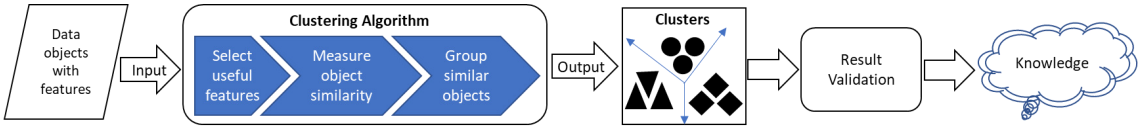

Figure 1. Process of data clustering

Clustering algorithms can be of various types, such as density-based clustering, grid based clustering, hierarchical clustering, partitioning methods, and model-based clustering. Model based clustering algorithm is a robust algorithm [4], as it automatically determines clusters considering outliers and best fits data. expectation-maximization (EM), COBWEB, CLASSIT, and Kohonen self-organizing map (KSOM) includes model based clustering examples. The KSOM [5] is an unsupervised learning (data set with no pre-existing labels) technique able to process high-dimensional data. It is an artificial neural network (ANN) that describes a mapping from high-dimensional input space to lower-dimensional mapping. Using different map sizes, the data set is trained simultaneously to find out more acceptable map size. Consequently, this map size represents the clusters of data set accurately. However, overlapping cluster and non-linear separability are limitations of this algorithm. In overlapping cluster, a data point may belong to one or more clusters. Therefore, we propose an improved KSOM (I-KSOM) algorithm to overcome the overlapping problem which is a major concern in high-dimensional data.

We organize our paper as follows; firstly, we discuss related works in section 2. Original KSOM algorithm and another modified KSOM [6] is briefly described in section 3. In section 4, the proposed algorithm is explained elaborately. The following section 5 shows experimental setup and result analysis of I-KSOM along with other algorithms. Finally, section 6 concludes our paper with future plans.

2. RELATED WORKS

A lot of researches have been conducted using KSOM for clustering different high-dimensional data. In addition, reducing the drawbacks of this algorithm is also a major concern in research area. Fernandes *et al.* [7] proposed an ant-based method for clustering with less free parameters that takes benefit of the cooperative self-organizing map (SOM) with ant colony optimization (ACO) named as kohon ants. Kovacs and Ko [8] developed a pneumatic actuator to observe the actions of a usual component by using big data signal processing method. For recognizing and grouping identical signal patterns they used Ward's hierarchical clustering with multivariate Euclidean distance and KSOM. Mallick *et al.* [9] worked with humongous gene expression data sets for identifying the cancerous gene. They applied KSOM for optimizing these multidimensional data and for detecting cancer. An extension of SOM has been proposed by Onishi [10]. As the SOM cannot narrate the relevance between the data and the location of the node, they suggested a landmark map (LAMA) that is able to perform landmark-concentrated data visualization i.e., a new view of data and desired non-linear projection.

As anomaly detection is challenging for surveillance videos, Ratre *et al.* [11] has demonstrated a method for the localization and detection of anomalies by propounding hybrid tracking model and fractional Kohonen self-organizing map (FKSOM). After tracking the objects by the tracking model, they applied the FKSOM algorithm to identify normal and anomalous events. For unraveling the complicated problems like image and signal processing, other methods are incorporated with the KSOM which is also vastly used for dimensionality reduction. Lagus *et al.* [12] demonstrated WEBSOM (web based SOM) which organizes voluminous document collection. Rezaei *et al.* [13] applied four non-hierarchical clustering methods including KSOM for the segmentation of the virtual histology-intravascular ultrasound (VH-IVUS) images. The authors showed that, as the number of the cluster is unknown, KSOM can not be applied in VH-IVUS. However, KSOM has become a common and versatile clustering method because it can handle noisy and partial data [14]. Moreover, it can metamorphose multidimensional data space into two or three-dimensional output space.

For ennobling and optimizing the neighbourhood preservation, the growing self-organizing map

(GSOM) was introduced which has several flaws like incompetent for high-dimensional data and projection distortion[15]. Therefore, for handling high-dimensional data, Amarasiri *et al.* [16] suggested the high dimensional growing self-organizing map (HDGSOM) where the clustering processes is surpassing. After that, for conquering the projection distortion and handling the mixed data, Hsu *et al.* [17] introduced the generalized visualization-induced self-organizing map (GViSOM) which preserves the topological structure of the data. To handle unsupervised high-dimensional data set, Begum and Akthar [18] introduced an algorithm named KSOMKM which used KSOM with an improved load based initial centroid K-Means algorithm.

Ahmad and Yusof [19] suggested pheromone-based kohonen self-organizing map (PKSOM) algorithm to filter the dispersed data in the clusters. Some modifications have been made to original KSOM using Ant Clustering Algorithm to improve the cluster density. To abate overlap in clustering and non-linear separable problem, Ahmad and Yusof [6] suggested a modified KSOM (M-KSOM) which is stimulated by pheromone approach in Ant Colony Optimization. They calculated the distance amongst objects by Euclidean Distance measure approach. The focus of M-KSOM has some similarities with our proposed algorithm. The authors proposed some modifications of the original KSOM algorithm towards performance improvement. However, this M-KSOM is also flawed as there were still some overlapped clusters. Furthermore, clusters were still congested and the separation boundaries were not clear. Hence, we proposed an improved KSOM (I-KSOM) algorithm to overcome the overlapping problem.

