Influence of the Number of Neighbours on the Clustering Metric by Oscillatory Chaotic Neural Network with Dipole Synaptic Connections

Vasyl Lytvyn[1], Dmytro Dudyk[1], Ivan Peleshchak [1], Roman Peleshchak[1], Petro Pukach[1]

[1] *Lviv Polytechnic National University, 12 Stepana Bandera Street, Lviv, 79013, Ukraine*

Abstract

Clustering is indispensable for addressing practical challenges across diverse domains in today's data-driven environment. Given the pivotal role of technology in managing vast amounts of data, effective data grouping has become indispensable for successful operations across various domains. For instance, in marketing, clustering aids in identifying customer segments for personalized marketing, while in medicine, it supports accurate diagnosis and treatment. Similarly, in financial analysis, it is vital for detecting anomalies or fraud, and in organizing textual data, it helps uncover fundamental trends. The emergence of oscillatory chaotic neural networks with dipole interactions offers a promising novel approach to clustering, leveraging self-organizing properties to group data effectively. Understanding how the number of nearest neighbours influences clustering metrics in this method is crucial for optimizing its efficiency and applicability.

The study aims to calculate and analyse the evaluation of clustering metric values, including the Adjusted Rand Index (ARI) and silhouette coefficient (SC), concerning the number of nearest neighbours and clustering resolution to determine the optimal number of nearest neighbours for enhancing clustering quality.

Oscillatory chaotic neural networks with dipole synaptic connections between neurons were employed. To ensure a comprehensive analysis, four diverse datasets were utilized, each chosen for its distinct characteristics, representing different complexities commonly encountered in real-world data scenarios: Atom (linear inseparability), WingNut (small inter-cluster/large intra-cluster distances), TwoDiamonds (weak link connecting clusters), and EngyTime (overlapping clusters of different densities). Clustering was performed across different ranges of nearest neighbour values (Atom: 1-300, WingNut: 1-800, TwoDiamonds: 1-400, EngyTime: 1-1000) and resolution levels to comprehensively assess the influence of nearest neighbour selection on clustering quality across various data complexities.

The study revealed a significant impact of the number of nearest neighbours on clustering efficiency when employing oscillatory chaotic neural networks. Networks with dipole synaptic connections exhibited less sensitivity to changes in the number of nearest neighbours compared to those with Gaussian-based synaptic connections, indicating their robustness. Additionally, the optimal number of nearest neighbours varied across datasets and resolution levels, highlighting the need for tailored parameter selection to maximize clustering quality.

The results confirm the importance of selecting the optimal number of nearest neighbours to enhance clustering quality using an oscillatory chaotic neural network. Further research could explore additional factors influencing clustering performance.

Keywords

Data clustering, oscillatory chaotic neural network, nearest neighbours, dipole synaptic connections.

1. Introduction

In the modern scientific world, the clustering problem is used in solving practical problems across various domains. In marketing and audience segmentation, for example, effective customer clustering allows you to identification of groups of consumers with similar behavioural and

COLINS-2024: 8th International Conference on Computational Linguistics and Intelligent Systems, April 12–13, 2024, Lviv, Ukraine

vasyl.v.lytvyn@lpnu.ua (V. Lytvyn); dmytro.dudyk.mnsam.2022@lpnu.ua (D. Dudyk); ivan.r.peleshchak@lpnu.ua (I. Peleshchak); roman.m.peleshchak@lpnu.ua (R. Peleshchak); petro.y.pukach@lpnu.ua (P. Pukach)

0000-0002-9676-0180 (V. Lytvyn); 0009-0005-3831-0826 (D. Dudyk); 0000-0002-7481-8628 (I. Peleshchak); 0000-0002-0536-3252 (R. Peleshchak); 0000-0002-0359-5025 (P. Pukach)

purchasing habits, facilitating more accurate and personalized marketing strategies. In medicine, clustering patients based on medical indicators help in accurate diagnosis and individualised treatment. In financial analysis, grouping financial transactions to detect anomalies or fraud is an integral part of ensuring financial security. In telecommunications, clustering subscribers based on their service usage helps optimise the network and improve customer service. In text analytics, thematic or categorical clustering is a crucial step in identifying and understanding the main trends.

Clustering, which is a key machine learning method aimed at grouping similar objects, is of particular importance in the context of an oscillatory chaotic neural network (OCNN). The OCNN clustering method utilizes the oscillatory properties of neurons to group objects and exhibits self-organizational characteristics, enabling it to dynamically adapt to changes in input data.

The research aims to calculate and analyse the values of clustering metrics depending on the number of nearest neighbours and the resolution of data clustering to find the optimal number of nearest neighbours that will improve the quality of clustering. The selection of the optimal number of nearest neighbors not only influences the quality of clustering but also impacts the execution time of the algorithm. Finding the optimal number of nearest neighbours can improve the quality of clustering and reduce execution time. Determining the optimal resolution for data clustering aims to find clusters with similar characteristics more accurately.

The object of research is the process of clustering by an oscillatory chaotic neural network, and the subject is the influence of the number of nearest neighbours on the value of clustering metrics by an oscillatory chaotic neural network.

2. Literature review

Cluster analysis is a valuable tool in many scientific and applied fields that allows you to divide a large sample of objects into groups with similar properties. Traditional methods such as k-means and hierarchical clustering have proven to be effective, but there are challenges associated with determining the quality of clustering, choosing the number of clusters, and selecting appropriate metrics [1].

In modern research, there is an interest in using cluster synchronization in complex dynamic systems, in chaotic neural networks, to identify clusters. Cluster synchronization allows for the identification of groups of neurons that interact and share similar dynamic characteristics. Of particular interest is the phenomenon of cluster synchronization, where groups of connected dynamic systems synchronise without fully synchronising all network neurons. Cluster synchronization is relevant in neurobiology, where, for example, cluster synchronization can occur in brain neural networks, where certain groups of neurons interact and show joint activity, while other groups can function independently [2].

