Northeastern University

Early Childhood Development Clustering Algorithm

Final Project

Students: Mitchell Hornsby, Juliette Kamtamneni, and David Riggle

Course: Algorithms CS5800

Professor: Justin Kennedy

December 3, 2023

Table of Contents

Introduction and Team Members:

During the first years of life, children embark on an extraordinary journey of development. The human infant's brain doubles in size during the first year of life, forming 1 million new neural connections every second. This period is critical for laying the foundations of cognitive, emotional, physical, and social development. The process by which human babies learn is often sequential, such as rolling, sitting, and standing before walking. Similarly, for slightly older children, hand strength and pre-writing strokes come before tracing letters. These domains can be conceptualized as multiple ladders, representing cognitive, social and emotional, speech and language, fine motor, and gross motor skill development. Children climb each ladder at their own pace, and many skills within these domains are typically learned in sequence.

Team Members and Task Division:

Mitchell Hornsby: Designed and generated mock data and helped implement the sample program. Also helped with the design of the program.

Juliette Kamtamneni: Project idea, research, final project paper, and powerpoint.

David Riggle: Designed and implemented the sample program.

Hypothesis:

Our hypothesis is that the developmental "ladders" as described above, are interconnected. Contrary to the more traditional view of isolated skill progression, some skills across the major areas of early human development are often, but not always, learned closely together. For example, a child has just started to identify symbols, an exciting cognitive milestone. Does this new skill relate to simultaneous advancements in another developmental domain, such as the development of a specific emotional regulation milstone? Could the mastery of a new skill in one area prime the brain for significant developments in another? If validated, these interdependencies could reshape our understanding of early childhood development, and how we can best support young children.

Objective:

The goal of our project is to explore and lay a foundation for validating this hypothesis by using the predictive power of hierarchical clustering to identify complex relationships and predict the exact next new skill(s) a young child is poised to learn. The long-term goal is to empower parents by providing deep insights into their child's development. Caregivers could potentially maximize the rate of learning during the time in life when the brain has the highest plasticity.

Methodology:

Clustering is a technique in computer science used for discovering meaningful groups and patterns within an unlabeled dataset. It is a powerful data analysis tool, used when researchers do not have clearly labeled data. Interpreting the results of clustering is critical to make meaningful connections. Often, an expert in the field is needed to assist in correctly identifying meaningful clusters for specific problems.

The most popular clustering algorithms include k-means clustering, hierarchical clustering, density-based clustering, and distribution-based clustering. A few examples of where clustering is used include targeted marketing strategies, cancer analysis and treatment, genetic data analysis, anomaly detection in cybersecurity and insurance fraud, traffic management, and academic research. This project falls under the category of academic research, as we aim to explore data to create a framework for testing a hypothesis and finding patterns in data.

Algorithm:

The type of clustering we chose to use for this project is Hierarchical Clustering. We chose hierarchical clustering mainly for two reasons. First, a dendrogram provides a clear and intuitive visual representation of the relationships between skills. It will show how skills cluster together based on the age of acquisition. Parents can easily see which skills are closely linked and which are more distinct. This is helpful for us to see the potential relationship between skills and to provide meaningful insights.

Second, there is no need to specify the number of clusters. Hierarchical clustering can handle the complexity and diversity of a large dataset with many different skills. It doesn't require you to predetermine the number of clusters, which is advantageous given the wide range of skills we are considering.

Other reasons why hierarchical clustering was chosen as the best approach include:

- Hierarchical clustering can reveal both broad and fine-grained patterns in skill acquisition. For example, it can show groups of skills that are typically learned around similar ages, as well as more subtle relationships, like skills that form a transition between two broader skill sets.
- It is scalable to large datasets.

- It provides data visualization that is easy to interpret for non-expert users.

The time complexity for hierarchical agglomerative clustering is O(n^3), which makes it a slower algorithm. It can take either an agglomerative (bottom-up) approach, where each data point is initially considered a single cluster and then, or a divisive (top-down) approach, where all observations start in a single cluster before being divided into smaller ones. The most common approach is the agglomerative approach, which is the algorithm we used in this project.

