Steve Jaimes
Ren Hao Wong
Zheng Wei Ng
Erik Rodriguez

K-means and Self-Organizing Maps on Topological Preservation and their Performance on Increased Number of Clusters and Nodes.

Abstract

This project compares the performance of two clustering algorithms, Kmeans and Self-Organizing Maps. While k-means is a popular and widely used clustering algorithm, SOM offers a different approach to clustering that may have advantages in certain applications. This is done by examining the effects of an increasing number of clusters(k-means) and nodes(SOMS) and evaluating their overall performance. The data set used for this comparison is the Iris dataset.

Introduction

Clustering is an important task in data mining and machine learning that helps group similar data points together. Two of the most popular clustering algorithms are the K-means and Self-Organizing Maps. The K-means is a non-topologically preserving algorithm that partitions data into a defined number of clusters. The SOM algorithm is a topologically preserving algorithm that maps data onto a low-dimensional grid. This project compares the performance of K-means and SOM on the Iris dataset, examining the clustering outcomes and investigating the impact of varying the number of output SOM nodes and K-means clusters. The Iris dataset, widely used as a benchmark in machine learning, contains measurements of iris flower attributes. By applying K-means and SOM to this dataset, this study aims to analyze their clustering results and gain insights into their behavior under different configurations. This project builds on related work such as a study by Sueli A. Mingoti and Joab O. Lima titled "Comparing SOM neural network with Fuzzy c-means, K-means, and traditional hierarchical clustering algorithms", published by European Journal of Operational Research. Their research provides insights into the performance comparison of SOM with other clustering algorithms, including K-means, which aligns with the objective of this project. Another study relevant to this is a study by Chen, Y., Qin, B., Liu, T., Liu, Y., and Li, S. titled "The comparison of SOM and K-means for text clustering" in the journal Computer and Information Science. Their study focuses on clustering text and also compares the effectiveness of SOM and Kmeans. By incorporating insights from related works, such as the study on text clustering by Chen et al., this project expands on the existing research by investigating and comparing the performance of K-means and SOM algorithms on the Iris dataset, providing valuable insights for the selection of suitable clustering techniques in various domains.

Methods

Using MATLAB R2022b, the performance of K-means was tested with different values of clusters, k, and the performance of Self-Organizing Maps (SOMs) was tested with different sizes of square dimensions. The data set used is the well-known Fisher's Iris data set from 1936, which is included in the MATLAB Statistics Toolbox. The data set contains 150 iris specimens from 3 species (setosa, versicolor, virginica), with 50 specimens from each species. Each specimen has 4 features that the algorithms may utilize for clustering.

The K-means approach is tested with k = 2, 3, 4,..., 100 whereas the SOMs approach is tested with square dimensions of 4, 9, 16,..., 100. Additional tests for SOMs were also done with dimensions using multiples of 1, 2, and 3. All elements within each cluster are given shared labels (species), and the new labels are compared against their originals to evaluate the models' accuracy.

To assess the accuracy of the algorithms across different numbers of clusters and nodes, graphs were created to show their corresponding clustering, and their accuracy was recorded by comparing the percentage of correct labeling of the elements with their actual species. The percentage of elements per cluster/node is also calculated by dividing the mean number of elements in each cluster/node by the sample size of the data set.

For the implementation of the K-means algorithm, the beginning clusters were chosen at random. Then each point is assigned to its nearest centroid based on the calculated Euclidean distance. The mean of the data points belonging to each cluster was then computed to update the centroid of that cluster. This process was iterated through a loop, continually reassigning points. This process was done for each number of K, 2-100.

An SOM is calculated by first initializing a grid in the specified dimensions, then repeatedly updating the weight of each node of the grid to move them closer to data points. Our approach made use of the selformap() and train() functions, and tested with dimensions of:

$Squares$: $(2 \times 2,\ 3 \times 3,\ 4 \times 4,\ldots,\ 10 \times 10)$,
$x \times n; n = 1$: $(2 \times 1,\ 3 \times 1,\ 4 \times 1,\ldots,\ 10 \times 1)$,
$x \times n; n = 2$: $(2 \times 2,\ 3 \times 2,\ 4 \times 2,\ldots,\ 10 \times 2)$,
$x \times n; n = 3$: $(2 \times 3,\ 3 \times 3,\ 4 \times 3,\ldots,\ 10 \times 3)$,
$n \times y; n = 1$: $(1 \times 2,\ 1 \times 3,\ 1 \times 4,\ldots,\ 1 \times 10)$,
$n \times y; n = 2$: $(2 \times 2,\ 2 \times 3,\ 2 \times 4,\ldots,\ 2 \times 10)$,
$n \times y; n = 3$: $(3 \times 2,\ 3 \times 3,\ 3 \times 4,\ldots,\ 3 \times 10)$

Where grids of transposed dimensions were also tested to see the effects of transposed grids on SOMs.