When employing this approach, we direct the chaotic dynamics of the network so that neurons organise themselves into synchronised clusters. Note that neurons within each cluster oscillate in a coordinated manner, while neurons within different clusters may oscillate independently or demonstrate different synchronised patterns.

The use of cluster synchronisation in the context of chaotic neural networks can be useful for cluster data analysis, especially in the presence of heterogeneities. This approach opens opportunities to identify internal structures and relationships that may be difficult to discern using other methods.

Oscillatory Chaotic Neural Network (OCNN) is a novel model of artificial neural networks. Chaos is a phenomenon of complex unpredictable and random behaviour arising from simple deterministic nonlinear systems. Leveraging the principles of chaos and neural networks allows us to solve complex problems in various fields. Aihara and his colleagues have developed a chaotic neural network model that exhibits the nonlinear characteristics inherent in artificial neural networks while also demonstrating the ergodic properties associated with chaotic systems, which allows it to be used for intelligent information processing [3].

In [4, 5], a group of Italian scientists showed that OCNN can be used to solve clustering problems. In this network, each data record is associated with an oscillatory neuron, and synaptic connections between each pair of neurons are calculated by the Gaussian function (1) based on the Euclidean distance between the corresponding data records.

Synaptic connections play a crucial role in transmitting information between neurons in neural networks, which is key to brain function and the learning process. Traditionally, their operation is explained using chemical and electrochemical mechanisms.

Experimental studies [6, 7] have established a connection between neurons in the human brain and microtubules of the cytoskeleton. These works indicate that microtubules of the cytoskeleton, composed of tubulin molecules, serve as corresponding substrates for "quantum-statistical computations" in brain neurons. Each tubulin molecule possesses a dipole moment of approximately 100D and forms a dimer consisting of α- and β-tubulins connected by a thin bridge. The tubulin dimer can exist in two different geometric configurations (conformations), corresponding to two states described in Boolean algebra as 0(↓) and 1(↑) (or -1(↓), +1(↑)). Additionally, it has been reported in [6, 8 – 10] that microtubules of the cytoskeleton optically flicker during metabolic activity, and the resonant frequencies of tubulin molecules are approximately 1011–1013Hz, indicating that neurons have their frequencies.

Based on this information, we replace the synaptic weight function of the OCNN from the Gaussian (1) to dipole (2). Each neuron in the OCNN with introduced dipole synaptic connections plays the role of a tubulin molecule.

However, one of the key parameters that affect the clustering process in the OCNN is the number of nearest neighbours. This number influences the structure and, consequently, the dynamics of clusters formed by the network during training. The number of nearest neighbours can significantly affect the clustering results in traditional methods, but its impact on the OCNN has not been studied in detail.

3. Materials and methods

The characteristics of the Oscillatory Chaotic Neural Network (OCNN) are as follows:

- The neural network is single-layered, recurrent, and fully connected.
- The network nodes are neurons with a transfer function that exhibits chaotic behavior, such as the logistic map.
- It possesses the property of non-attractiveness, meaning the neural network does not have explicit stable states or points that would attract its dynamics.
- The network's output results are hidden in the dynamics of neuron outputs, meaning that the network's operation results are reflected in the evolving neuron outputs over time rather than in stable states predetermined at the start of the network's operation.
- Each element of the dataset corresponds to one neuron in the OCNN.

In the works [4, 5], a group of Italian scientists led by L. Angelini uses the Gaussian function to calculate the synaptic weights between the neurons of the OCNN:

$$w_{ij} = \exp\left(\frac{-dist\left(r_i, r_j\right)^2}{2a^2}\right), \tag{1}$$

where $dist\left(r_i, r_J\right)$ is the Euclidean distance $dist\left(r_i, r_j\right)$between the *i*-th and *j*-th data points in a D-dimensional space, and a is the scaling constant that is the mean (typically arithmetic) of the distances between k nearest neighbours of each neuron.

In this work, the synaptic connections between neurons in the OCNN with dipole interaction are given by the function:

$$w_{ij} = \frac{a^3}{a^3 + dist\left(r_i, r_j\right)^3}. \tag{2}$$

The dynamics of an oscillatory chaotic neural network is given by evolutionary equation (3):

$$x_i(t+1) = \frac{1}{C_i}\sum_{j \neq i}^{N} w_{ij} f\left(x_j(t)\right) \quad (3)$$

where:

- N is the number of neurons,
- $f(\cdot)$ is the transfer function,
- $x_i(t)$ is the value of the i-th neuron at discrete time t, which lies in the range [-1,1],
- $C_i = \sum_{j \neq i}^{N} w_{ij}$ is the normalizing coefficient.

The logistic mapping is used as the transfer function $f(x) = 1 - bx^2$, where b is a parameter typically set to 2. This mapping demonstrates chaotic dynamics arising from sensitivity to initial conditions and nonlinearity. Using the logistic map as the transfer function for each neuron results in chaotic oscillations within the neural network.

Starting from a random initial configuration $x_i(0) \in [-1,1]$, equation (3) is computed iteratively T times. There are two-time intervals during which the system operates: a transient period, consisting of $T_p\left(0 < t \leq T_p\right)$ iterations, and the subsequent $T_n\left(T_p < t \leq T_n\right)$ iterations, which $T_p\ \left(0 < t \leq T_p\right)\ T_n\ \left(T_p < t \leq T_n\right)$ serve to gather statistical information about the oscillations of each neuron. Information about neuron activity is translated into a sequence of bits using a threshold function. This function assigns a value of 1 if the output of the neuron exceeds the threshold of 0, and 0 otherwise, indicating whether the neuron fires or not.

