Data Clustering using Nature inspired Kin Recognition System

Asha Yadav[1] and Seema[2*]

ABSTRACT

Data clustering is a widely employed method in the area of data analysis, finding applications in several domains such as data mining, pattern recognition and picture analysis.

Researchers have made it a continual goal to enhance the performance of clustering algorithms and find solutions to the difficulties that come with the management of large datasets during clustering. On the other hand, traditional clustering algorithms might not be up to the task of reaching a greater level of accuracy when it comes to the classification of enormous datasets. Consequently, the application of clever algorithms is essential in order to successfully cluster difficult data. In this paper, a method of clustering is presented that takes its cues from natural phenomena, namely the phenomenon of kin recognition among groups of trees. The experimental results showed the accuracy and capability of proposed algorithm to data clustering.

Keywords: *Clustering, Kin recognition, Phylogenetic measure.*

1. Introduction

Clustering is a widely acknowledged methodology employed for the purpose of knowledge discovery in diverse scientific domains. Notably, it finds application in clustering gene expression data (Cui et al., 2013), medical image analysis (An & Doerge, 2012), air pollution data analysis (Moolgavkar et al., 2013), and power consumption data analysis (Iglesias & Kastner, 2013). The primary objective of the clustering process is to group comparable data points together inside a cluster, while ensuring that dissimilar data points are assigned to separate clusters. Currently, one of the primary issues in the field of database management pertains to the increasing diversity of data types, which has resulted in heightened complexity and scope. In the context of data complexity, the measurement of distance also emerges as a significant consideration in the field of clustering. Several clustering methods have been utilised to create clusters in datasets. These algorithms include k-means, k-medoids, hierarchical clustering, as well as more modern approaches such as fuzzy c-means and rough clustering. Researchers have consistently aimed to improve the performance of clustering algorithms and address the challenges associated with managing huge datasets in clustering. However, conventional clustering algorithms may not be adequate for achieving a higher level of accuracy when grouping huge datasets. Therefore, the utilisation of intelligent algorithms is necessary in order to cluster complex data. This study presents a clustering technique that draws inspiration from nature, namely the concept of kin recognition among groups of trees, in order to cluster datasets.

Plants possess the ability to identify and acknowledge their genetic relations, and as a result, they exhibit alterations in their functional characteristics in response to the presence of nearby kin. In a forest ecosystem, several species of trees and fungi establish symbiotic relationships, forming a cooperative system through which they trade vital resources. The cooperative organisational framework of plants facilitates the sharing, communication and exchange of diverse vital nutrients, hence enhancing their survival(Books, 2017). Simard's research revealed the utilisation of a cooperative mechanism by Douglas fir trees, enabling them to engage in communication, protection, and defence of their kin and neighbouring trees. This mechanism involves the sharing of several resources, including phosphorus, nitrogen, zinc and water (Beiler et al., 2010).

Although trees may appear to be solitary entities, the soil beneath our feet reveals a contrasting narrative. It has been suggested that trees engage in covert communication, exchange resources and engage in conflict with one another. This is accomplished by the utilisation of a network of fungi that proliferate in the vicinity of and inside their root systems. Fungi play a vital role in facilitating nutrient exchange with plants, whereby they

1. Department of Computer Science, School of Open Learning, University of Delhi
2. Department of Computer Science, SRCASW, University of Delhi
* *Corresponding Author* ✉ seema.seema@rajguru.du.ac.in

Received: 10 October, 2023
Available online: 31 December, 2023

offer essential nutrients to trees while receiving sugars in reciprocation. Through the process of connecting to the mycelial network, trees are able to exchange resources with one another. The individuals are engaging in conversation, exchanging both knowledge and culinary items. Forests should not be regarded just as an assemblage of individual trees. Forests are intricate systems comprised of interconnected hubs and networks that provide communication among trees. These structures enable feedback mechanisms and adaptive processes, hence enhancing the resilience of the forest ecosystem. Mycorrhizal fungi establish a symbiotic and mutually beneficial reciprocal association with plants. Mycorrhizae assist the growth of individual plants. Fungal networks have been observed to enhance the immune systems of their host plants. This phenomenon occurs due to the colonisation of plant roots by fungi, which then induces the synthesis of defence-associated compounds. Fungi have been referred to as the natural networking system of the Earth. Mycelium serves as a means of interconnection among various individuals within forest ecosystems, facilitating not only intra-species interactions but also inter-species interactions.