3. KOHONEN SELF-ORGANIZING MAP (KSOM)

T. Kohonen demonstrated the original KSOM [5] with data containing features of 500 dimensions. KSOM is a feed forward single layer neural network used in clustering large, complex, incomplete and noisy data sets [20]. It has the ability to transform multidimensional data space into two or three-dimensional output space [21]. There are two steps of the algorithm, train phase and recall phase. Figure 2 shows training procedure of the algorithm. In training phase, input vectors x_i ($i = 1, 2, ..., n$ where, n= no. of instances of each data set) are connected to the input layer and then summed up with weight vectors w_ij, ($i = 1, 2, 3, ..., n, j = 1, 2, 3, ..., n$) to calculate the competitive layer vectors.

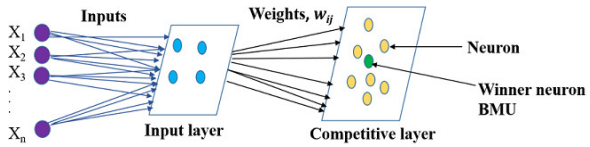

Figure 2. KSOM algorithm

The distance among these vectors in this layer is measured using Euclidean distance [6] presented in (1) and then the winner is selected named best matching unit (BMU) with minimum distance. The new weight of the BMU is then updated and after one training loop learning rate α is updated by some fraction. Finally, the difference between the updated weights of output node and the input vector is calculated and the learning process is continued if the difference is grater than some stopping value. Otherwise, learning process is stopped. After train phase, in recall phase, the weight vector is transposed and finally multiplied with input vector to get the final cluster,

$$D_(i,j) = \sqrt{\sum_{i=1,}^{n}(x_i - w_ij)^2} \quad (1)$$

here, j is output node, $D_(i,j)$ is Euclidean distance between all input nodes and the output node. The performance of KSOM algorithm is measured using two measurements that commonly used in evaluating and measuring the self-organization algorithm: topological error and quantization error [22].

3.1. The modified KSOM (M-KSOM) algorithm

Ahmad and Yusof [6] proposed the algorithm which applied pheromone-density measure (PDM) approach in ACO [23]. Additionally, PDM uses euclidean distance and some extra parameters to calculate distance between input and output nodes which is represented in (2),

$$f(i) = \frac{1}{\delta^2} \sum_{i=1}^{n} 1 - \frac{D_(i,j)}{\alpha(t)} \quad (2)$$

where, δ is the number of neighborhood nodes and $\alpha(t)$ is the learning rate. The convergence process is affected by selection of this learning rate value. If $\alpha(t)$ is extremely large, the convergence occurs very quickly and the learning process acts in incorrect way. On the other hand, the learning process becomes very slow, if the value is extremely small. Additionally, $f(i)$ measures the average distance of the current node with all neighbors, to find the most similar cluster accurately. The more small value of $f(i)$ of the current node indicates more similarity with its neighboring nodes. However, there exists some overlapping clusters and the separation boundaries among them are very close.

4. PROPOSED IMPROVED KSOM (I-KSOM) ALGORITHM

4.1. Problem statement

Inspite of having usefulness, the original KSOM has some major disadvantages like overlapping cluster and non-linearly separability problem. Therefore, all these issues require to consider the improvement of the introduced problems. In clustering, the choice of distance measure is a critical issue. Euclidean and Manhattan distances are most widely used methods in this regard. The original KSOM uses the Euclidean distance to find the distance which shrinks the data space with the increase of dimensionality [6], [24]. Hence, we propose an improved version of KSOM algorithm using another distance measurement approach to achieve better clustering for high-dimensional data. In our approach, we use EISEN Cosine correlation distance equation to measure closest neighbor which reduces the problem of overlapping clusters. The other process are kept same as original KSOM algorithm. Cosine similarity is a method of similarity measurement between two points in positive space, where the result is nearly bounded in [0, 1]. It is most commonly used in high-dimensional data spaces which measures similarity of two data points in terms of their attributes [25]. Another distance measure approach is dissimilarity measurement like correlation-based distance which is derived from the subtraction of correlation coefficient from 1. Pearson correlation distance, EISEN Cosine, Spearman and Kendall are examples of such correlation-based distances [26], [27]. EISEN Cosine is derived from subtracting Cosine similarity from 1 presented in (3). We inspired from the work of similarity measurement for text classification and clustering [28].

4.2. Contribution: I-KSOM algorithm

In this paper, we introduce EISEN Cosine distance measure based algorithm named I-KSOM for clustering to reduce the overlapping problems in high-dimensional data. The overall procedure is divided into two phase like original KSOM, the train phase and the recall phase. The algorithm is summarised below where step 1 to step 8 present the train phase and the rest of the steps represent recall phase of our algorithm.