Based on these T_n iterations, the information matrix I is calculated, which contains mutual information for each pair of neurons. The mutual information I_{ij} between the i-th and j-th neurons is determined by the formula $I_{ij} = H_i + H_j - H_{ij}$, where H_i is the Shannon entropy for the sequence of obtained bits $I_{ij} = H_i + H_j - H_{ij}$ of the i-th neuron; H_{ij} is the joint Shannon entropy for the sequences of bits of the i-th and j-th neurons [4].

After calculating the information matrix, I, further analysis allows for solving the clustering problem. If the i-th and j-th neurons oscillate synchronously, then the value of the information I_{ij} reaches its maximum value of ln2. In the case where their oscillations exhibit chaotic behaviour, the information I_{ij} decreases to zero [4]. This approach allows for the separation of different types of dynamic interactions between neurons and determines clusters in the neural network based on the nature of their oscillations.

Clusters are formed as connected components of a graph, where connections are established between all pairs of i-th and j-th neurons for which $I_{ij} > \theta$. $I_{ij} > \theta$. The threshold value of the information matrix θ controls the resolution with which the dataset is clustered. If the value of θ is close to the minimum value in the matrix $I\left(\theta \approx \min I_{ij}\right)$, all points belong to the one cluster, and if it is close to the maximum value in $I\left(\theta \approx \max I_{ij}\right)$, all points form their clusters. However, the most interesting case for the clustering task is the intermediate value of θ, as it allows observing the formation of groups of neurons that oscillate synchronously $I\ \left(\theta \approx \min I_{ij}\right)\ I\ \left(\theta \approx \max I_{ij}\right)$.

A significant aspect of studying the impact of the number of neighbours on the clustering process of OCNN is the choice of metrics for assessing the clustering results. In this work, the Adjusted Rand Index (ARI) and the Silhouette Coefficient (SC) are used [11].

ARI is a key metric that considers the agreement between the true classes and the clusters. The uniqueness of ARI lies in its ability to adjust for random agreements, making it a reliable indicator of clustering accuracy, even in cases of heterogeneous class distribution [11]. This metric is an adjusted version of the Rand index (4), which measures the degree of overlap between two sections.

$$RI = \frac{2(p+m)}{N(N-1)}, \quad (4)$$

where p is the number of pairs of objects with the same labels and are in the same cluster, m is the number of pairs of objects with different labels that are in different clusters, and N is the number of objects in the sample.

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}, \quad (5)$$

where E is the operator of mathematical expectation.

The silhouette coefficient measures how compact and well-separated the objects within clusters are. This metric provides an assessment of both the shape and the distance between clusters. A high SC indicates successful clustering with clear distinctions between groups [11].

$$SC = \frac{1}{N}\sum_{i=1}^{N} \frac{b_i - a_i}{\max(a_i, b_i)}, \quad (6)$$

where a_i is the average distance from the i-th object to other objects in the same cluster, b_i is the average distance from the i-th object to objects in the nearest neighbouring cluster, and N is the number of objects in the sample.

The datasets used in this paper are Atom (Fig. 1(a)), WingNut (Fig. 1(b)), TwoDiamonds (Fig. 1(c)), and EngyTime (Fig. 1(d)) from the article [12].