The trees contain brain-like structures and processes which helps them to communicate through the social network at the root tips. In the context of arboreal ecosystems, it has been shown that a mother tree, characterised by its bigger size and advanced age, possesses the ability to discern its offspring through the intricate network of mycorrhizal fungi. Furthermore, it is worth noting that a juvenile tree possesses the ability to recognise the maternal tree through the intricate mycorrhizal network. The mycelium of the root functions similarly to the human nervous system in terms of transmitting information. Mother trees have the ability to distinguish between seedlings that are related to them and those that are unrelated. Mother trees establish connections with their offspring through larger mycorrhizal networks. They sequester a greater amount of carbon in subterranean environments. They actively decrease competition among their own roots in order to create space for their relatives. When a mother tree sustains injury or approaches the end of her life cycle, it conveys messages of wisdom to the succeeding generation of seedlings.

The phenomenon of kin recognition and the understanding of kin-specific chemical signals have been widely acknowledged in several microbial and animal models. Recent studies have demonstrated that plants possess the capacity to recognise neighbouring plants by discerning their relatedness and identity. Despite lacking mobility, plants are not considered passive entities (Biedrzycki & Bais, 2010). Plant-plant interactions have been extensively studied and documented, encompassing a range of interactions that can be either negative or positive. Negative interactions such as allelopathy, involve the production of chemicals by one plant that inhibit the growth or development of neighbouring plants. On the other hand, positive interactions involve the release of volatile compounds by a plant to alert nearby plants of potential threats or dangers. This exchange of information among plants is a key. The annual plant Cakile edentula demonstrated kin recognition by exhibiting increased root production when grown in the presence of non-related individuals (plants of the same species but originating from different maternal sources) compared to being grown alongside genetically related individuals (plants originating from the same maternal source) (Dudley & File, 2007).

Kin recognition refers to the capacity to distinguish individuals who are genetically related from those who are not, irrespective of the underlying process or evolutionary purpose (Penn & Frommen, 2010). Phenotype matching sometimes referred to as the 'armpit effect', represents a prevalent method for kin recognition. Phenotype matching is a process wherein an individual assesses phenotypic clues shown by another individual within the population in order to ascertain their kinship. The effects of kin recognition extend beyond the mere reduction of resource competition among genetically related plants,

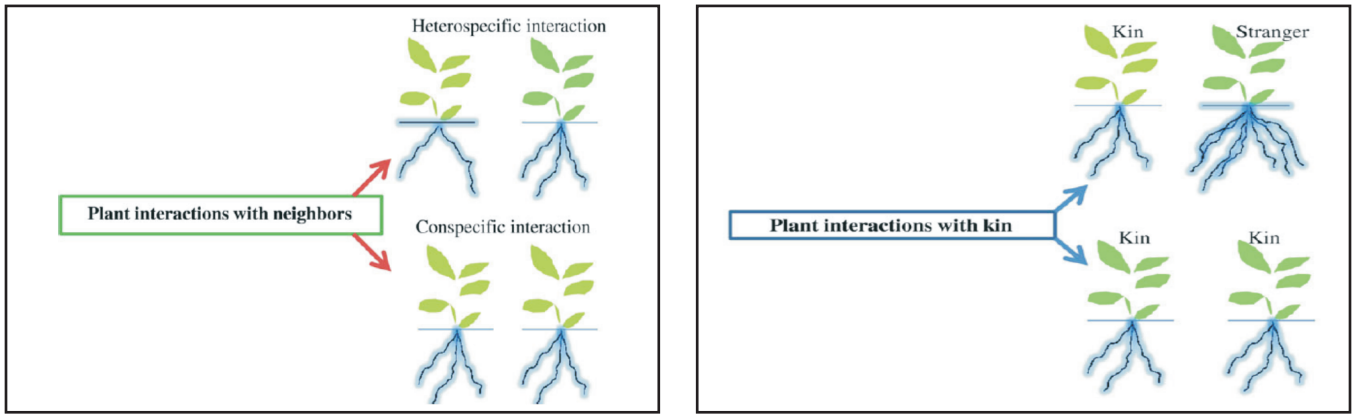

Figure1: *Plant interaction with kin and neighbours*

encompassing interactions with symbiotic organisms such as mycorrhizal networks.

Plants that are cultivated alongside both family members and unrelated individuals have been seen to exhibit adaptability in terms of their root and aerial growth. Fig 1 shows the effect of plant interaction with kin and neighbour. In contrast, plants cultivated alongside conspecific and heterospecific individuals exhibit root flexibility in response to fluctuating nutritional circumstances. Plants cultivated alongside conspecifics in a shared habitat have been observed to elicit defence-related chemicals to a greater extent when compared to interactions with heterospecific. The user's text does not provide any information or context to be rewritten in an academic manner.

