A Re-examination of the Causal Relationship between Prototype Generation and Category Representation from the Perspective of Machine Intelligence

Andrew H. Zhang

University of Toronto St. George

COG250Y1: Introduction to Cognitive Science

Dr. J. Vervaeke

Last Revised: December 06, 2021

Abstract

The prototype theory of concepts is a theory of categorization which proposes that in human cognition, prototype generation causally facilitates category representation. Key criticisms of prototype theory, namely its inability to explain the 'pet-fish problem' and the instability of conceptual prototypes under varying contexts, are examined from the perspective machine intelligence where both supervised and unsupervised classification algorithms are used as tools to provide insight into whether prototype generation or category representation is likely to be the more fundamental process in human cognition.

Under the assumption that human cognition is mechanizable, recent advances in machine intelligence strongly imply that a theory of categorization in which category representation enables prototype generation has greater explanatory power over the reverse.

The nearest centroid classifier supervised classification algorithm as a mechanizable abstraction of prototype theory where prototype generation causes category representation demonstrates the key merits of prototype theory while suffering from the same inability to explain the 'pet-fish problem' and the instability of conceptual prototypes under varying contexts. With prototype generation being caused by category representation, Gaussian mixture discriminant analysis as a supervised classification algorithm and dynamic k-means clustering as an unsupervised classification algorithm are both able to retain the key merits of prototype theory while each respectively being able to demonstrate the 'pet-fish problem' and the instability of conceptual prototypes under varying contexts.

Introduction

As put forward by the works of Elenor Rosch in the early 1970s, the prototype theory of categorization proposes that the cognitive process of category representation arises from one's ability to generate a prototype for each conceptual category that one perceives. In her theory, Rosch defines a 'prototype' as an abstract "central tendency" of the distribution of concrete instances of a category across the space of features that they possess, in such a way that "(prototypes) become foci of organization for categories" *(Rosch, 1973)*.

Despite its success in accommodating the breakdown of the deductive taxonomic structure of classic Aristotelian category in human cognition, prototype theory faces two major phenomena in human cognition that it fails to explain, namely the 'pet-fish problem' and the instability of conceptual prototype under varying context.

The pet-fish problem as demonstrated by Osherton and Smith in 1981 shows that for two distinct conceptual categories (pet and fish), features of prototypical exemplars of each individual category (prototypical pets are fluffy; prototypical fish lives in the ocean) may both be absent in a prototypical exemplar of the composite concept category formed by both (goldfish is a prototypical pet-fish).

The instability of conceptual prototypes under varying context as demonstrated by the experiments published by Roth and Shoben in 1983 showed that under differing contexts, the

prototypical exemplars of the same conceptual category changes, and so do the typicality gradient of all other exemplars of the category.

The ways in which the aforementioned two phenomena challenge prototype theory is conducive to a re-examination of the causality between prototype generation and category representation as described by prototype theory, namely whether our ability to generate prototypes is the effect of our ability to categorize, instead of its result as implied by the theory.

Under the assumption that human cognition is mechanizable, examining various classification machine learning algorithms that seek to reproduce human categorization can offer valuable insight into which process out of category representation and prototype generation is more fundamental.

Justification of Assumptions and Logical Structure of Arguments

The primary assumption underlying this paper is that human cognition, or at least human categorization, is theoretically mechanizable. Underpinned by the materialist theory of mind, this assumption is fundamental to connecting the algorithms of machine categorization to insights about prototype theory as a theory of human categorization, as it facilitates the establishment of mechanical and probabilistic models of cognition as a set of tools to analyze the assumptions that underlie human cognitive processes and how these processes can take place at multiple levels of abstraction *(Griffiths, 2009)*. Objections to this assumption would arguably take away the only objective tools that we have to model the processes of our mind, as the introspective structuralist school of psychology has been historically proven to be notoriously unreliable.