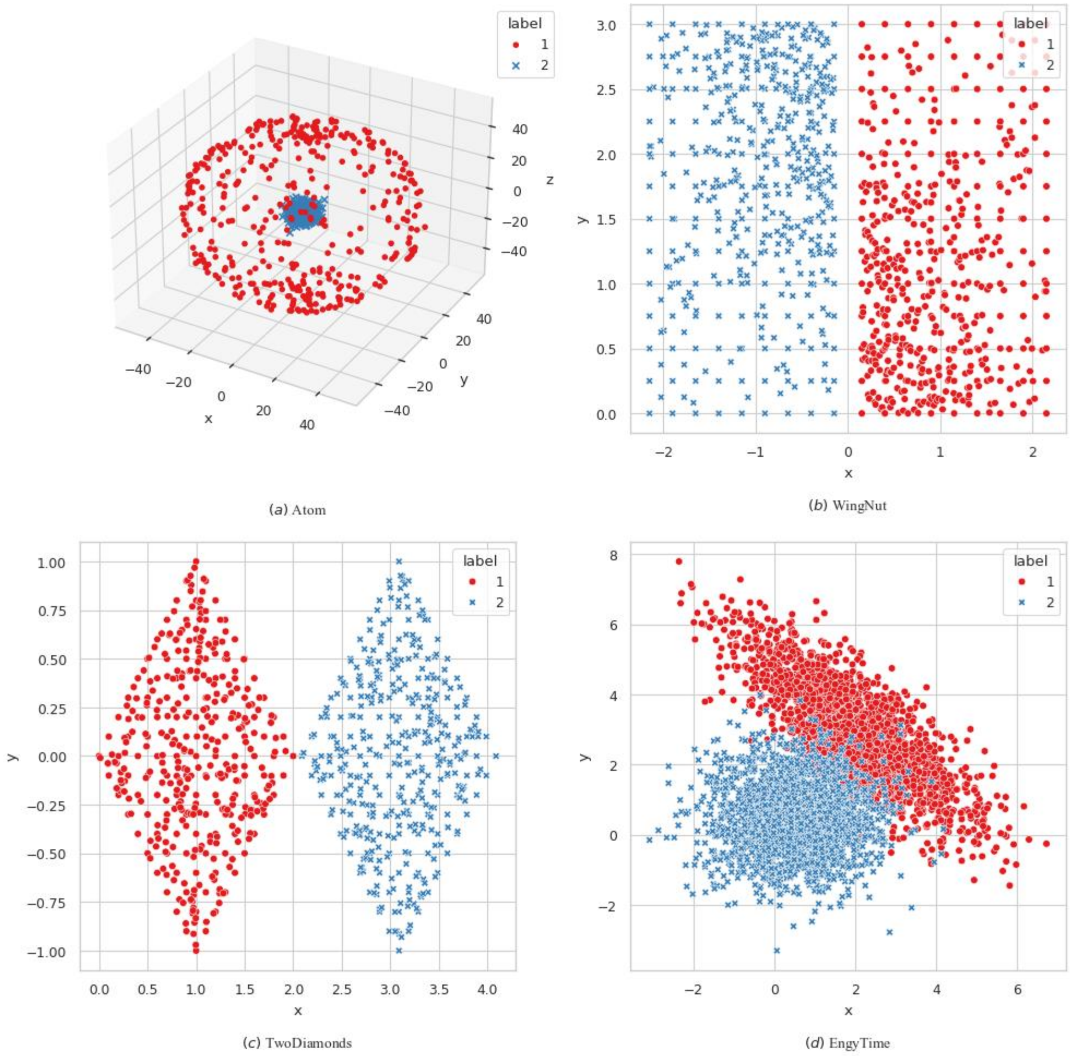


Figure 1: Datasets Atom (a), WingNut (b), TwoDiamonds (c), and EngyTime (d).

The Atom dataset consists of 400 kernel points and 400 shell points in three-dimensional space R^3. In the Cartesian metric space, the dataset is defined as linearly inseparable, with the kernel cluster entirely encompassing the shell cluster. Additionally, the density of the kernel points is significantly higher than the density of the shell points [12].

The WingNut dataset comprises two subsets of data, each containing 500 points. Each subset represents an overlay of a square grid with cells of length 0.2 and randomly positioned points with a gradually increasing density in one of the corners. Both subsets are mirrored and shifted to ensure a distance between them exceeding 0.3, providing greater spacing between than within the subsets [12].

The TwoDiamonds dataset consists of two clusters of two-dimensional points. Within each diamond, 300 points are uniformly distributed. The clusters almost touch at their corners, complicating the detection of this weak link and making this dataset challenging [12].

Another dataset, EngyTime, contains 4096 points belonging to two clusters in R^2. EngyTime is a two-dimensional mixture of Gaussian distributions. The clusters overlap, and the cluster boundaries can only be determined using density information, as there is no space between the clusters [12].

4. Computer experiment

In this section, we conduct a computer experiment to evaluate the impact of the number of neighbours on clustering metrics deterioration in networks with synaptic connections between neurons (1) and (2).

The purpose of the experiment is to determine the optimal number of neighbours for each network and dataset. The optimal number of neighbours is the number at which the network detects clusters with the finest resolution window achieving the maximum value of the defined clustering metrics.

We compute the average values of each clustering metric using 5 random initial conditions. These initial conditions are specified in equation (3) for each clustering process. The goal of using average values is to reduce the influence of randomness on the results of our experiment. In other words, we conduct multiple clustering processes with different random initial conditions and then take the average values of the metrics to obtain more robust and reliable results that account for randomness.

The first dataset is Atom, which is challenging due to its linear inseparability in the Cartesian space.

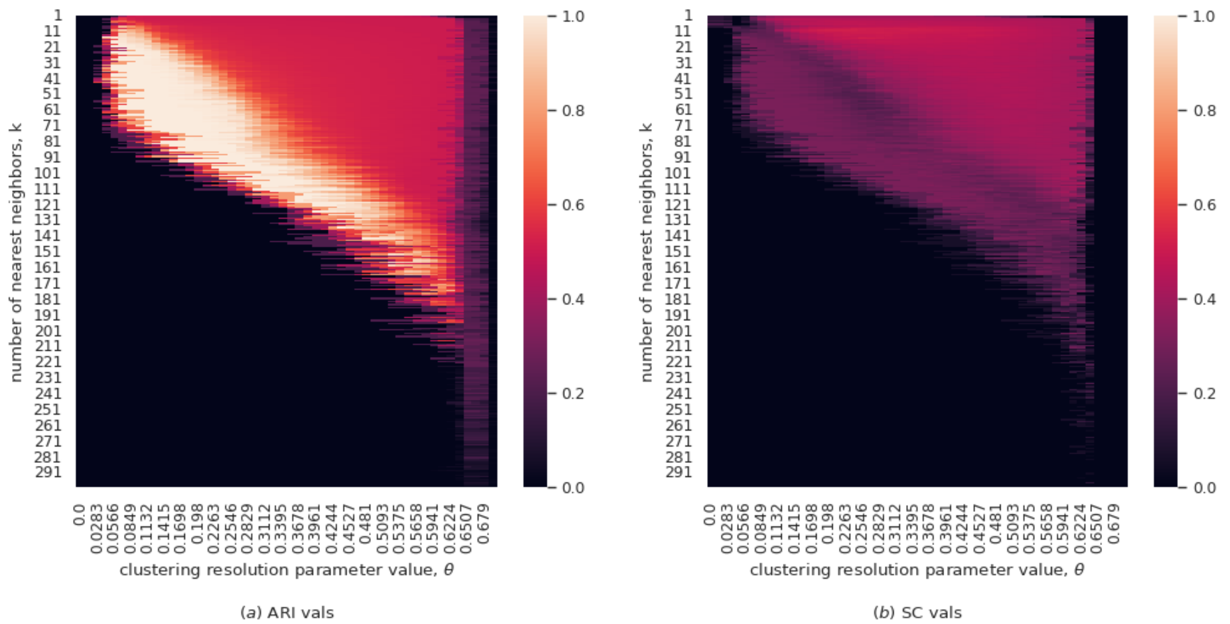
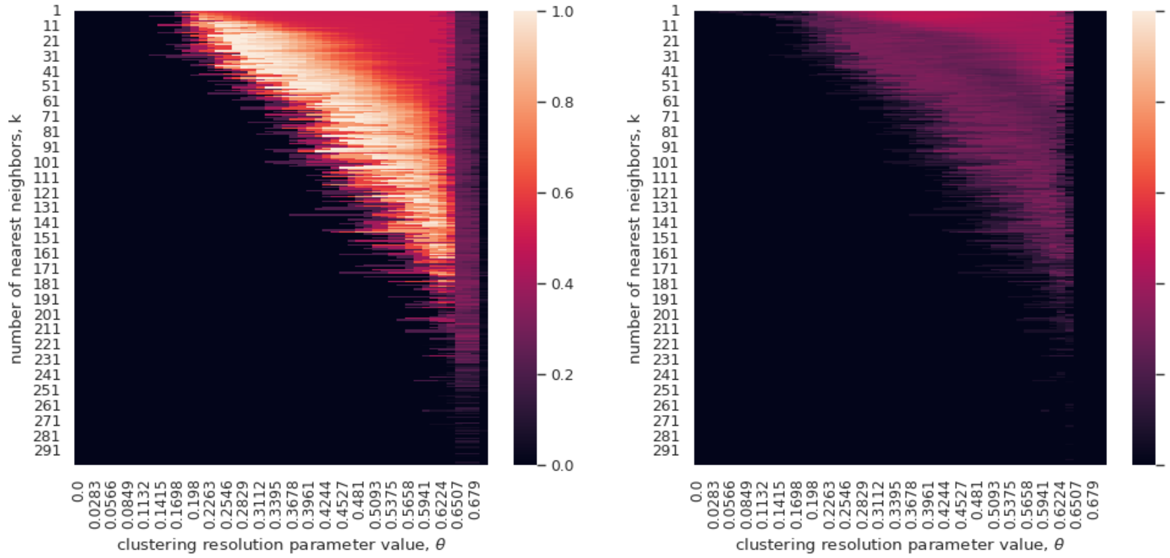