2. Literature Review

The present analysis focuses on conducting a comprehensive literature study. The utilisation of clustering techniques allows for the extraction of concealed insights pertaining to intricate networks, which may be challenging to discern from rudimentary observations. In the context of a co-authorship network, clustering can be interpreted as a manifestation of researchers collaborating within a specific field of study (Daud et al., 2017). Similarly, within a communication network, clustering can signify the presence of tightly-knit social circles consisting of friends, family members, or colleagues. In the context of a software source code network, clustering may indicate the existence of distinct subsystems. Lastly, in a citation network, clustering can denote the grouping. Extensive study has been conducted across various fields, such as computer science, statistics, biology, physics, business intelligence, pattern recognition, online search and astronomy in order to identify and explore clusters. This is attributed to the significant significance of clustering and its wide range of applications. The partitioning methods, such as k-means and K-medoids involve the relocation of points by transferring them from one category to another based on their distance. Number of clusters needs to be set in advance in these approaches and they are sensitive to starting cluster. Hierarchical clustering methods employ recursive principles to partition the data through either a top-down or bottom-up approach. Density-based methods such as DBSCAN and Denclue, have the capability to identify clusters of arbitrary shape. Soft computing encompasses many evolutionary methodologies that are commonly employed for addressing clustering difficulties. Various algorithms, including the Genetic Algorithm (GA), Artificial Bee Colony (ABC), and Particle Swarm Optimisation (PSO) have demonstrated the ability to effectively optimise the objective function (Jiang & Wang, 2014). The purpose of this introductory section is to provide an overview of the topic at hand. The incorporation of K-nearest neighbour (KNN), principal component analysis (PCA) and fuzzy weighted K-nearest neighbour into the density peaks method (DPC) addresses the constraints of the Density Peak Clustering algorithm (Xie et al., 2016). A novel clustering technique, referred to as PDPC (Cai et al., 2020), was proposed to address the challenges associated with local optima and random selection of initial centroids. This approach builds upon the notions of DPC and particle swarm optimisation.

The majority of current clustering methods utilise pairwise distance measurements between items as input without incorporating the benefits of phylogenetic principles. For instance, the renowned UCLUST algorithm is designed to generate clusters that aim to minimise the hamming distance between a sequence and its corresponding cluster centroid, while simultaneously maximising the hamming distance between centroids (Edgar, 2010). Utilising phylogenetic relatedness as a basis for grouping offers two potential advantages. Phylogenetic analysis aims to infer the evolutionary history of organisms. Therefore, clustering based on phylogeny has the potential to not only reflect the evolutionary distance (branch length), but also the relationship in terms of tree topology and kin recognition. The ClusterPicker algorithm (Ragonnet-Cronin et al., 2013) constructs a cluster sequence by considering the distances between clusters, while using a phylogenetic tree as a constraint. However, it should be noted that ClusterPicker primarily relies on sequence distance rather than tree distance, and it also incorporates scaling factors (Ragonnet-Cronin et al., 2013).

In Nye's publication titled "Tree of Trees" (2008), the author gives a summary of the evolutionary similarities seen among genes. This summary is shown as a meta-tree, wherein each tip corresponds to a tree produced from multi-locus data, and each internal node represents the consensus of its respective child trees. The meta-tree is derived using intertree Robinson-Foulds distances (Gori et al., 2016) using an approach similar to the neighbour joining method employed in phylogenetic tree reconstruction (Saitou & Nei, 1987).While the neighbour joining approach is known for its computational efficiency, it is important to note that the compression of sequences into distances might result in information loss. Additionally obtaining accurate estimates of pairwise distances for divergent sequences can pose challenges.

In a similar vein, the Conclustador algorithm (Leigh et al., 2011) employs intertree distances as a fundamental component for the purpose of clustering. In the study conducted by Leigh et al., a unique Euclidean distance metric is employed to compare trees, taking into account bipartitions weighted by bootstrap support. Additionally, for the purpose of clustering, the authors utilise a modified

version of the k-means algorithm and a spectral clustering approach (Agarwal et al., 2005). PhyBin (Newton & Newton, 2013) is a method that shares conceptual similarities, since it is capable of either detecting genes with trees that are topologically identical or conducting hierarchical clustering based on the Robinson-Foulds distance matrix of each tree.