1. Initialize the weight matrix w_ij, ($i = 1, 2, 3, ..., n, j = 1, 2, 3, ..., n$, and n=no. of instances of each data set), for all output node with random numbers bounded in [0, 1] so that all values are normalized. Then, assign a small learning rate α, where initially $\alpha(t)$=0.5 and iteration t=0 to total no of iterations T.
2. Repeat step 3 to 8, until the stopping condition is true.
3. For every element in the input vector x_i=x_1, x_2, x_3,..., x_n. repeat step 4-6.
4. For each output node j, calculate EISEN Cosine correlation distance between all input nodes and the output node j.

$$D_(i,j) = 1 - \frac{\sum_{i=1}^{n} x_i w_i j}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} w_i j^2}} \quad (3)$$

5. The output node having the minimum $D_(i,j)$ is selected as the best matching unit (BMU_t). Therefore, we have to link the input element with this output node. $\|BMU_t\|$ = min $\|D_(i,j)\|$
6. As an input element linked with a output node, the weights of this node is updated to make it more representative of all input elements connected with this node. The new weight of the winning output node is updated according to the learning rate by the following relation $w_ij(t+1)$ = $w_ij(t)$ +α $[x_i - w_ij(t)]$
7. After one training loop for all input element need to update the learning rate α by some fraction. So that the weights of output nodes move gradually close to the input vector. $\alpha\ (t+1) = 0.5\alpha\ (t)$

8. Check the difference between the updated weights of output node and the input vector. If the difference is grater than some stopping value continue the learning process. Otherwise, stop weight learning.
9. The final updated weight vector $w_ij(t+1)$ is then transposed $w_ij(t+1) = w_ij(t+1)^T$
10. The transposed weight vector is multiplied with input vector to get the final cluster. The nodes with similar value of J_i contain in the same cluster. $J_i = w_ij(t+1) \times x_i$

5. EXPERIMENTAL RESULTS

5.1. System setup

We carry out experiments on five common high-dimensional data sets [27] *i.e.* Iris, Seeds, Glass, Vertebral Column, and Wisconsin Breast Cancer. The data sets with their number of attributes, classes, and instances are summarized in Table 1. We implement our algorithm in Python and used NumPy's support for map processing. In our implementation, we set the parameters as learning rate, $\alpha = 0.5$, weight w_ij is randomly chosen bounded in [0, 1], and the total no of iterations, T varied from 1000 to 5000 with a difference of 1000 at each step.

Table 1. Data sets used in experiments

Serial No.	Data set Name	No. of Attributes	No. of Classes	No. of Instances
1	Iris	4	3	150
2	Seeds	7	3	210
3	Glass	10	6	214
4	Vertebral Column	6	2	210
5	Wisconsin Breast Cancer	32	2	569

5.2. Performance measurement

In this subsection, we present the performance measurement methods of the SOMs. Performance of SOMs can be measured in many ways. Average quantization error and topographic error are most common approaches in this regard [5], [29].

- Quantization error (QE) : The QE is a statistical measure of variance that is used for map resolution measurement. The low quantization error refers to better resolution of map. The (4) represents the mathematical expression of QE obtained from average distance between input data vector x_t and their corresponding BMU_t at each iteration t.

$$QE = \frac{1}{T}\sum_{t=1}^{T} ||x_t - BMU_t|| \tag{4}$$

- Topographic error (TE) : It defines how well the structure of the map is designed. The small value of the error indicates the well structure of the map. Usually, the best-matching unit (BMU) and second-best-matching unit (BMU') in the map for each input vector are evaluated for finding the error. The error is represented in (5), in terms of input vector error $d(x_t)$ at training time t divided by total number of iterations T .

$$TE = \frac{1}{T}\sum_{t=1}^{T} d(x_t), \qquad d(x_t) = \begin{cases} 1 & \text{if } BMU_t \text{ and } BMU'_t \text{ are not adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

5.3. Result analysis

Here, we present the detailed analysis of our algorithm result along with the original KSOM and M-KSOM [6]. We evaluate five high-dimensional data sets and compared with KSOM and M-KSOM to explore the effectiveness of our I-KSOM algorithm. Figures 3-7 shows the results for all data sets respectively for 4000 (a)-(c) and 5000 (d)-(f) iterations using all algorithms. Each different color represents different clusters of each data set, where outliers are presented using circles. Figure 3 shows the result of Iris data sets with three group of clusters; Setosa, Virginica, and Versicolor. From the figure, we can see that all the algorithms are able to produce all appropriate clusters. However, for both cases, (4000 and 5000 iterations) in KSOM and M-KSOM the data are dispersed on the topological map and clusters are closely located with some overlapping. Moreover,

M-KSOM provides more separable clusters in 5000 iterations than 4000 ones. On the other hand, I-KSOM is able to cluster the data with more distinguishable boundary in both cases. Therefore, I-KSOM is more efficient to produce accurate clusters than other two algorithms in this case.