Analysis and Results

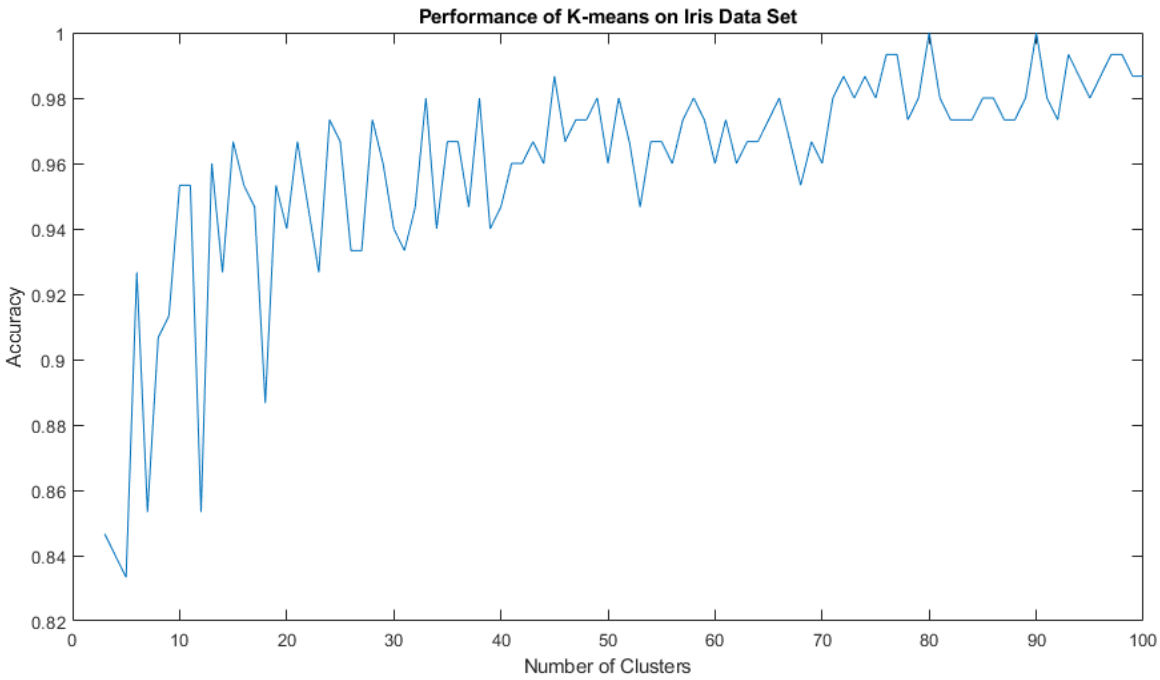


Figure 1: Performance of K-means across the different number of nodes

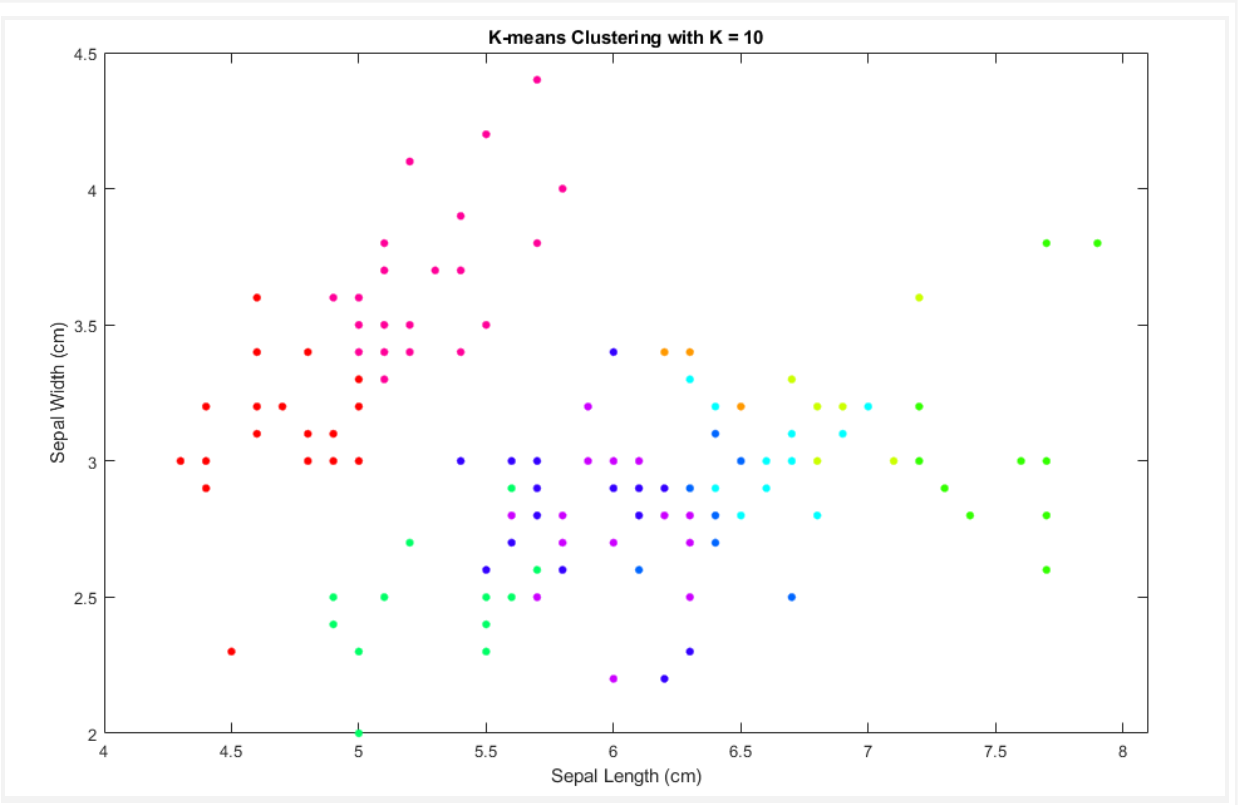


Figure 2: K-means Clustering with 10 Clusters

Applying the k-means algorithm on the iris dataset showed that the increase in clusters led to a more accurate classification of the data points. Testing the number of clusters from 3 to 100 revealed that the accuracy of the classification was shown to increase steadily. Starting with a cluster number of 3, the accuracy of classification was 84.67%. The accuracy slowly increases and hits 100% when the number of clusters reaches 80. This result indicates that a bigger value of K would be more suitable for this dataset, as it enables a more precise and accurate clustering of the data. Beyond a certain point, however, increasing the number of clusters may not always result in accuracy improvements and may even cause overfitting. A drawback of using a larger value of K is that it can result in overfitting the data, leading to poorer performance on unseen data. A higher amount of clusters can result in clusters that are too small. This makes it difficult to draw meaningful conclusions from the resulting clusters. It is best to select a meaningful K, to have a balance between accuracy and generalizability. Having something like K = 10 would be more meaningful as it has an accuracy of over 95% while also keeping the clustering somewhat distinguishable as shown in Figure 2. While it may be possible to calculate the distances between the cluster centroids, it is still hard to visualize the relationship between the clusters obtained through the K-means algorithm. This data set represents the specimens with 4 features, meaning they exist in a four-dimensional space, and it is hard to perceive the cluster positions in a multidimensional environment.