In viewing machine learning algorithms as tools to analyze the causal relationship between prototype generation and categorization representation, we will first examine an algorithm where classification is completely causally dependent on prototype generation, and investigate whether it displays the same merits, namely allowing varying degrees of membership to each conceptual category, and pitfalls, namely the inability to explain the pet-fish problem and instability of conceptual prototype under varying context, as prototype theory. Then, we will examine more algorithms where prototype generation is in reverse enabled by categorization, and investigate whether they could retain the same merits of prototype theory while accommodating some of the theory's major pitfalls. This should allow us to deduce which approach is likely a better model of human categorization, and thus provide insight into whether category representation or prototype generation should be more fundamental in a theory of human cognition.

Nearest Centroid Classifier as a Mechanizable Abstraction of Prototype Theory

The nearest centroid classifier is a supervised classification algorithm, where a set of data points each belonging to a specific class labelled by humans with the correct classification are provided to the algorithm. This is called the 'training set'. The algorithm is 'trained' by computing and recording the Euclidean barycenters ('centroids') of all data points labelled as belonging to each class in the training set. New unseen and unlabelled data points are then classified into the class whose centroid that it is closest to in terms of Euclidean distance in feature space *(Johri et al., 2021)*. This process can be visualized on *Figure 1*.

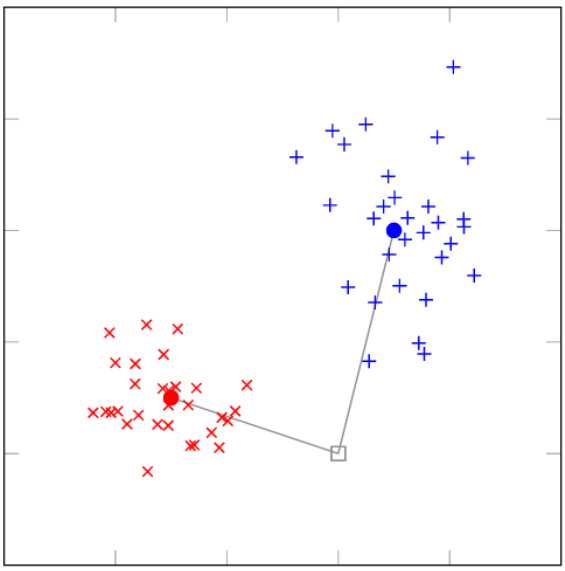

Figure 1. Nearest Centroid Classifier in 2D Feature Space
(Adcock et al., 2015)

On this figure: crosses are the training set, circles are centroids, the square is an unseen unlabelled new data point. In this case it will be classified into the 'red' class since it is closer to the red centroid.

Empirically, nearest centroid classifier has applications in tasks such as mosquito genus classification, where approaches incorporating this algorithm can yield 97.2% average classification accuracy *(Alar & Fernandez, 2021)*.

As a "central tendency", centroids in this algorithm can be considered as the functional equivalent to the prototypes as described in prototype theory. With such definition of prototype, nearest centroid classifier is a rather faithful mechanizable abstraction of prototype theory, where the algorithm's ability to represent different categories in feature space is completely causally dependent on the prototypes that it generates. As such, the merits of prototype theory are easily seen in this algorithm, where the varying degrees of membership of an exemplar (new data point) to a class or category can be quantified by its Euclidean distance to the centroid in feature space, thus accommodating for the breakdown of deductive taxonomic structure imposed by Aristotle's classical theory of categorization.

However, it is also quite apparent that this algorithm suffers from the same major pitfalls as prototype theory. As the algorithm performs classification, the location of centroids remains rigid, thus making it difficult to explain the instability of prototypes under varying contexts in human cognition. This rigidity also makes using this algorithm to represent compound categories difficult, which would imply that the pet-fish problem also could not be readily demonstrated by this algorithm.

Gaussian Mixture Discriminant Analysis and Demonstrating the Pet-fish Problem

Gaussian mixture discriminant analysis is also a supervised classification algorithm, but as opposed to nearest centroid classifier, it is a generative model that generates category representation before prototypes are produced. The gaussian mixture discriminant analysis algorithm is trained on the training set by generating a weighted gaussian distribution of numeric likelihood that any given data point in feature space belongs to a specific class, for every class in the training set. These distributions of likelihoods being functions of the entire feature space enables the classification of a new unseen unlabeled data point to the class whose gaussian distribution function returns the highest numeric likelihood for it. The decision boundaries, or 'discriminants', are thus produced by the intersections of the Gaussians for different classes. This process can be visualized on *Figure 2*.