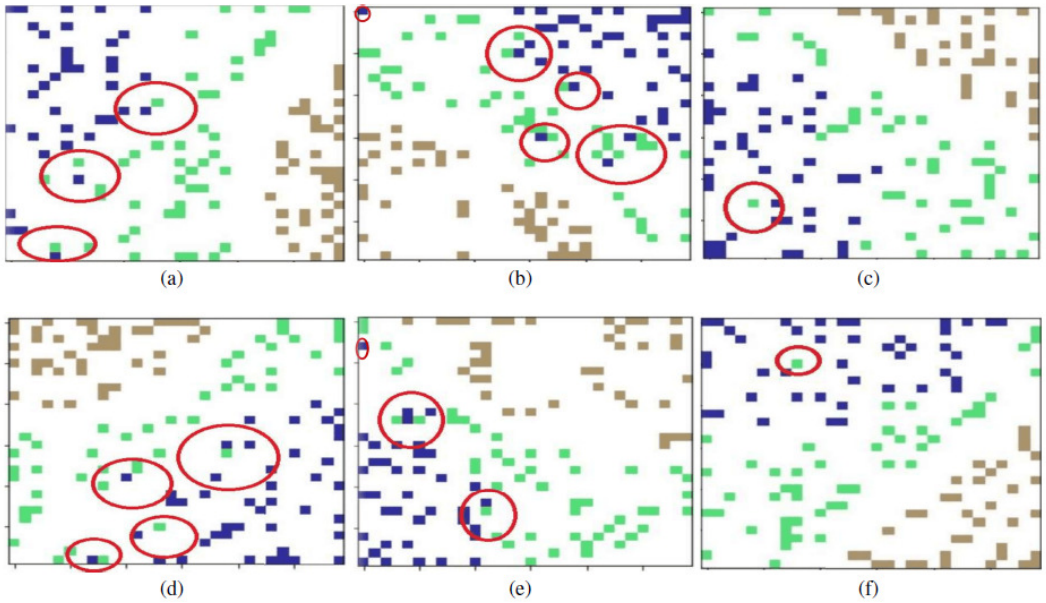

Figure 3. Results for iris data set after: (a) 4000 iterations of KSOM, (b) 4000 iterations of M-KSOM, (c) 4000 iterations of I-KSOM, (d) 5000 iterations of KSOM, (e) 5000 iterations of M-KSOM, and (f) 5000 iterations of I-KSOM

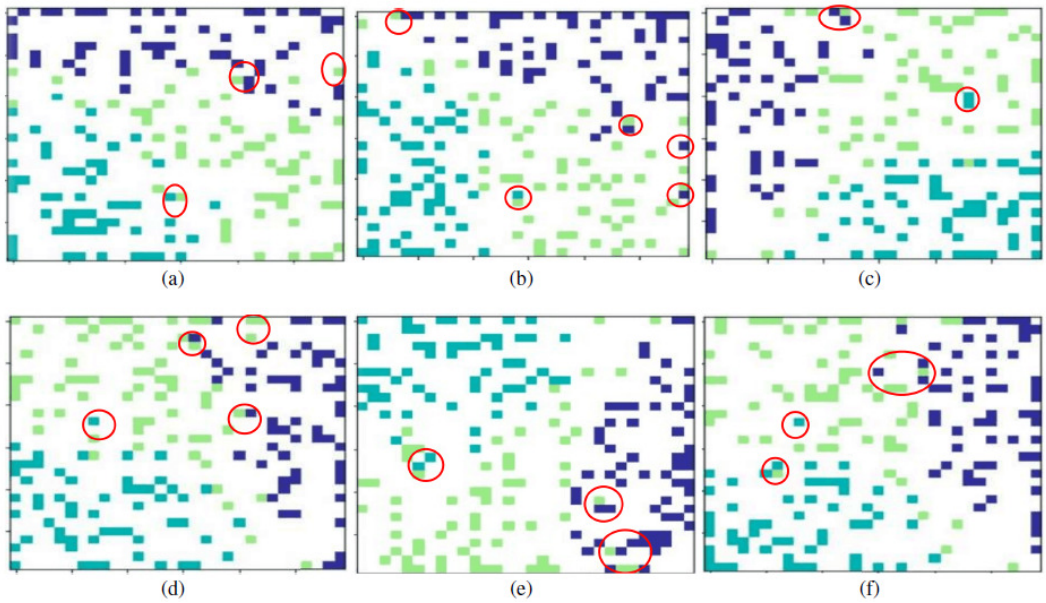

Figure 4. Results for seeds data set after: (a) 4000 iterations of KSOM, (b) 4000 iterations of M-KSOM, (c) 4000 iterations of I-KSOM, (d) 5000 iterations of KSOM, (e) 5000 iterations of M-KSOM, and (f) 5000 iterations of I-KSOM

Furthermore, seeds data set contains three clusters: Kama, Rosa, and Canadian; shown in Figure 4. In addition, Figure 6 represents Vertebral Column data set which has two clusters named normal and abnormal. In both cases, I-KSOM shows superior results than other algorithms whereas crystal and clear clusters appear for all iterations. In these cases, some overlapping clusters still remain in original KSOM and M-KSOM.