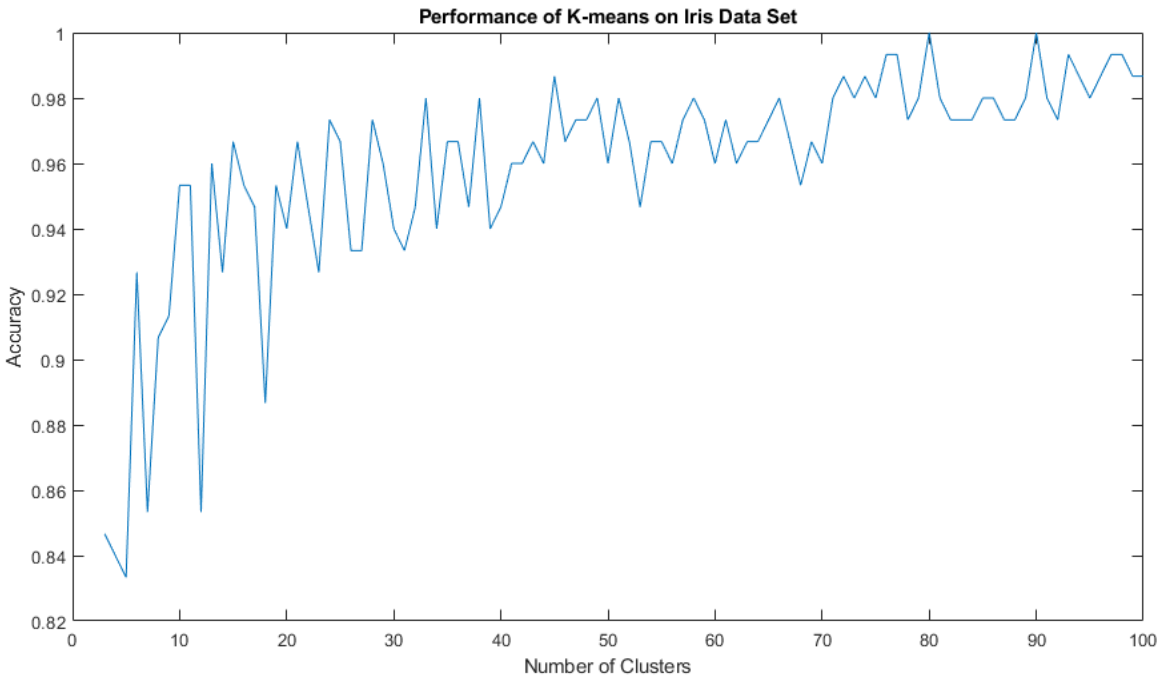


Figure 1: Performance of K-means across the different number of nodes

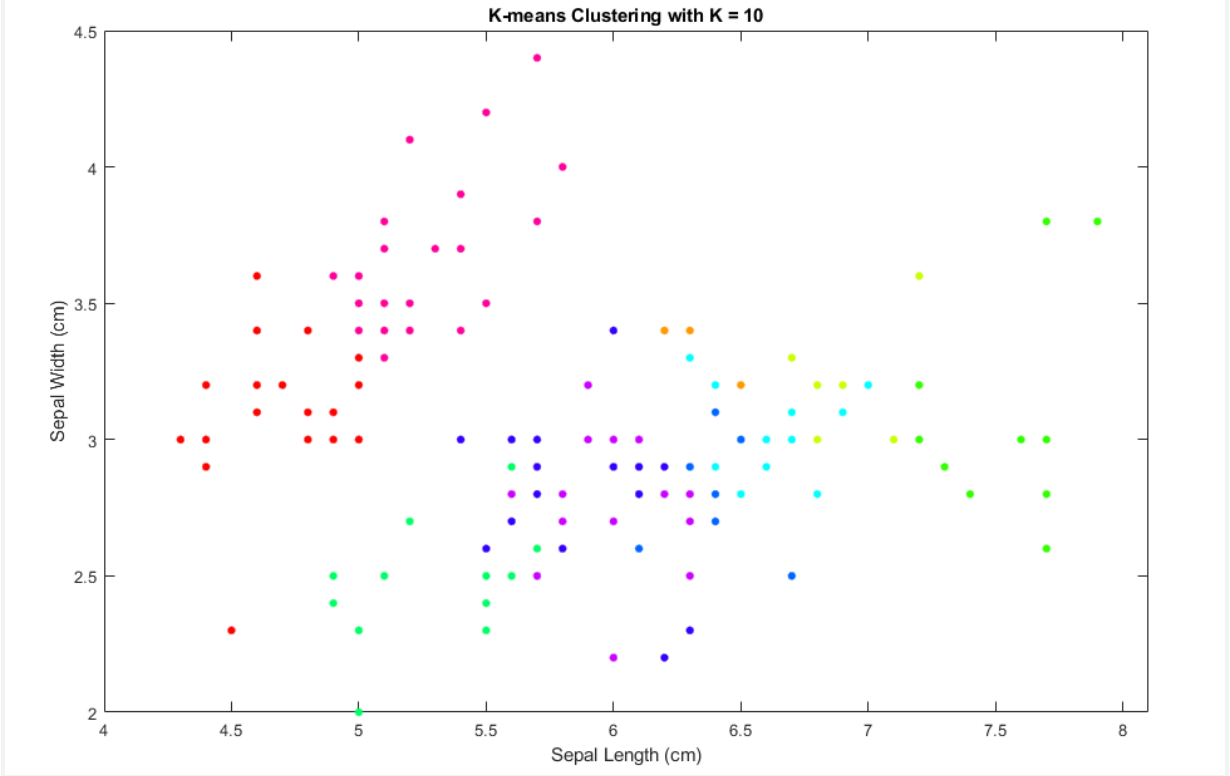


Figure 2: K-means Clustering with 10 Clusters