Phylogenetic inference has the ability to yield a cluster with higher accuracy. The utilisation of Bayesian inference has been employed in the estimation of phylogenetic relatedness among different species. The utilisation of Bayesian phylogenetic approaches has significantly transformed the landscape of analysis. These analyses encompass several methodologies, such as phylogeographic analysis to examine the spread of viruses in human populations, the inference of phylogeographic history and migration patterns between different species, the investigation of rates of species diversification, the estimation of divergence times and the inference of phylogenetic relationships across species or populations (Nascimento et al., 2017). The rise in popularity of Bayesian approaches can be attributed to two primary factors: firstly, the advancement of robust data analysis models; and secondly the accessibility of user-friendly computer programmes that effectively execute these models (Nascimento et al., 2017).

The Bayesian approach is a methodology for statistical inference. The primary characteristic of this approach involves the utilisation of probability distributions to depict the level of uncertainty associated with all variables that are not known, including the parameters of the model. The statistical model for the data in phylogenetics is determined by the combination of the tree topology and the substitution model. Various tree topologies correspond to distinct models, with the branch lengths or divergence durations, as well as the substitution parameters serving as parameters inside the model.

3. Research Methodology

Kin recognition can be applied to clustering because based on similarity measures clustering algorithm can differentiate similar and dissimilar objects in the groups. Similarly kin recognition concept can be used to identify stranger and neighbour tree based on the phylogenetic signal relatedness and Phenotypic similarity traits. Bayesian inference has been used to estimate phylogenetic relatedness among species using three nuclear and plastid genes and phenotypic similarity between species has been quantified using the Euclidean distance across traits (Fitzpatrick et al., 2017).

Equation 1 calculate the Euclidean distance

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad \ldots (1)$$

Where

d	:	represents dimension of dataset
p, q	:	objects of dataset
i	:	current object of dataset
n	:	number of iterations

Input: Data set with dimension D and Number of clusters K

Output: Clusters

Step 1: Initialize the data set D, Number of clusters K.

- Calculate the maximum M and minimum m values of the data set D(pseudocode), then assign the M or m values as initial centroid,

Step 2: Calculate Phylogenetic relatedness and phenotypic similarity measure

- Bayesian inference to estimate phylogenetic relatedness
- Phenotypic similarity between species using the Euclidean distance across all measured traits

Step 3: Update the centroid using K-neighbour

Step 4: Recalculate Phylogenetic relatedness and phenotypic similarity measure

Step 5: Repeat steps 2 ,3, 4 until there are no changes in centroid values.

Figure 2 : *Pseudocode*

Experiments have been performed on four data sets including Iris, WDBC, Sonar and Wine that were selected from standard data set UCI. The Fig.2 represents pseudocode used for phylogenetic measure for clustering.

4. Results Analysis and Discussion

The methods and reasons behind kin recognition in animal and bacterial species exhibit significant diversity and may or may not be observed within a given species. Moreover, it is possible that this kin recognition does not result in any kind of kin bias or discrimination, or it may only result in kin bias when certain advantageous circumstances are present. The results of the experiments demonstrated that the suggested algorithm is accurate and capable of clustering data when applied it.

One of the primary attributes of a clustering method is its capacity to reduce clustering error. A suitable clustering methodology should possess the capability to minimise clustering error and accurately allocate data vectors to their respective clusters, in addition to considering intra-cluster error. Table 1 presents the clustering error results of K-means, K-medoids and Phylogenetic measure for the Iris, WDBC, Sonar and Wine datasets.

Table 1: *Comparison of Clustering Error between Different Methods for all dataset*

Data Set	K-means	K-medoids	Phylogenetic measure
Iris	15.05±10.10	10.64±4.50	9.58±7.67
WDBC	18.12±9.22	13.18±1.80e-15	11.18±1.81e-15
Sonar	54.95±0.97	46.60±0.42	44.98±0.84
Wine	32.38±6.08	28.74±0.39	25.59±0.47

5. Conclusion and Future Direction

This article presents a method of clustering that derives its cues from natural events, namely the phenomenon of kin recognition among groups of trees. It was demonstrated by the results of the experiments that the proposed algorithm is both accurate and capable of clustering data. Investigation into the domain of plant kin recognition will contribute to a more comprehensive knowledge of interplant communication and the broader scope of plant development. The discovery of additional plant species capable of kin recognition, particularly within the context of crop species, tree variants, or perennial species, would be highly intriguing. The examination of the mechanisms and signalling components implicated in kin recognition, encompassing root secretions and volatile emissions, would be of utmost importance and could potentially yield insights into plant relationships at a broader multitrophic level.