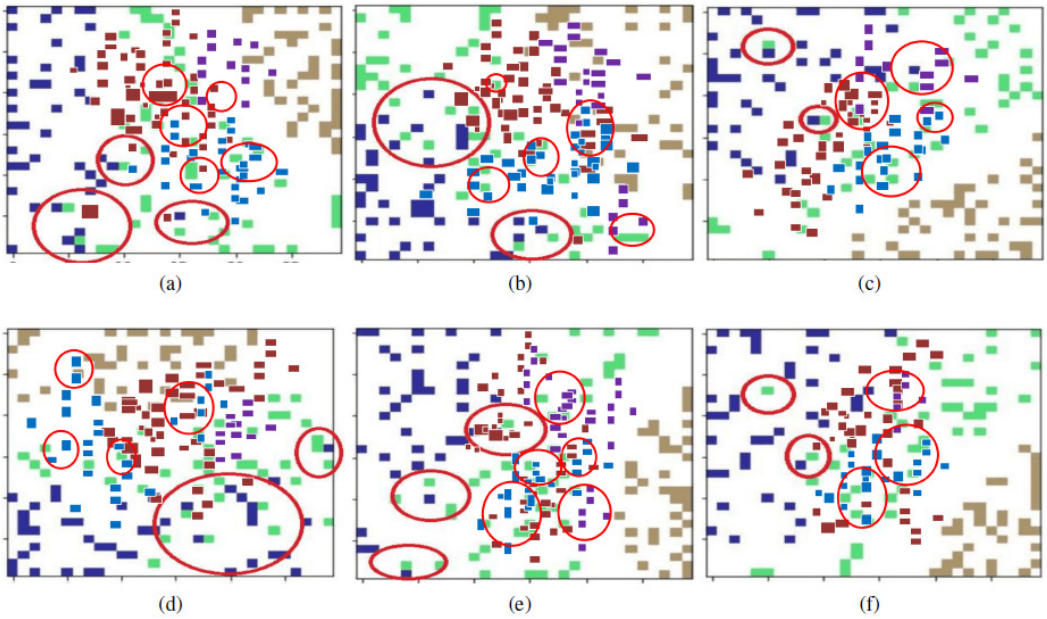

Figure 5. Results for glass data set after: (a) 4000 iterations of KSOM, (b) 4000 iterations of M-KSOM, (c) 4000 iterations of I-KSOM, (d) 5000 iterations of KSOM, (e) 5000 iterations of M-KSOM, and (f) 5000 iterations of I-KSOM

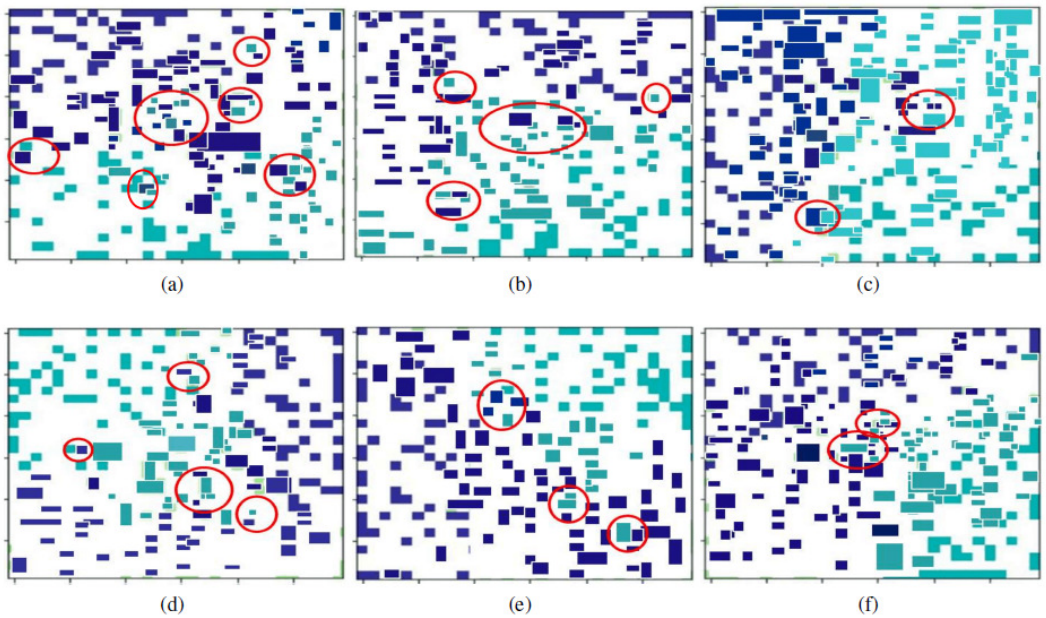

Figure 6. Results for vertebral column data set after: (a) 4000 iterations of KSOM, (b) 4000 iterations of M-KSOM, (c) 4000 iterations of I-KSOM, (d) 5000 iterations of KSOM, (e) 5000 iterations of M-KSOM, and (f) 5000 iterations of I-KSOM

Figure 5 and 7 represents the result for Glass and Wisconsin Breast Cancer data sets separately. Glass data set has six clusters named build wind non float, containers, headlamps, tableware, vehic wind float, and wind float. On the other hand, Wisconsin Breast Cancer data set contains two clusters: benign and malignant. In these two cases, the results contain more outliers even though our algorithm shows superior results than other algorithms. This happens because of curse of high-dimensionality of data.