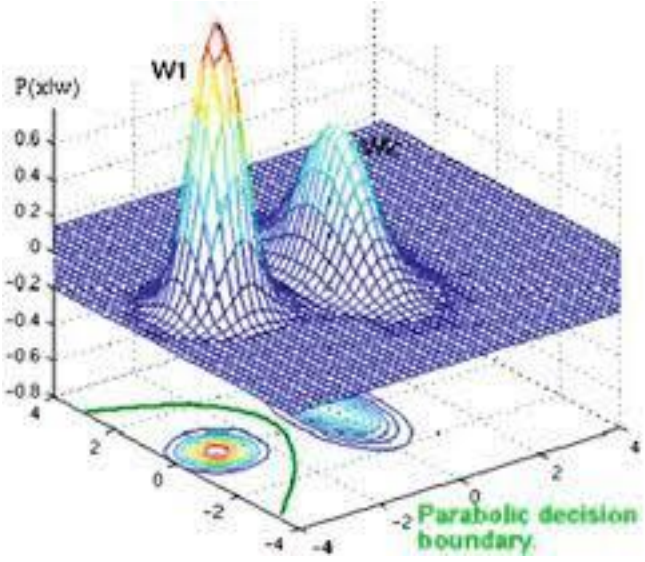

Figure 2. Gaussian Discriminant Analysis in 2D Feature Space (Dougherty, 2012)

Data points to the left of the green decision boundary will be classified as class "W1", whose Gaussian distribution yields a greater likelihood. Vice-versa for Class "W2".

Empirically, when applied to problems such as diagnosis of different stages of Alzheimer's disease, it is able to achieve an average classification accuracy of 87.43% *(Fang et al., 2017)*. Addressing the possible argument that this empirical accuracy being lower than nearest centroid classifier's 97.2% empirical accuracy in classifying mosquito genuses makes Gaussian mixture discriminant analysis an inferior classification algorithm, diagnosing Alzheimer's disease stages is an objectively more difficult classification task than mosquito genus classification even for humans, which makes this argument entirely fallacious.

The prototypes viewed as "central tendencies" that Gaussian mixture discriminant analysis can be said to generate are the various statistical measures of center for each Gaussian distribution associated to each class. It is a proven mathematical theorem that for a standard Gaussian distribution, the mean, mode, and local maxima as statistical measures of centre all take the exact same value. In the following arguments we will consider the local maxima of generated Gaussian distributions as the functional analog of the prototype in prototype for reasons that will become apparent. Various degrees of membership to each category as a key merit of prototype theory can be manifested in this algorithm as how each point in the feature space are assigned varying likelihoods of belonging to each class by the Gaussians of each class.

The essence of Gaussian mixture discriminant analysis lies in its ability to create compound classes through 'mixing' the likelihood distributions of several individual classes by adding them together and reweighing them. In the example of diagnosing Alzheimer's disease, this is exemplified by creating a mixed Gaussian from all the individual Gaussians representing different stages of Alzheimer's disease, and using it to distinguishing a brain afflicted with

Alzheimer's disease of any stage from a healthy one. This along with our previous definition of prototype allows us to demonstrate the pet-fish problem with Gaussian mixture discriminant analysis as seen in *Figure 3*.