Figure 2: Average values of clustering results metrics for the network (1) of the Atom dataset under the condition of using 5 random initial conditions for iterative equation (3): (a) Adjusted Rand Index (ARI); (b) Silhouette Coefficient (SC).

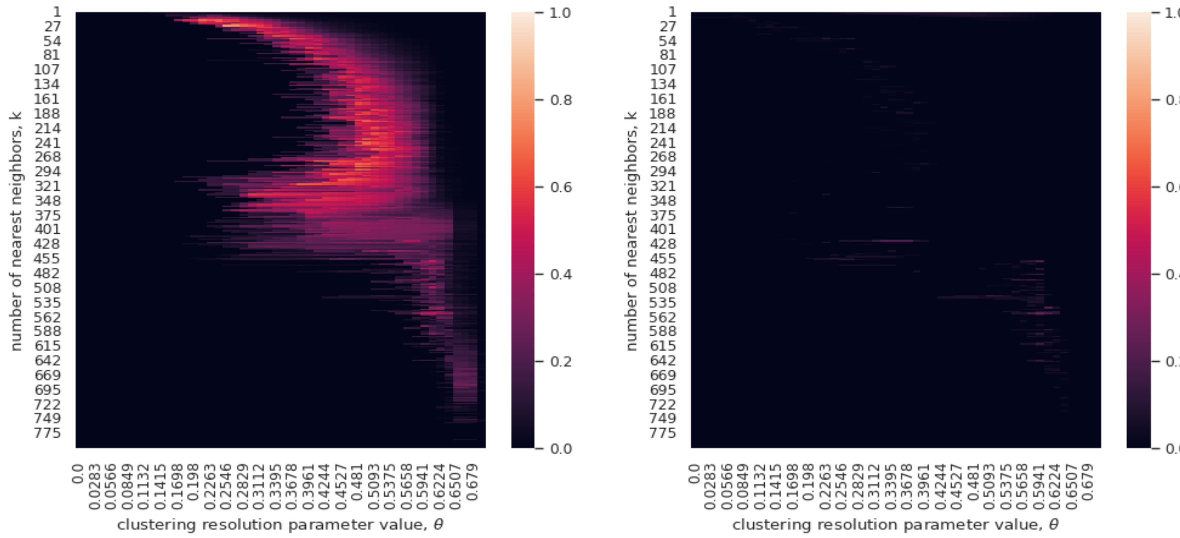
In Figure 2 (a), it is shown that as the number of nearest neighbours increases, it is necessary to increase the clustering resolution so that the ARI metric value still is unchanged. This means that with more neighbours, network (1) can detect finer differences between data points and forming clearer clusters.

The size of the parameter window θ, at which the maximum value of the ARI metric is reached, increases in the network with synaptic connections (1) as the number of nearest neighbours increases. However, upon reaching a certain value (approximately k=75), the size of the window begins to decrease, showing the onset of excessive network complexity and the possibility of falsely detected clusters.



(a) ARI vals (b) SC vals

Figure 3: Average values of clustering results metrics for the network (2) of the Atom dataset under the condition of using 5 random initial conditions for iterative equation (3): (a) Adjusted Rand Index (ARI); (b) Silhouette Coefficient (SC).