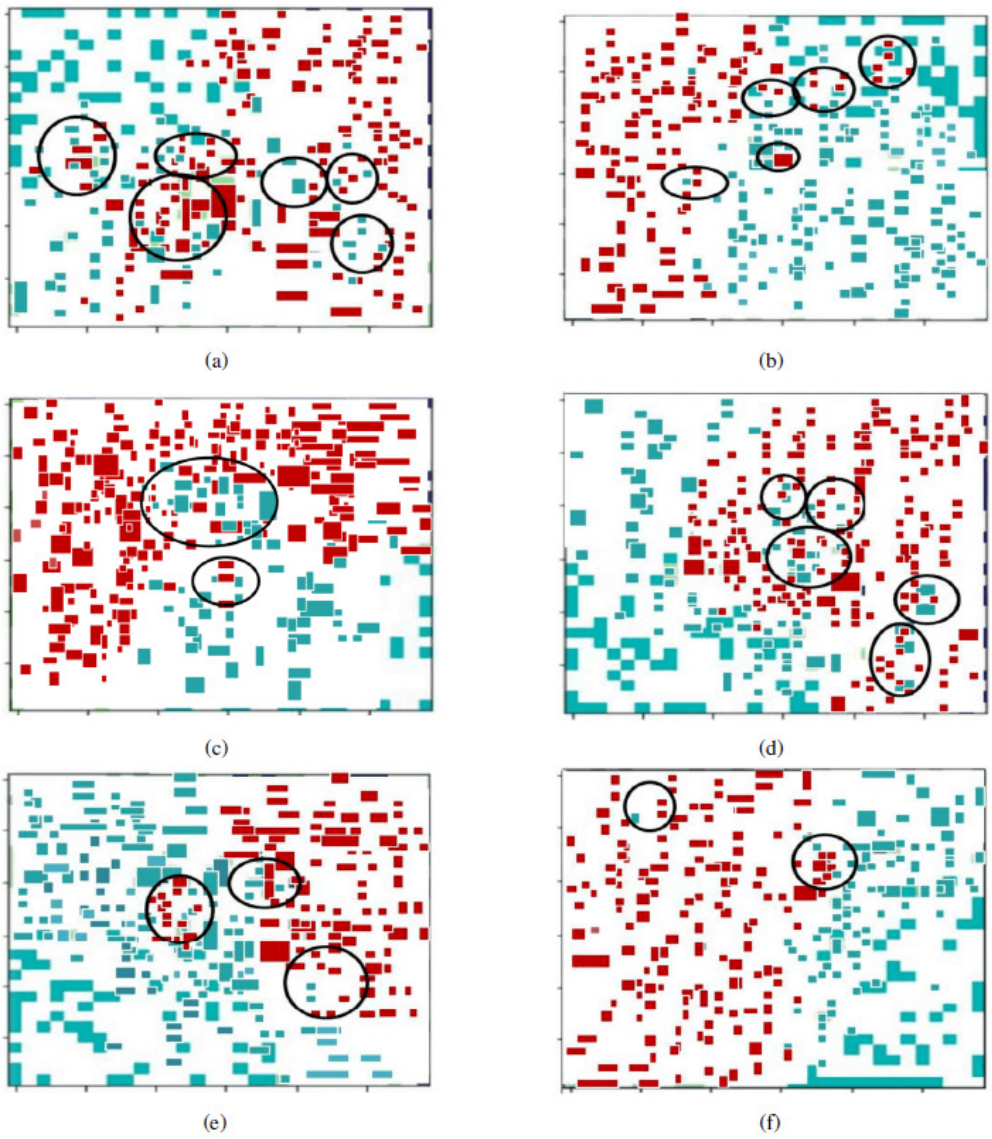

Figure 7. Results for wisconsin breast cancer data set after: (a) 4000 iterations of KSOM, (b) 4000 iterations of M-KSOM, (c) 4000 iterations of I-KSOM, (d) 5000 iterations of KSOM, (e) 5000 iterations of M-KSOM, and (f) 5000 iterations of I-KSOM

Quantization errors of all the algorithms for each data set are represented in Figure 8. The X-axis of each graph presents number of iterations and the Y-axis shows the error of the algorithm. In all cases, we vary the iteration # from 1000 to 5000 increased by 1000. From this figure, we can see that each algorithm outperforms with the increase of number of iterations. Additionally, I-KSOM gives less errors than others in Iris (a) and Seeds (b). For Glass set, though M-KSOM gives better result, our algorithm provides slighter error than original KSOM. Subsequently, each algorithm gives more errors in Vertebral and Breast Cancer data sets than other data sets.

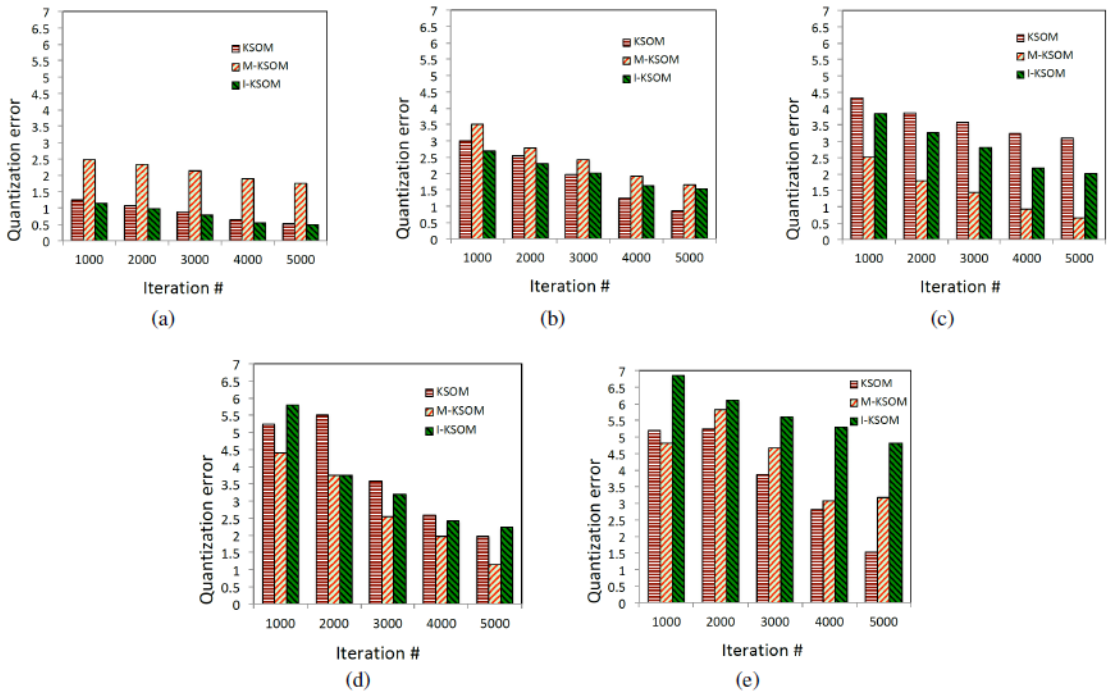

Figure 8. Results of quantization error for different data sets: (a) iris, (b) seeds, (c) glass, (d) vertebral column, and (e) Wisconsin breast cancer