References

1. Agarwal, S., Lim, J., Zelnik-Manor, L., Perona, P., Kriegman, D., & Belongie, S. (2005). Beyond pairwise clustering. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2*, 838–845.
2. An, L., & Doerge, R. W. (2012). Dynamic clustering of gene expression. *International Scholarly Research Notices, 2012*.
3. Beiler, K. J., Durall, D. M., Simard, S. W., Maxwell, S. A., & Kretzer, A. M. (2010). Architecture of the wood - wide web: *Rhizopogon* spp. genets link multiple Douglas - fir cohorts. *New Phytologist*, *185*(2), 543–553. https://doi.org/10.1111/j.1469-8137.2009.03069.x
4. Biedrzycki, M. L., & Bais, H. P. (2010). Kin recognition in plants: A mysterious behaviour unsolved. *Journal of Experimental Botany*, *61*(15), 4123–4128.
5. Books, W. (2017). *Summary and Analysis of The Hidden Life of Trees: What They Feel, How They Communicate — Discoveries from a Secret World: Based on the Book by Peter Wohlleben*.
6. Cai, J., Wei, H., Yang, H., & Zhao, X. (2020). A novel clustering algorithm based on DPC and PSO. *IEEE Access*, *8*, 88200–88214.
7. Cui, W., Wang, Y., Fan, Y., Feng, Y., & Lei, T. (2013). Localized FCM clustering with spatial information for medical image segmentation and bias field estimation. *Journal of Biomedical Imaging*, *2013*, 13–13.
8. Daud, A., Aljohani, N. R., Abbasi, R. A., Rafique, Z., Amjad, T., Dawood, H., & Alyoubi, K. H. (2017). Finding Rising Stars in Co-Author Networks via Weighted Mutual Influence. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 33–41. https://doi.org/10.1145/3041021.3054137
9. Dudley, S. A., & File, A. L. (2007). Kin recognition in an annual plant. *Biology Letters*, *3*(4), 435–438. https://doi.org/10.1098/rsbl.2007.0232
10. Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461.
11. Fitzpatrick, C. R., Gehant, L., Kotanen, P. M., & Johnson, M. T. J. (2017). Phylogenetic relatedness, phenotypic similarity and plant–soil feedbacks. *Journal of Ecology*, *105*(3), 786–800. https://doi.org/10.1111/1365-2745.12709
12. Gori, K., Suchan, T., Alvarez, N., Goldman, N., & Dessimoz, C. (2016). Clustering genes of common evolutionary history. *Molecular Biology and Evolution*, *33*(6), 1590–1605.
13. Iglesias, F., & Kastner, W. (2013). Analysis of similarity measures in times series clustering for the discovery of building energy patterns. *Energies*, *6*(2), 579–597.

14. Jiang, B., & Wang, N. (2014). Cooperative bare-bone particle swarm optimization for data clustering. *Soft Computing*, *18*, 1079–1091.
15. Leigh, J. W., Schliep, K., Lopez, P., & Bapteste, E. (2011). Let them fall where they may: Congruence analysis in massive phylogenetically messy data sets. *Molecular Biology and Evolution*, *28*(10), 2773–2785.
16. Moolgavkar, S. H., McClellan, R. O., Dewanji, A., Turim, J., Luebeck, E. G., & Edwards, M. (2013). Time-Series Analyses of Air Pollution and Mortality in the United States: A Subsampling Approach. *Environmental Health Perspectives*, *121*(1), 73–78. https://doi.org/10.1289/ehp.1104507
17. Nascimento, F. F., Reis, M. dos, & Yang, Z. (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution*, *1*(10), 1446–1454.
18. Newton, R. R., & Newton, I. L. (2013). PhyBin: Binning trees by topology. *PeerJ*, *1*, e187.
19. Penn, D. J., & Frommen, J. G. (2010). Kin recognition: An overview of conceptual issues, mechanisms and evolutionary theory. In P. Kappeler (Ed.), *Animal Behaviour: Evolution and Mechanisms* (pp. 55–85). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-02624-9_3
20. Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Brown, A. J. L., & Lycett, S. (2013). Automated analysis of phylogenetic clusters. *BMC Bioinformatics*, *14*(1), 317. https://doi.org/10.1186/1471-2105-14-317
21. Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425.
22. Xie, J., Gao, H., Xie, W., Liu, X., & Grant, P. W. (2016). Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K-nearest neighbors. *Information Sciences*, *354*, 19–40.

❖ ❖ ❖