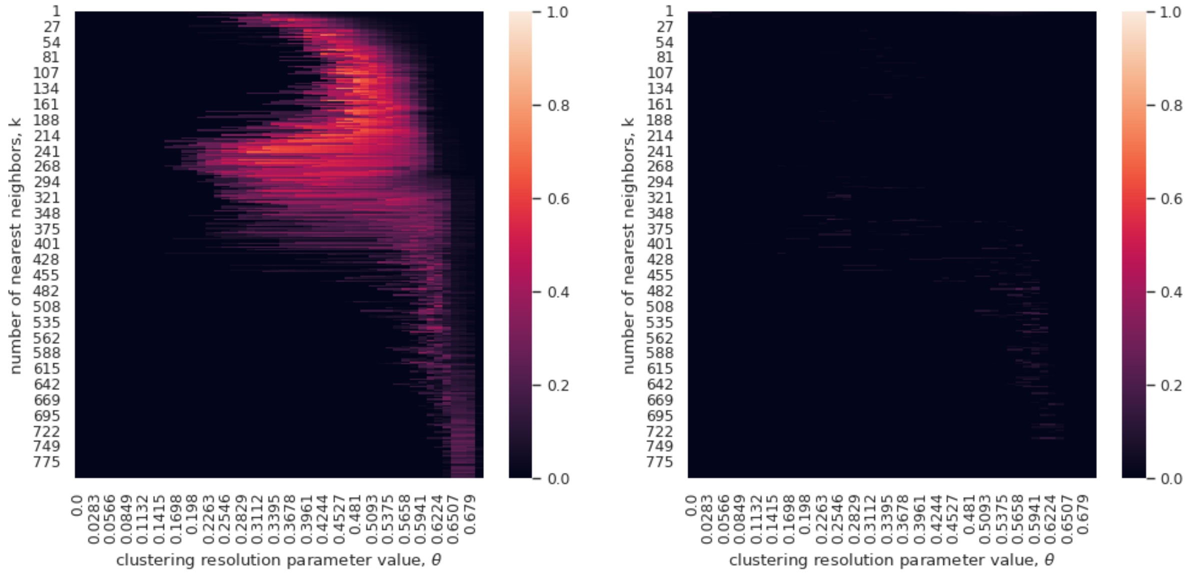


(a) ARI vals (b) SC vals

Figure 4: Average values of clustering results metrics for the network (1) of the WingNut dataset under the condition of using 5 random initial conditions for iterative equation (3): (a) Adjusted Rand Index (ARI); (b) Silhouette Coefficient (SC).

The network with dipole connections (2) is less sensitive to the number of nearest neighbours; however, the largest window size is smaller than that for the network with synaptic connections (1). This means that the network with dipole connections can only detect coarser differences between this kind of data.

The next experimental dataset will be WingNut, which is complex due to the small inter-cluster distance compared to the large intra-cluster distance [12].

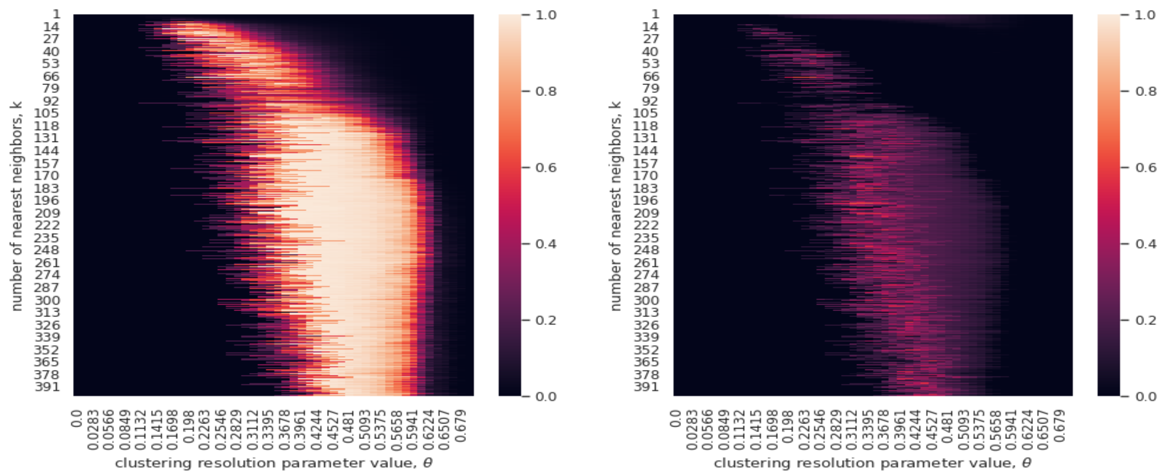


(a) ARI vals (b) SC vals

Figure 5: Average values of clustering results metrics for the network (2) of the WingNut dataset under the condition of using 5 random initial conditions for iterative equation (3): (a) Adjusted Rand Index (ARI); (b) Silhouette Coefficient (SC).

Due to the nature of the WingNut data, the silhouette coefficients in Fig. 4 (b) and Fig. 5 (b) are small, meaning they are negative or close to 0 in most of the clustering results.

On these data, the network with dipole synaptic connections (2) forms clusters with a higher ARI value. This is because network (2) assigns more weight to the nearest points compared to network (1).



(a) ARI vals (b) SC vals

Figure 6: Average values of clustering results metrics for the network (1) of the TwoDiamonds dataset under the condition of using 5 random initial conditions for iterative equation (3): (a) Adjusted Rand Index (ARI); (b) Silhouette Coefficient (SC).