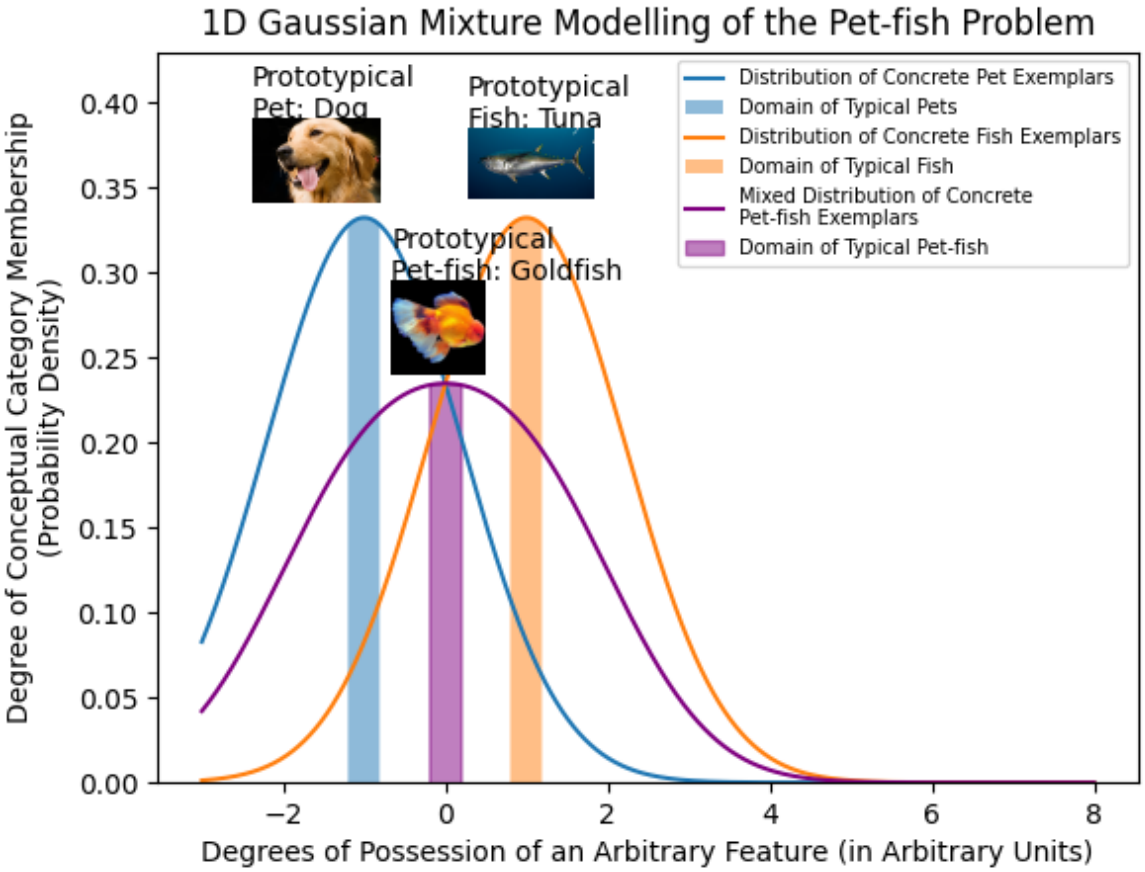

Figure 3. Demonstrating the Pet-fish Problem with Gaussian mixture modelling

*This figure is self-generated. See **appendix** for the Python code that produced this figure.*

In *Figure 3* we define some arbitrary domain in the feature space centered around the local maximum of each distribution as the domain of exemplars for each category that we consider typical. As we have successfully demonstrated with the purple mixture distribution, it is possible

to recreate a scenario in which the domain of typicality of the mixed distribution does not intersect the domain of typicality of the individual Gaussians that it is generated from.

Furthermore, in calibrating the parameters of the code that generated *Figure 3*, an arguably more interesting phenomenon emerged that revealed another behaviour of Gaussian mixture distributions, which is displayed on *Figure 4*.