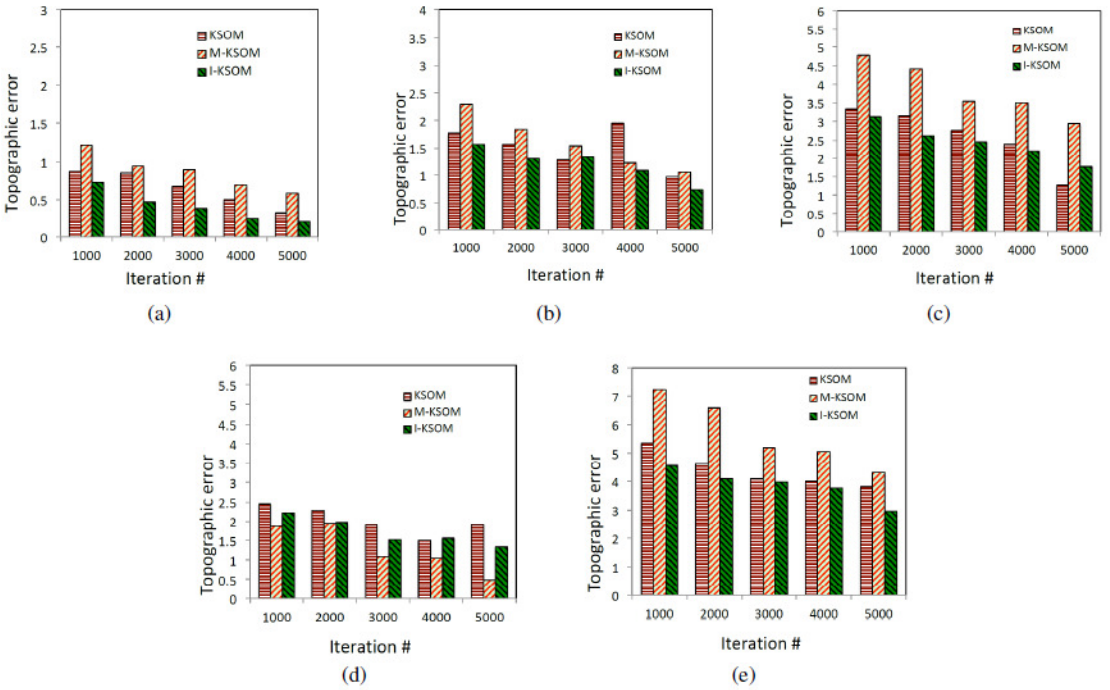

Figure 9. Results of topographic error for different data sets: (a) iris, (b) seeds, (c) glass, (d) vertebral column, and (e) Wisconsin breast cancer

Figure 9 demonstrates the topographic errors of each algorithm for all data sets as well. In this case, the graphs also are represented as like Figure 8. From the figure, we see that our algorithm shows overall finer

results than other algorithms for all data sets. Overlapping and outliers are critical issues in data clustering, classification, and decision making in the field of high-dimensional data analysis. Hence, we can conclude that our algorithm can be used significantly in this field.

6. CONCLUSION

In this paper, we introduce an improved KSOM (I-KSOM) algorithm using EISEN Cosine correlation distance to cluster the high-dimensional data sets more efficiently. Our algorithm performs better for clustering all five data sets.In addition, the errors are less than existing other algorithms. It should be reported that neural network works with initial weight that is chosen randomly which may affect the results of high-dimensional data clustering. Our future plan is to optimize initial weights of the KSOM and work on different categorical data for finding out the percentage of overlapping. Our plan also includes to apply Pearson correlation distance measure to original KSOM and compare the results with our I-KSOM for high-dimensional data sets.

REFERENCES

[1] S. Guha, R. Rastogi, and K. Shim, "Cure: an efficient clustering algorithm for large databases," *ACM SIGMOD Record*, vol. 27, no. 2, pp. 73-84, 1998, doi: 10.1145/276305.276312.

[2] W. Wang and J. Yang, "Mining High-Dimensional Data," In: Maimon O., Rokach L. *Data Mining and Knowledge Discovery Handbook*, Springer, Boston, MA, pp. 793-799, 2005, doi: 10.1007/0-387-25465-X_37.

[3] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, pp. 107-145, 2001, doi: 10.1023/A:1012801612483.

[4] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American statistical Association*, vol. 97, no. 458, pp. 611-631, 2002, doi: 10.1198/016214502760047131.

[5] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological cybernetics*, vol. 43, pp. 59-69, 1982, doi: 10.1007/BF00337288.

[6] A. Ahmad and R. Yusof, "A modified kohonen self-organizing map (ksom) clustering for four categorical data," *Jurnal Teknologi*, vol. 78, no. 6-13, pp. 75-80, 2016, doi: 10.11113/jt.v78.9275.

[7] C. Fernandes, A. M. Mora, J. J. Merelo, V. Ramos, and J. L. J. Laredo, "Kohonants: a self-organizing ant algorithm for clustering and pattern classification," *arXiv preprint arXiv:0803.2695*, 2008.