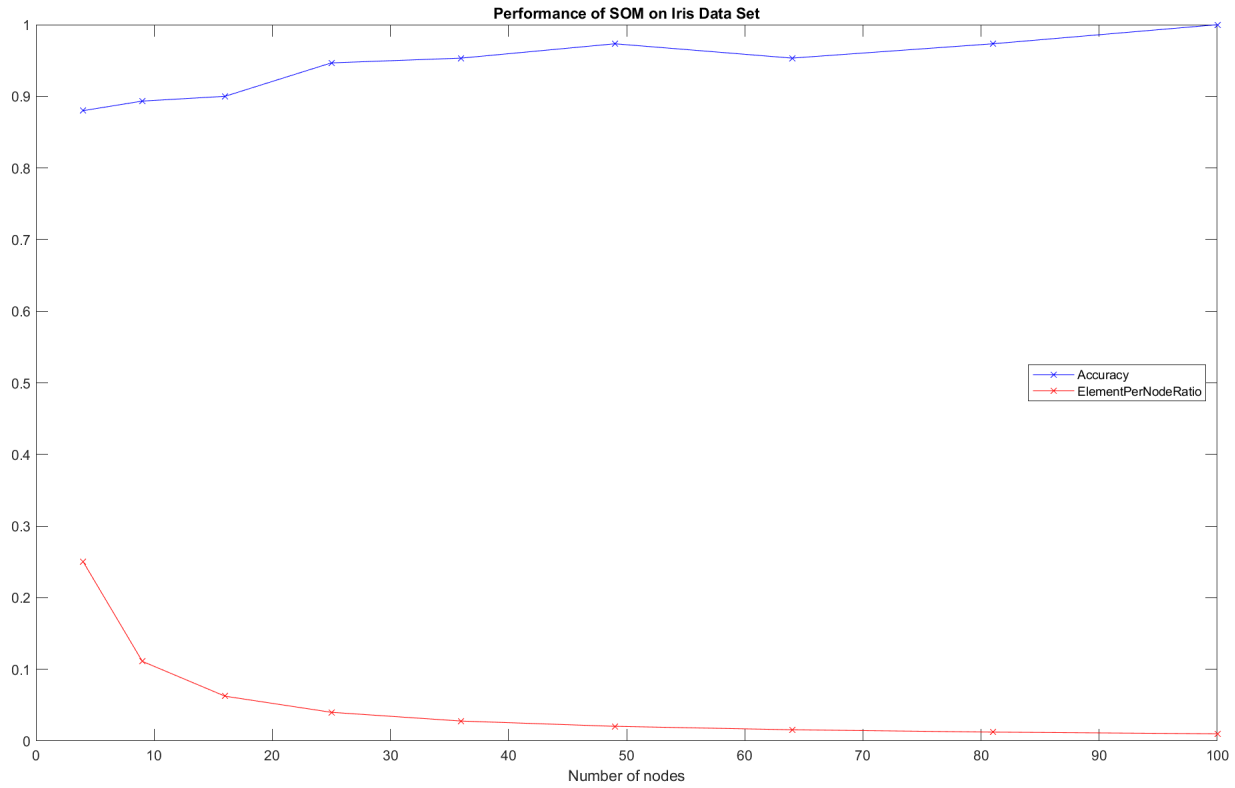


Figure 3: Performance of SOM across the different numbers of nodes

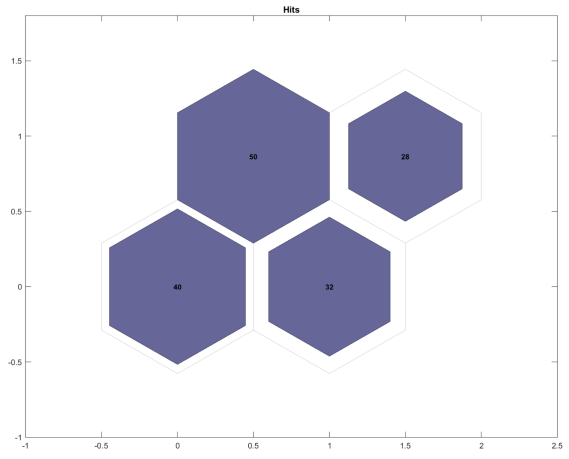


Figure 4: Element distribution in each SOM node for a 2-by-2 dimension