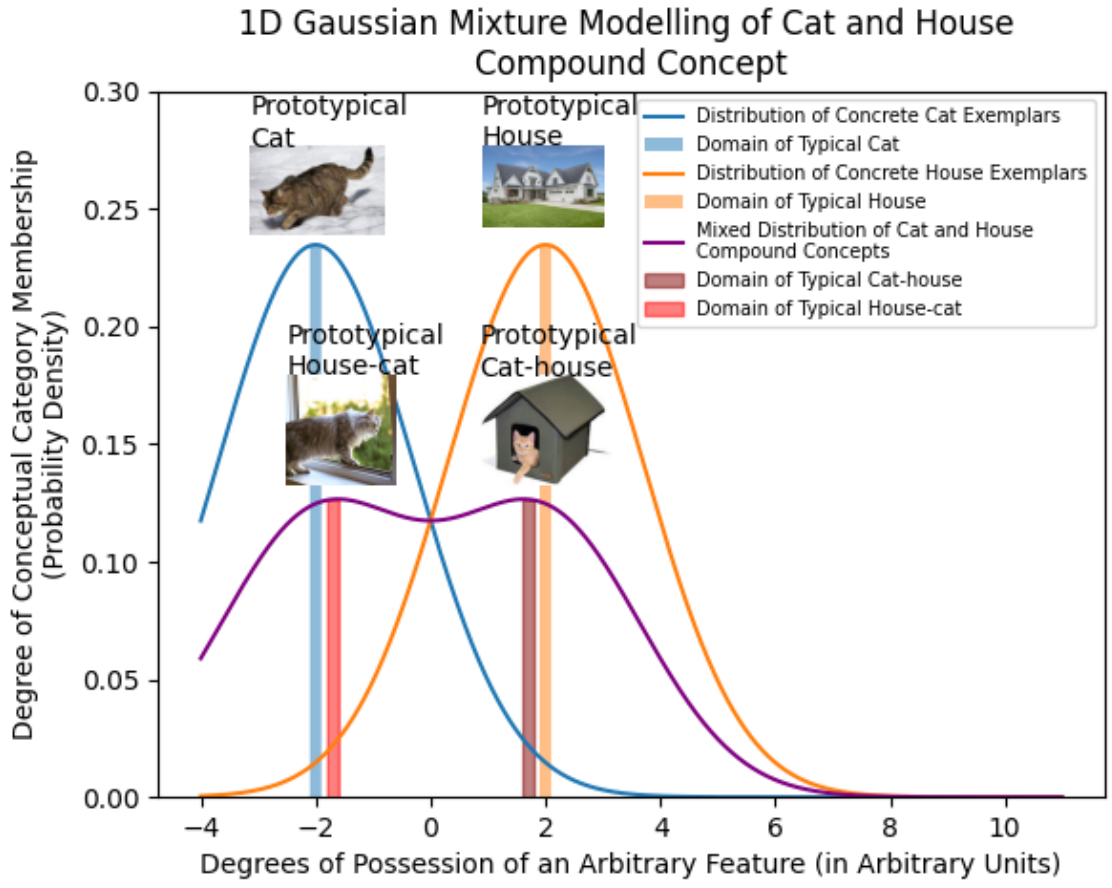

Figure 4. An example of Bi-maxima Mixed Gaussian Distribution

*This figure is self-generated. See **appendix** for the Python code that produced this figure.*

Figure 4 shows that if the individual Gaussians are sufficiently distant from one another, the resulting mixture will have two local maxima. This happens to be a quite accurate representation

of how human cognition combining two semantically distant concepts can generate two modes of combined concepts. Using the example in *Figure 4*, the conceptual combination of 'cats' and 'houses' has two modes of combined concepts 'cat-houses', and 'house-cats', where 'cat-houses' are houses made for cats more semantically close to 'houses' and 'house-cats' are cats that reside in houses more semantically close to 'cats'. Each of these modes then have their distinct domain of prototypical exemplars in feature space that the model on *Figure 4* is able to demonstrate.

The above analysis suggests that Gaussian mixture discriminant analysis is a superior model of human categorization than the nearest centroid classifier.

Dynamic k-means Clustering and Demonstrating Instability of Conceptual Prototype Under Varying Contexts

It is worth considering the counter-argument that supervised learning algorithms such as Gaussian mixture discriminant analysis and nearest centroid classifier are poor models of human categorization in general, as human categorization can happen spontaneously without the need for training. To address this, we will now look at an unsupervised classification algorithm.