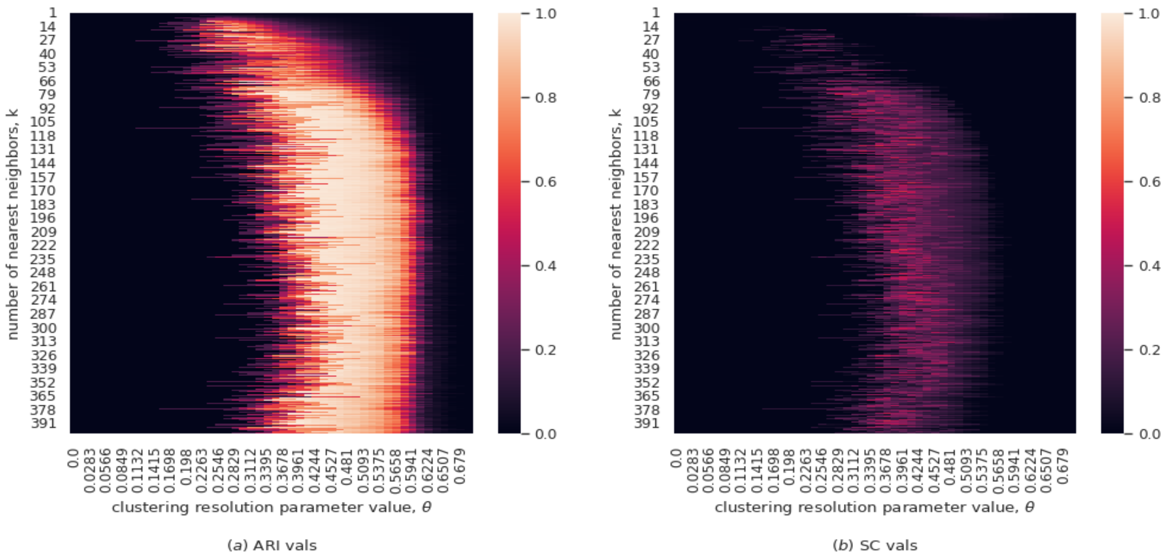


Figure 7: Average values of clustering results metrics for the network (2) of the TwoDiamonds dataset under the condition of using 5 random initial conditions for iterative equation (3): (a) Adjusted Rand Index (ARI); (b) Silhouette Coefficient (SC).

Although network (1) has a larger clustering resolution window for the TwoDiamonds dataset, network (2) requires fewer nearest neighbours, meaning it can identify clusters with high accuracy even with a small number of neighbours.

The next dataset, EngyTime, can be correctly clustered based solely on density since the data from different classes intersect.

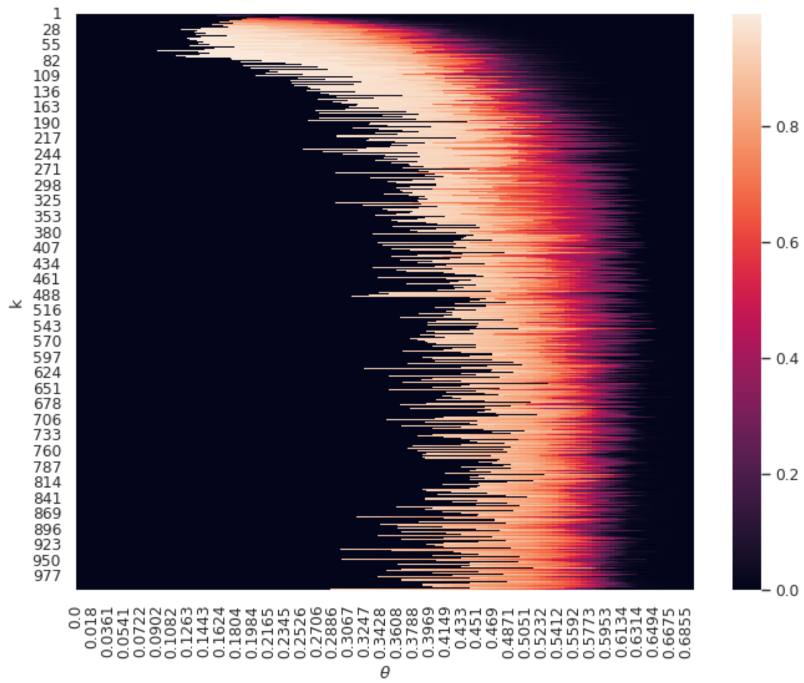


Figure 8: Average values of the ARI metric for the clustering results of the network (1) on the EngyTime dataset under the condition of using 5 random initial conditions for the iterative equation (3).

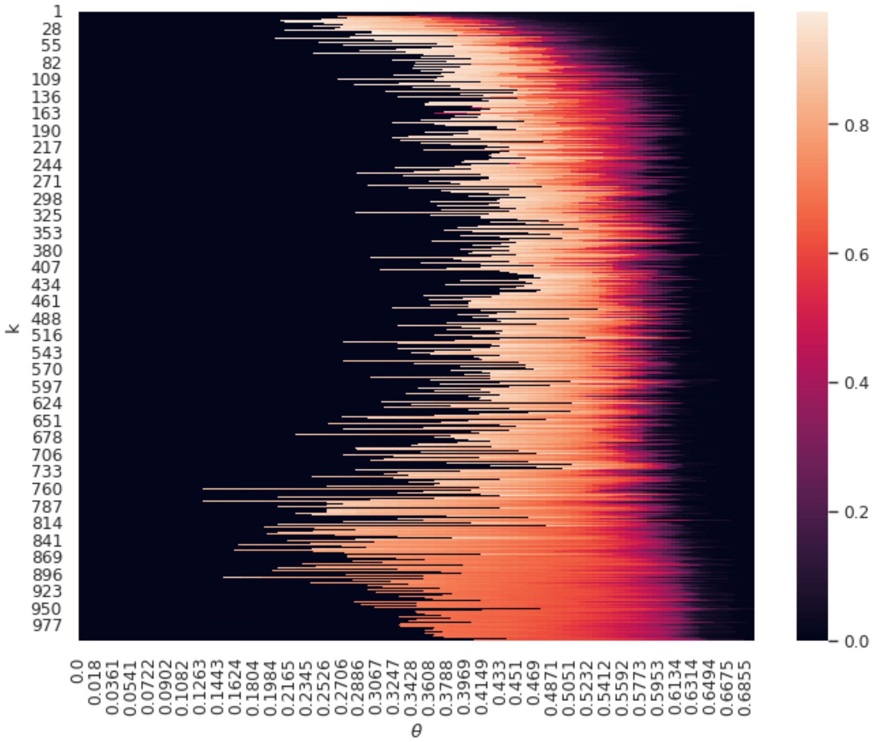


Figure 9: Average values of the ARI metric for the clustering results of network (2) on the EngyTime dataset under the condition of using 5 random initial conditions for the iterative equation (3).