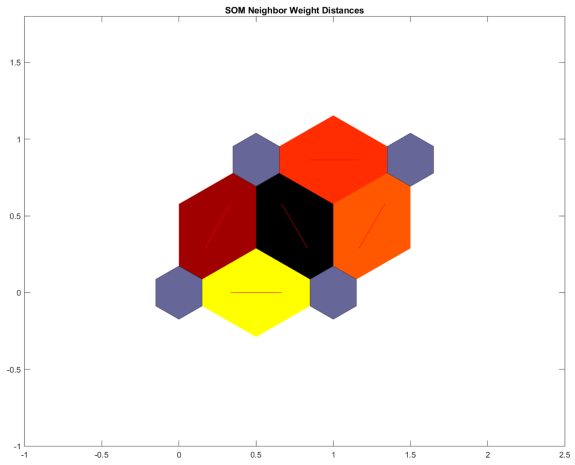
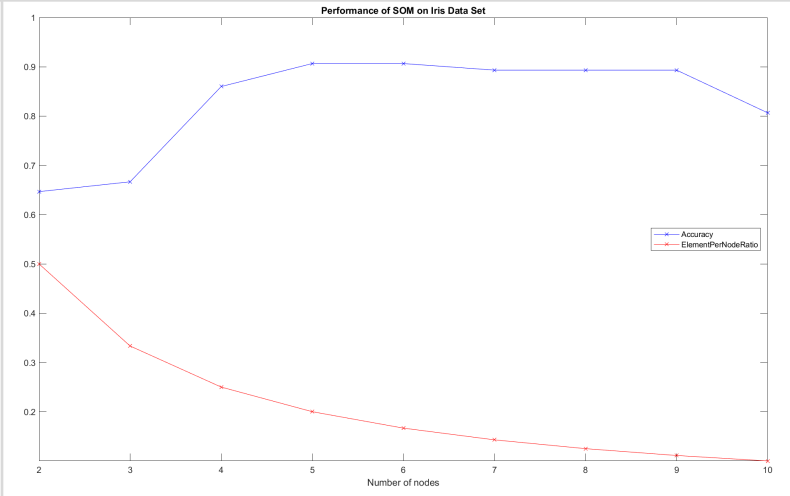
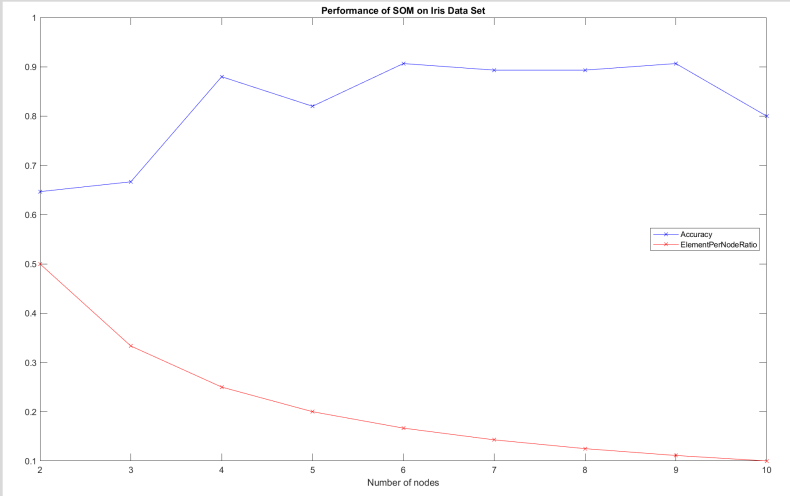


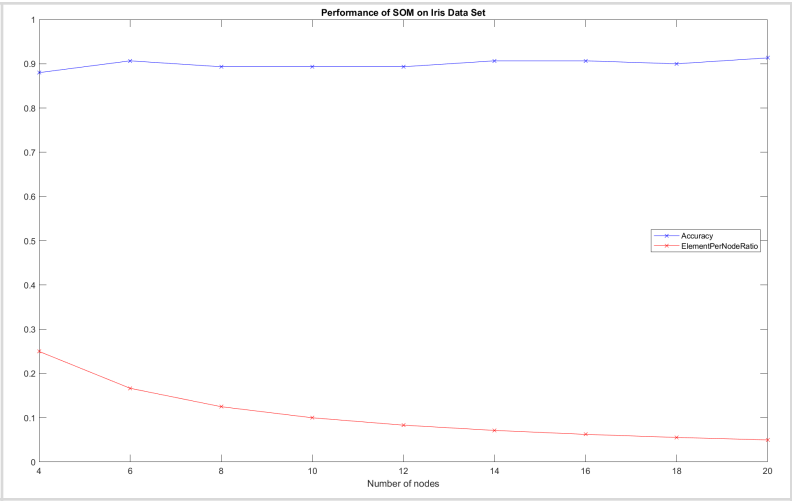
Figure 5: Neighbor weight distances between SOM nodes for a 2-by-2 dimension