Dynamic k-means clustering is based on the k-means clustering algorithm, which only needs as parameter the number of classes, “k”, that it is expected to discern from a set of unlabelled data points in features space to be able to perform classification without training. How this algorithm works is similar to the nearest centroid classifier. The algorithm begins its first iteration by randomly generating k-number of centroids in the feature space, upon which the nearest centroid classifier algorithm is performed to preliminarily classify these data points according to the random centeroids. In the next iteration, the algorithm uses the previous preliminarily classified data points as the training set to recalculate new locations for the k-number of centroids, and re-classify the data points accordingly. This iteration continues until convergence, which means the change of distance between old and new centroids of each class across two consecutive iterations becomes zero or sufficiently small, where the algorithm halts and is said to have found the optimum classification of the initially unlabelled dataset. *(Hossain et al., 2019)* This process can be visualized in *Figure 5.*

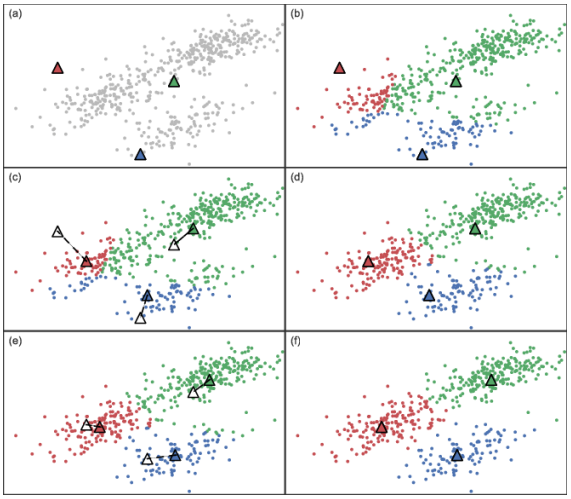

Figure 5. k-means Clustering for k=3 in 2D Feature Space
(Benavente et al., 2017)

On this figure: triangles represent the centroids of each class. Three iterations of the k-means clustering algorithm is demonstrated in the figure in the order of (a) to (f). In (a) three random centroids are generated, in (c) (e) the centroids are updated, in (d) the classification is updated, and in (f) optimal classification is achieved.

Empirically, k-means clustering is employed in tasks such as classifying health states of newborns based on features of their mothers in low income areas with poor medical infrastructure, which is able to achieve an average predictive accuracy of 67.58% *(Abbas et al., 2020)*. The same defense against fallacious arguments that lower predictive accuracy in a task requiring greater general intelligence implies an inferior algorithm used earlier can be applied here as well without loss of generality.

As k-means clustering also employs its final iteration centroids as a measure of "central tendency" of categories, the logic behind how the properties of centroids allows the nearest centroid classifier to demonstrate the merits of prototype theory can also be applied without loss of generality to dynamic k-means clustering if we also regard its final iteration centroids as prototypes. However, what differentiates dynamic k-means clustering from nearest centroid algorithm is that its prototype generation depends on category representation. An immediate counter argument to this claim that might initially seem intuitive is that the random centroids are generated first in this algorithm before the algorithm attempts classification. However, consider the fact that the initial locations of the randomly generated centroids do not have any deterministic predicting power over the optimal classification result, and can not be considered "central tendencies" of the algorithm's final category representation at all. Conversely, regardless of where we initialize the first iteration's centroids, we will always have the same optimal classification classes as result upon convergence. Further consider the final iteration of the algorithm, the optimal classification needs to be represented first before the final centroids could be positioned to check for convergence. This shows that centroids are merely indicators for whether the algorithm's current state of classification is optimal, and thus prototype generation depends on category representation for k-means clustering.

From how k-means clustering works, one can already discern how the instability of conceptual prototypes is demonstrated by how the location of the centroid for each class changes with each iteration towards convergence. The dynamic k-means clustering algorithm takes this even further by enabling the algorithm to further evolve with respect to newly added unseen unlabelled data points after convergence has been achieved *(Shafeeq & Hareesha, 2012)*. As for human

cognition defining contexts implies the provision of additional information, the additional new data points to a dynamic k-means clustering algorithm that has reached convergence can be considered a functional equivalent of variation of context. In this case, the addition of new data points to the system will cause its initial convergence classification to be no longer optimal, where the algorithm will seek a new convergence by changing the location of its classification and location of centroids again. This mirrors the instability of conceptual prototype and typicality gradient in human categorization, where one can observe the model continuously adapting to new contexts much like a human does. This shows that even for unsupervised learning, algorithms in which category representation enables prototype generation informs a more powerful model of human categorization.