Agglomerative hierarchical clustering builds clusters step by step. The distance between two clusters is defined as the shortest distance between any two points in each cluster. It works on the principle of iteratively merging clusters based on the minimum distance between them. As the algorithm is applied, data points are iteratively combined and can then be easily visualized in a dendrogram. The crucial step in hierarchical clustering is to interpret and cut the dendrogram into distinct clusters. This step is usually done by a data scientist or other expert working on the project.

Algorithm steps:

1. The algorithm starts by initializing each data point as a separate cluster.
2. Compute distance between all clusters.
3. Identify closest clusters.
4. Merge closest clusters. At each merge, total clusters decrease by one.

5. Update distance matrix. Recalculate distances between new clusters with all other clusters.
6. Repeat until all data points are merged into a single cluster.
7. Build a dendrogram.

Pseudocode:

```
function AgglomerativeClustering(data_points)
   initialize each data point as a single cluster
   while there is more than one cluster
      find the two closest clusters A and B (based on minimum distance)
      merge A and B into a new cluster C
      update distances between C and all other clusters
   end while
   return the cluster hierarchy
end function
```

The final product of the algorithm is a dendrogram, which visually represents the merging process and the structure of the clusters. When applied to our data, the dendrogram would show how closely related different skills are based on the average age at which children acquire them, providing a clear and intuitive visual representation of the relationships between skills. The finished dendrogram will show groups of skills that are typically learned around similar ages, as well as more subtle relationships, like skills that form a transition between two broader skill sets.

Most Prominent Drawbacks to our Approach:

- Efficiency: Agglomerative clustering can be computationally expensive for large datasets, as it requires recalculating distances and finding the minimum at each iteration.
- Sensitivity to Noise and Outliers: In single-linkage, the algorithm can be sensitive to noise and outliers, as it relies on minimum distances, which can be skewed by such anomalies.

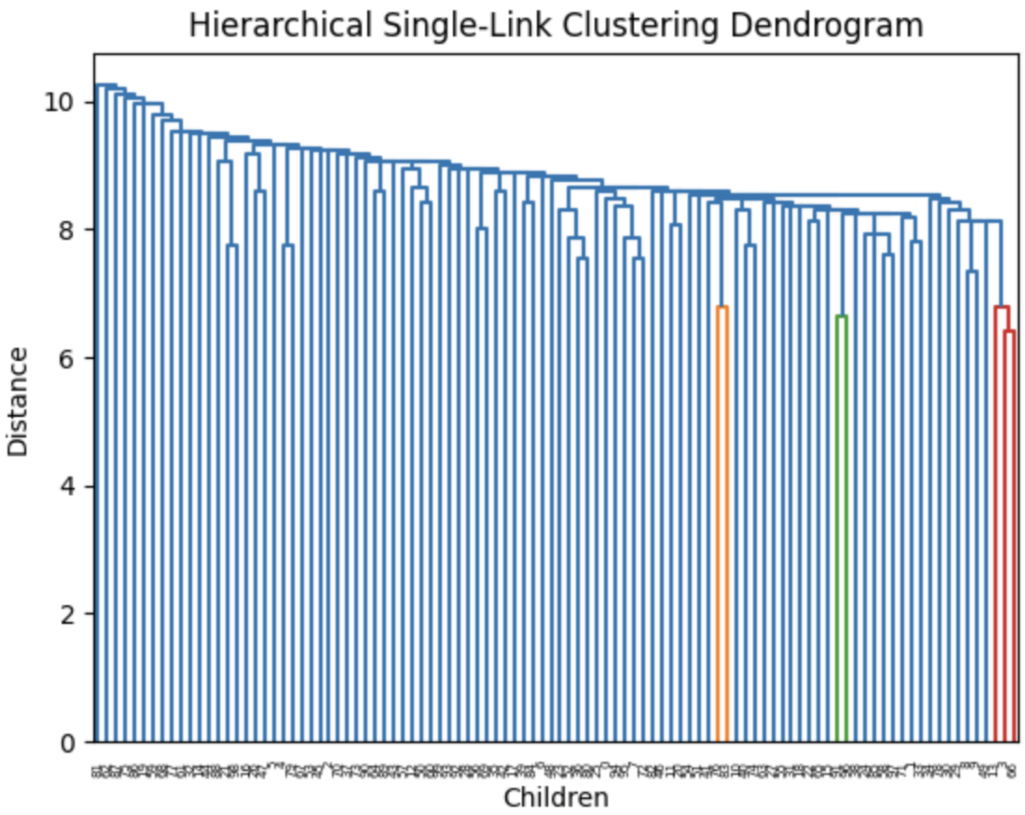
Mock Data:

Our mock data was created to be flexible for the number of children, number of skills and the max age for analysis. This data is populated by using a random distribution of the inputs. For our final analysis we decided to use a size of 100 children to show the breadth of the capability of the analysis. We created sample data for 20 generic skills they could attain by the age of 6. This array is populated by random integers to represent the ages that each child has achieved this skill with zero indicating they had not achieved it yet. Given the usage of the single linked cluster we needed to tune the threshold value in order to create clusters that were larger than 1 child. We created part of the program that would analyze the size and amount of clusters to determine what would make a nice tradeoff for informative sizes. We wouldn’t want clusters that only have one child and we wouldn’t want clusters where every child already shares the entirety of the skillset. This allows us to inform predictions and recommendations.

Description of Program:

Our proposed program is a comprehensive developmental tracking and recommendation system for young children, designed to analyze and predict skill acquisition patterns. It begins by collecting and inputting data on various developmental skills and the ages at which children typically acquire them, incorporating both broad skill categories and individual child data. Utilizing hierarchical clustering, it identifies clusters of skills commonly learned together, creating a dendrogram to visualize these relationships. Based on this analysis, the program compares individual children's skill acquisition timelines to the identified clusters, pinpointing areas of advanced development or potential delays. A recommendation engine then tailors activities and suggestions to support each child's next skill development, adapting to their unique learning journey. Our prototype program envisages that it might be appropriate to eventually integrate some kind of monitoring and alert systems, creating a feedback loop that allows for regular updates and adjustments based on the child's progress and new data patterns. An alert system, generating progress reports for parents or educators and alerting them to significant milestones or when additional services might be beneficial. Instead of implementing these additional steps, our program focuses on how clustering the skill acquisition data can allow you to predict the next skill.

Dendrogram Output:



Analysis of Data:

As our data was not real, it was difficult to create an accurate dendrogram that would reflect real data. It would take years and hundreds of children to yield an accurate dendrogram with meaningful clusters. But, in the case of real data, below are the steps on how we would process the data.

1. Extract Cluster Information: First, we need to extract meaningful information from the clustering results. This involves identifying which skills are grouped together and at what age range these clusters tend to appear. The interpretation of a dendrogram is usually performed by a data scientist or other expert in the field.

2. Individual Child's Skill Assessment: For each child in the program, we would assess their current skill set. Caregivers would record the age at which the child acquired each skill, and help determine the child's current skill set and those they have not yet acquired.

3. Comparing with Cluster Patterns: Compare the child's skill acquisition timeline with the cluster patterns. If we had years worth of data and well understood clusters, the program could help parents identify any advanced development or any potential delays.

4. Predicting the Next Skill: Based on the child’s current skills and the cluster analysis, the program could identify the next cluster and specific skill. The algorithm could identify skills the child hasn’t acquired yet, and suggest the most likely next skill.

5. Recommend Activities: Based on the skills identified, the program could recommend developmental activities tailored to practicing the next new skill the child is due to learn.