For the EngyTime dataset, at low values of nearest neighbours, the network (1) performs better in clustering than network (2). This can be attributed to the complexity of the EngyTime dataset, which requires considering density. Therefore, network (1) is a more effective clustering method for datasets characterized by linear inseparability or complex topology.

5. Discussion

The research results have shown that the number of neighbours can influence the clustering effectiveness. For some tasks, it was proven that the clustering efficiency increases with the increase in the number of neighbours. This is because increasing the number of neighbours allows neurons to form denser connections between them, which can lead to more clearly defined clusters.

The network with dipole connections (2) is more flexible and less sensitive to the number of neighbours compared to the network with synaptic connections (1). This makes it more effective for a wider range of datasets that are complex due to small inter-cluster distances compared to large intra-cluster distances. For example, the network with dipole connections can be an effective clustering method for datasets having data points with varying densities [13].

Furthermore, network (2) is less sensitive to the number of neighbours. This means that it can detect clusters with high accuracy even with a small number of neighbours.

The network with synaptic connections (1) has a larger clustering window compared to the network with dipole connections (2). This allows it to detect finer differences between data points. This makes it more effective for datasets that are complex due to linear inseparability or topology. For example, the network with synaptic connections can be an effective clustering method for datasets having intersecting data points.

Conclusions

- It has been established that oscillatory chaotic neural networks with dipole synaptic connections between neurons are novel networks that can solve clustering tasks for a wider range of datasets, regardless of their complexity, compared to networks with Gaussian synaptic connections between neurons.
- It has been demonstrated that a network with dipole connections (2) proves to be more flexible and less sensitive to the number of nearest neighbours compared to the network with synaptic connections (1). This characteristic makes it particularly effective for datasets where it is important to consider complexity due to the small inter-cluster distances compared to the large intra-cluster distances.
- It has been determined that networks with synaptic connections (1) have a larger clustering window and higher resolution compared to networks with dipole connections. This allows them to detect even small differences between data points. They are effective for datasets where complexity is due to linear inseparability or special structure.
- It has been identified that the optimal number of neighbours for each network and dataset is crucial to achieving the maximum resolution window and the maximum value of the clustering quality metric.

References

[1] G. J. Oyewole, G. A. Thopil, Data clustering: application and trends, Artif. Intell. Rev. (2022). doi:10.1007/s10462-022-10325-y.

[2] V. Baruzzi, M. Lodi, F. Sorrentino, M. Storace, Bridging functional and anatomical neural connectivity through cluster synchronization, Sci. Rep. 13.1 (2023). doi:10.1038/s41598-023-49746-2.

[3] K. Aihara, T. Takabe, M. Toyoda, Chaotic neural networks, Phys. Lett. A 144.6-7 (1990) 333–340. doi:10.1016/0375-9601(90)90136-c.

[4] L. Angelini, F. De Carlo, C. Marangi, M. Pellicoro, S. Stramaglia, Clustering Data by Inhomogeneous Chaotic Map Lattices, Phys. Rev. Lett. 85.3 (2000) 554–557. doi:10.1103/physrevlett.85.554.

[5] L. Angelini, Chaotic neural network clustering: an application to landmine detection by dynamic infrared imaging, Opt. Eng. 40.12 (2001) 2878. doi:10.1117/1.1412623.

[6] R. Penrose, Shadows of the mind, Random House of Canada, Limited, 1995.

[7] S. R. Hameroff, Quantum coherence in microtubules: a neural basis for emergent consciousness?, J. Conscious. Stud. 1.1 (1994) 91–118.

[8] J. A. Brown, J. A. Tuszynski, A review of the ferroelectric model of microtubules, Ferroelectrics 220.1 (1999) 141–155. doi:10.1080/00150199908216213.

[9] J. A. Tuszyński, S. Hameroff, M. V. Satarić, B. Trpisová, M. L. A. Nip, Ferroelectric behavior in microtubule dipole lattices: Implications for information processing, signaling and assembly/disassembly, J. Theor. Biol. 174.4 (1995) 371–380. doi:10.1006/jtbi.1995.0105.

[10] C. Hunt, H. Stebbings, Role of MAPs and motors in the bundling and shimmering of native microtubules from insect ovarioles, Cell Motil. Cytoskelet. 27.1 (1994) 69–78. doi:10.1002/cm.970270108.

[11] R. O. Sinnott, H. Duan, Y. Sun, A Case Study in Big Data Analytics, in: Big Data, Elsevier, 2016, pp. 357–388. doi:10.1016/b978-0-12-805394-2.00015-5.

[12] M. C. Thrun, A. Ultsch, Clustering benchmark datasets exploiting the fundamental clustering problems, Data Br. 30 (2020) 105501. doi:10.1016/j.dib.2020.105501.

[13] Stochastic Pseudo-Spin Neural Network with Tridiagonal Synaptic Connections, 2021 IEEE Int. Conf. Smart Inf. Syst. Technol. (SIST) (2021) 1–6. doi:10.1109/SIST50301.2021.9465998.