[8] T. Kovacs and A. Ko, "Monitoring pneumatic actuators' behavior using real-world data set," *SN Computer Science*, vol. 1, no. 196, 2020, doi: 10.1007/s42979-020-00202-2.

[9] P. Mallick, O. Ghosh, P. Seth, and A. Ghosh, "Kohonen's Self-organizing Map Optimizing Prediction of Gene Dependency for Cancer Mediating Biomarkers,"*Emerging Technologies in Data Mining and Information Security. Advances in Intelligent Systems and Computing*, 2019, pp. 863-870, doi: 10.1007/978-981-13-1501-5_75.

[10] A. Onishi, "Landmark map: An extension of the self-organizing map for a user-intended nonlinear projection," *Neurocomputing*, vol. 388, pp. 228-245, 2020, doi: 10.1016/j.neucom.2019.12.125.

[11] A. Ratre and V. Pankajakshan, "Tucker visual search-based hybrid tracking model and fractional kohonen self-organizing map for anomaly localization and detection in surveillance videos," *The Imaging Science Journal*, vol. 66, no. 4, pp. 195-210, 2017, doi: 10.1080/13682199.2017.1396405.

[12] K. Lagus, T. Honkela, S. Kaski, and T. Kohonen, "Websom for textual data mining," *Artificial Intelligence Review*, vol. 13, no. 5-6, pp 345–364, 1999, doi: 10.1023/A:1006586221250.

[13] Z. Rezaei, M. D. Kasmuni, A. Selamat, M. S. M. Rahim, G. Abaei, and M. R. A. Kadir, "Comparative study of clustering algorithms in order to virtual histology (vh) image segmentation," *Jurnal Teknologi*, vol. 75, no. 2, pp. 133–139, 2015, doi: 10.11113/jt.v75.4994.

[14] J. Mira and A. Prieto, "Connectionist Models of Neurons, Learning Processes, and Artificial Intelligence," *6th International Work-Conference on Artificial and Natural Neural Networks, IWANN 2001 Granada, Spain, June 13–15, 2001 Proceedings, Part 1*, 2001, doi: 10.1007/3-540-45720-8.

[15] T. Villmann and H.-U. Bauer, "Applications of the growing self-organizing map," *Neurocomputing*, vol. 21, no. 1-3, pp. 91-100, 1998, doi: 10.1016/S0925-2312(98)00037-X.

[16] R. Amarasiri, D. Alahakoon and K. A. Smith, "HDGSOM: a modified growing self-organizing map for high dimensional data clustering," *Fourth International Conference on Hybrid Intelligent Systems (HIS'04)*, 2004, pp. 216-221.

[17] C.-C. Hsu, K.-M. Wang, and S.-H. Wang, "GViSOM for Multivariate Mixed Data Projection and Structure Visualization," *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, 2006, pp. 3300-3305, doi: 10.1109/IJCNN.2006.247327.

[18] M. Begum and M. N. Akthar, "KSOMKM: An Efficient Approach for High Dimensional Dataset Clustering," *International Journal of Electrical Energy*, vol. 1, no. 2, pp. 102-107, 2013, doi: 10.12720/ijoee.1.2.102-107.

[19] A. Ahmad and R. Yusof, "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*, 2014, pp. 107-113.
[20] V. -E. Neagoe, R. -M. Stoica, A. -I. Ciurea, L. Bruzzone and F. Bovolo, "Concurrent self-organizing maps for supervised/unsupervised change detection in remote sensing images," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 8, pp. 3525-3533, Aug. 2014, doi: 10.1109/JSTARS.2014.2330808.
[21] H. Merdun, "Self-organizing map artificial neural network application in multidimensional soil data analysis," *Neural Computing and Applications*, vol. 20, pp. 1295-1303, 2011, doi: 10.1007/s00521-010-0425-1.
[22] Z. M. Zin, M. Khalid, E. Mesbahi, and R. Yusof, "Data clustering and topology preservation using 3d visualization of self organizing maps," in *Proceedings of the World Congress on Engineering*, vol. 2, 2012, pp. 696-701.
[23] X. Liu and H. Fu, "An effective clustering algorithm with ant colony." *Journal of Computers*, vol. 5, no. 4, pp. 598-605, 2010, doi: 10.4304/jcp.5.4.598-605.
[24] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *Database Theory — ICDT'99*, C. Beeri and P. Buneman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 217-235, doi: 10.1007/3-540-49257-7_15.
[25] Z. Qin, H. Lian, T. He, and B. Luo, "Cluster correction on polysemy and synonymy," in *2017 14th Web Information Systems and Applications Conference (WISA)*, 2017, pp. 136-138, doi: 10.1109/WISA.2017.45.
[26] R. Gentleman, B. Ding, S. Dudoit, and J. Ibrahim, "Distance measures in dna microarray data analysis," *(eds) Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, New York, pp. 189-208, 2005, doi: 10.1007/0-387-29362-0_12.
[27] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml
[28] Y.-S. Lin, Y. Jiang, and S.-J. Lee, "A similarity measure for text classification and clustering," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 7, pp. 1575-1590, July 2014, doi: 10.1109/TKDE.2013.19.
[29] K. Kiviluoto, "Topology preservation in self-organizing maps," in *Proceedings of International Conference on Neural Networks (ICNN'96)*, 1996, pp. 294-299 vol.1, doi: 10.1109/ICNN.1996.548907.