Discussion:

If our hypothesis proves to be true, and given unlimited time, resources, and access to data, we could create an exceptionally useful tool to help track and predict developmental milestones. With the use of a robust database, flawless clustering algorithm, and personalized predictive modeling, this idea has the potential to really help young children learn to the best of their abilities. Tailored recommendations would result in truly meaningful play to encourage development.

The most impactful method to engage parents would be to create a phone application with addictive features such that parents eagerly review newly acquired skills and recommended activities daily or weekly. The activities would be fun and rewarding for both children and parents. With consistent use, parents could earn points to unlock rewards within the application.

With all milestones updated in the application weekly, a well-written report could be sent to a pediatrician for review ahead of developmental appointments. In my (Juliette's) personal experience, pediatricians often miss crucial developmental milestones or differences. For example, yes a 3 year old has the capacity to share if nudged to do so while filling out a developmental form at the pediatricians office, but, do they share much more often or rarely compared to other children? A simple 'yes' or ‘no’ could indicate an advance or delay that is easily overlooked but could be nurtured in either case. Built-in alerts would be sent to the caregiver, offering recommendations for the next steps or access to an online specialist for discussion.

Although this is a small class project, if given unlimited resources, this program could revolutionize early childhood development by providing data-backed and deeply personalized insights and recommendations to caregivers. The current sample program is a tool for prediction, but with a large dataset, it could also evolve into a platform for further research.

Conclusion:

David: When I think about the original impetus for this project - Juliette's idea of having a robust and useful tool that can predict and suggest child development skills based on those acquired, and I think about the prototype tool we have now and the path between those two- it suggests to me the importance of good OOD principles, especially designing with modularity, encapsulation and abstraction of functions in mind so that we could iterate the design of the program and improve it by adding or replacing tools as we go along.

Mitchell: In conclusion, I was surprised by the amount of research that exists deploying similar clustering algorithms to classify young children for different purposes. I am hopeful that informative predictions can be made by similar strategies used by us in a clinical research environment. Every young child is different and milestones can be stressful for young parents; so having a sophisticated data driven framework for what to expect could be greatly beneficial to young families during an uncertain time in their lives.

Juliette: In conclusion, our project was small and foundational, but the goal of using hierarchical agglomerative clustering to analyze and predict early childhood development could be a significant step towards understanding exactly how children learn across the major developmental ladders. In the best case scenario, the program has the potential to provide caregivers with meaningful insights for a truly personalized tool to aid in the development of young children.

Citations:

1. *Ages and stages of development*. Ages and Stages of Development - Child Development (CA Dept of Education). (n.d.). https://www.cde.ca.gov/sp/cd/re/caqdevelopment.asp
2. Author links open overlay panelS.H.E. Dijkstra a, a, b, Highlights•Nine distinct clusters of learning behaviour were distilled with a data-driven approach.•The clusters differed significantly on learning metrics.•The clusters provide insight into cognitive knowledge and self-regulated learning., & AbstractWhen children are learning using adaptive learning technologies (ALTs). (2023, March 25). *Clustering children's learning behaviour to identify self-regulated learning support needs*. Computers in Human Behavior. https://www.sciencedirect.com/science/article/pii/S074756322300105X
3. Google. (n.d.). *What is clustering? | machine learning | google for developers*. Google. https://developers.google.com/machine-learning/clustering/overview
4. *Inbrief: The Science of Early Childhood Development*. Center on the Developing Child at Harvard University. (2020, October 29). https://developingchild.harvard.edu/resources/inbrief-science-of-ecd/
5. Sriram, R. (2020, June 24). *Why ages 2-7 matter so much for brain development*. Edutopia. https://www.edutopia.org/article/why-ages-2-7-matter-so-much-brain-development/
6. *Using Cluster Analysis to Interpret the Variability of Gross Motor Scores of Children With Typical Development*. Academic.oup.com. (n.d.). https://academic.oup.com/ptj/article/90/10/1510/2737783

7. Wikimedia Foundation. (2023, October 10). *Hierarchical clustering*. Wikipedia. https://en.wikipedia.org/wiki/Hierarchical_clustering