Conclusion

Assuming human cognition is mechanizable, insights provided by analysis of classification machine learning algorithms nearest centroid classifier, Gaussian mixture discriminant analysis, and dynamic k-means classifier strongly imply that a theory of human categorization in which category representation enables prototype generation has greater explanatory power over the reverse, as represented by Eleanor Rosch's prototype theory.

Critical Evaluation and Scope for Further Investigation

The analysis presented above is far from a comprehensive list of classification algorithms in machine intelligence. A stronger conclusion can be reached if meta-analysis of a more comprehensive collection of classification algorithms could be performed. The differing performance of the three classification algorithms on different tasks noted earlier in this essay also calls for future experiments to be done on the three algorithms performing the same classification task to more definitively conclude that Gaussian mixture discriminant analysis and k-means clustering are superior classification algorithms than nearest centroid classifier.

Essay Word Count: 2184[1]

[1] Excluding title page, abstract, bibliography, appendix, section titles, in text citation, image captions, and footnotes.

Bibliography

Abbas, S. A., Aslam, A., Rehman, A. U., Abbasi, W. A., Arif, S., & Kazmi, S. Z. (2020). K-means and K-medoids: Cluster analysis on birth data collected in City Muzaffarabad, Kashmir. *IEEE Access*, *8*, 151847–151855. https://doi.org/10.1109/access.2020.3014021

Adcock, J., Allen, E., Day, M., Frick, S., Hinchliff, J., Johnson, M., Morley-Short, S., Pallister, S., Price, A., & Stanisic, S. (2015, December 9). *Advances in Quantum Machine Learning*. arXiv.org. Retrieved December 6, 2021, from https://arxiv.org/abs/1512.02900.

Alar, H. S., & Fernandez, P. L. (2021). Accurate and efficient mosquito genus classification algorithm using candidate-elimination and nearest centroid on extracted features of wingbeat acoustic properties. *Computers in Biology and Medicine*, *139*, 104973. https://doi.org/10.1016/j.compbiomed.2021.104973

Benavente, P., Protopapas, P., & Pichara, K. (2017). Automatic survey-invariant classification of Variable stars. *The Astrophysical Journal*, *845*(2), 147. https://doi.org/10.3847/1538-4357/aa7f2d

Dougherty, G. (2012). Supervised learning. *Pattern Recognition and Classification*, 75–98. https://doi.org/10.1007/978-1-4614-5323-9_5

Fang, C., Li, C., Cabrerizo, M., Barreto, A., Andrian, J., Loewenstein, D., Duara, R., & Adjouadi, M. (2017). A novel gaussian discriminant analysis-based computer aided diagnosis system for screening different stages of alzheimer's disease. *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*. https://doi.org/10.1109/bibe.2017.00-41

Griffiths, T. L. (2009). Connecting human and machine learning via probabilistic models of cognition. *Interspeech 2009*. https://doi.org/10.21437/interspeech.2009-2

Hossain, M. Z., Akhtar, M. N., Ahmad, R. B., & Rahman, M. (2019). A dynamic K-means clustering for data mining. *Indonesian Journal of Electrical Engineering and Computer Science*, *13*(2), 521. https://doi.org/10.11591/ijeecs.v13.i2.pp521-526

Johri, S., Debnath, S., Mocherla, A., SINGK, A., Prakash, A., Kim, J., & Kerenidis, I. (2021). Nearest centroid classification on a trapped ion quantum computer. *Npj Quantum Information*, *7*(1). https://doi.org/10.1038/s41534-021-00456-5

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. https://doi.org/10.1016/0010-0285(73)90017-0

Shafeeq, A. B. M., & Hareesha, K. S. (2012). Dynamic Clustering of Data with Modified K-Means Algorithm. *2012 International Conference on Information and Computer Networks (ICICN 2012)*, *27*. Retrieved from https://www.researchgate.net/profile/Ahamed-Shafeeq/publication/267752474_Dynamic_Clustering_of_Data_with_Modified_K-Means_Algorithm/links/54599d740cf2cf516483d7b6/Dynamic-Clustering-of-Data-with-Modified-K-Means-Algorithm.pdf.

Appendix

For the Python code that generated *Figures 3 & 4*, see GitHub repository:

https://github.com/a663E-36z1120/cognitive-science