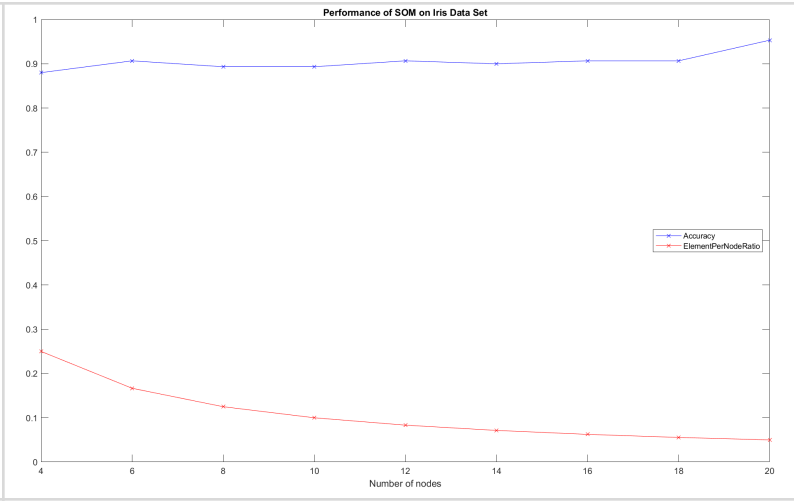
The accuracy of SOM is as expected where it increases when the number of nodes increases, with a great start of obtaining around 88% accuracy at the smallest tested dimension, a 2-by-2. The accuracy growth reaches diminishing returns after 5-by-5 nodes and eventually approaches 100% accuracy at 10-by-10 nodes. However, the sample size used is only 150 specimens, and at 100 nodes, each node only contains 1% of the sample size, around 1-2 elements per node. This observation at near 100% accuracy is not meaningful because the SOM simply clustered the elements at a near 1-to-1 grouping without really relating similar elements under the same node. With an exponentially decreasing graph of the number of elements per node, the distribution actually dropped to 4% (6 elements per node) and below starting at the dimension of 5-by-5. The trade-off between better accuracy and keeping data related meaningfully is very much worth considering when designing models using SOM. At 4 nodes with an accuracy of 88% and a high element-per-node ratio of around 25%, the square dimension of a 2-by-2 seems to be the optimal setting for SOM on the iris data set. From Figures 4 and 5, we can see that the majority of elements are grouped in the top left node, which is the most distinct node in relation to the other nodes (Darker color indicates further distance). The setosa irises are highly likely to be grouped into this node, as their features are more recognizable from the rest, whereas versicolor and virginica share some similarities in features so their nodes are closer to each other.

Table 1: Comparison between the Performances of SOM at Different Grid Transpositions

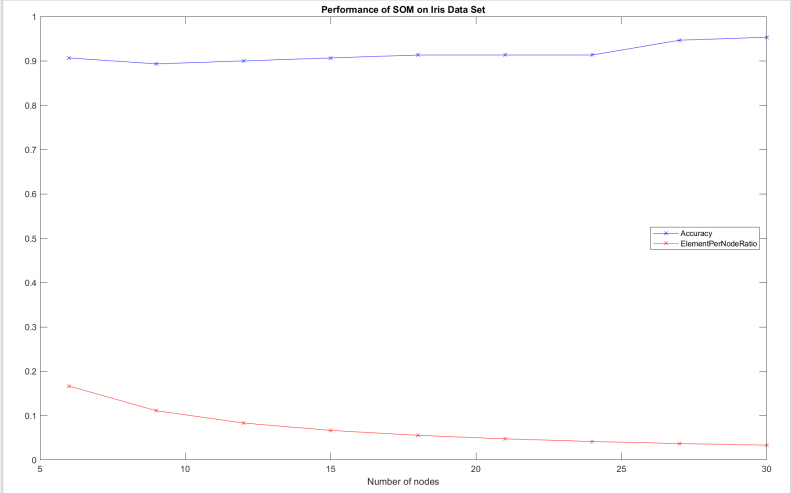
n	$x \times n$	$n \times y$
1		

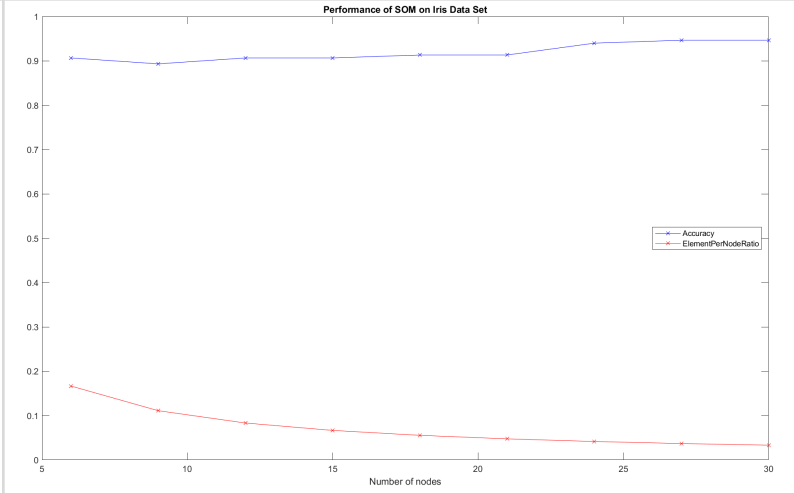
2





3





The new results from Table 1 tell that grid transposition can affect the accuracy of a SOM especially when the number of nodes is low. According to the first row of the table when $n = 1$, the difference in transposition shows clearer differences in accuracy and also encounters a drop in performance as the number of nodes increases. The tests here are essentially using a one-dimensional grid, and the poor performances due to the lack of a second dimension really show here. Looking at the second and third rows where $n = 2$ and $n = 3$, the additional dimensions helped stabilize the performance with respect to increasing nodes. The trend lines for these parameters are similar, but for the Iris data set, increments of nodes on the y-axis yielded slightly better results than increments at the x-axis. At the highest point when $n = 3$ with 27 and 30 nodes, the accuracies achieved is around 95%, coming very close to the initial results on square dimensions at a similar number of nodes, which is 25 nodes or 5×5. At 3 nodes, which is the number of species of Iris, the accuracies obtained from both the 3×1 and 1×3 grids are also worse than that of the results from k-means.

Discussion

The comparison between K-means and SOM shows the uniqueness of each algorithm. The topological preservation property of SOM enables it to capture and preserve neighborhood

relationships in the data. The SOM algorithm is constrained to a square grid, which can limit its flexibility, unlike the K-means where the number of clusters is freely determined. The choice between K-means and SOM depends on the specific requirements of the problem at hand that have other factors to consider such as computational efficiency, and the nature of the data. Increasing the number of clusters/nodes in K-means and SOM can have improving effects on their accuracy but finding the optimal balance between better accuracy and meaningful classification is crucial. The 2D neighbor weight distance visualization provided by a SOM algorithm also proved that it is topologically preserving and provides an alternative usage to the K-means algorithm that focuses on high dimensional distances. The K-means algorithm on the other hand shows higher flexibility in regard to its number of clusters, whereas a SOM is required to confine its number of nodes to a quad. Future research can further explore the impact of varying other parameters, such as the initialization method in K-means and the learning rate in SOM, on their performance. Additionally, investigating the applicability of these algorithms on different datasets with varying characteristics would provide a greater understanding of their strength on different datasets.

Taking into consideration the feedback received during the presentation of this study, additional tests were conducted on varying dimensions of the SOM grid. These tests showed that differences in results can be impacted based on not just the number of nodes, but also the dimensions on each axis. Scaling the size on each dimension for training a SOM is also dependent on the data set, referring to the findings of this study that the Iris data set observes differences on the y-axis slightly more than the x-axis.

Conclusion

This study compared the performance of K-means and Self-Organizsing Maps on the iris dataset on different numbers of increasing clusters and nodes. Our findings indicate that both algorithms can benefit from an increase in the number of clusters/nodes in terms of accuracy. However, it is essential to strike a balance between improved accuracy and meaningful classification. The SOM algorithm demonstrated its topological preservation property through the 2D neighbor weight distance visualization. The K-means algorithm showed its flexibility in terms of the number of clusters it can accommodate, allowing for a more extensive range of cluster configurations. In conclusion, this study highlights the potential advantages of increasing the number of clusters/nodes in both K-means and Self-Organizing Maps algorithms, emphasizing the need for a careful balance between accuracy improvement and meaningful classification.

Works Cited:

Sueli A. Mingoti and Joab O. Lima. “Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms”.European Journal of Operational Research, Volume 174, Issue 3, 2006, Pages 1742-1759, ISSN 0377-2217.

Chen, Y., Qin, B., Liu, T., Liu, Y., & Li, S. “The comparison of SOM and K-means for text clustering. Computer and Information Science”. 3(2). 2010.