Deciphering Learning Curve Characteristics via K-Means Clustering of Curve Model Parameters

Enes Arda Ozgur

Supervisors: Tom Viering and Taylan Turan

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 28, 2024

Name of the student: Enes Arda Ozgur
Student Email: E.A.Ozgur@student.tudelft.nl
Final project course: CSE3000 Research Project
Thesis committee: Tom Viering, Hayley Hung

Abstract

Learning curves illustrate the relationship between the performance of learning algorithms and the increasing volume of training data [1, 2, 3]. While the concept of learning curves is well-established, clustering these curves based on fitting parameters remains an underexplored area. Our study delves into this domain and leverages the Learning Curve Database (LCDB) to discover potential patterns. We investigate whether different curve models uncover distinct patterns, examine the impact of different datasets on these learners, and explore if various learners display unique characteristics and behaviors or adhere to a common pattern. Curve model analyses conclude that most of the data points are in a single cluster (dominant cluster), indicating a potential commonality. Certain learners, such as QuadraticDiscriminantAnalysis and PassiveAggressiveClassifier, exhibit unique traits and do not conform to this common pattern, regardless of dataset attributes. Moreover, while various learners demonstrate similar characteristics within a single curve model, distinct patterns emerged when comparing across different curve models, indicating internal similarity but external divergence in behavior.

1 Introduction

In the field of Machine Learning (ML), it is commonly anticipated that an increase in training data improves model performance [1]. However, data collection and training involve significant costs, often referred to as the bottleneck of ML [4]. To address this, a common strategy is building models based on data samples, but the optimal sample size is still a subject of ambiguity and ongoing debate [5]. This challenge underscores the importance of learning curve research. Understanding curve behaviors and characteristics can provide insights into the necessary data volume for achieving a certain level of model performance [1, 6, 7, 8].

It's commonly assumed that if a certain parametric model fits a group of learning curves well, then we expect these curves to exhibit similar behaviours. Motivated by this reason, learning curves are fitted into 20 parametric models and extrapolated [1]. Despite of this effort, the field has many unexplored aspects, such as clustering the fitting parameters of these curve models. This paper aims to bridge this gap by investigating the following research question: "Can distinct patterns be detected in learning curves within the given LCDB by clustering their curve fitting parameters with K-Means clustering algorithm?"

This paper is structured as follows: Section 2 talks about related work. Section 3 outlines our methodology, detailing the process and sources for obtaining the fitting parameters of curve models, data preprocessing, and how we obtain the K value for our K-means clustering. Section 4 presents our experimental setup. The implications of the experiment results are then critically examined and discussed in Section 5. Section 6 concludes the paper with a summary of our findings and suggestions for future research directions in this field. Lastly, Section 7 is devoted to responsible research practices, addressing ethical considerations.

2 Related Work

Since Ebbinghaus's work in 1885, numerous researchers have endeavored to define the characteristics of learning [9]. Researchers have attempted to decipher the mathematical formulation of learning curves to determine if learning efficiency either maintains a steady rate of improvement or gradually slows down [10]. Yet, the results were inconclusive, revealing that learning curves are diverse, and no universal model has been established [1].

[11] delves deeply into this issue, suggesting that various factors, such as different datasets, classification algorithms, and processes, can fundamentally influence the shape of learning. While it may seem simple to average learning curves from various subjects, this process is actually more complex than it appears. Simple averaging can result in a power function for a group's performance, even when each individual's learning curves are exponential. Thus, it is important to recognize that averaging can alter the shape of learning. Consequently, it's recommended to focus on specific application domains for learning curves.

To gain a clearer understanding of learning curves, some studies have concentrated on fitting these curves into existing models [1]. These fits then serve as a basis for extrapolating the curves. However, there appears to be a lack of research on clustering these fitting parameters. Therefore, our research aims to identify patterns by using these fitting parameters, isolating datasets and learners as advised in [11].

3 Methodology

This section details the process and sources for obtaining the fitting parameters of curve models utilized in our clustering analysis. It outlines the steps involved in data preprocessing and explains the rationale behind choosing the k-means algorithm, as well as the measures taken to overcome its drawbacks.

3.1 Fitting Parameters of Curve Models

To conduct our experiments, we used the LCDB, which is a comprehensive collection of 4,367 learning curves from 20 learners applied to 246 datasets. 20 different curve models were fitted to these learning curves. Among the models analyzed in [1], MMF4 and WBL4 emerge as particularly notable when sufficient curve data are available for fitting. Therefore, our research primarily focuses on these two models, MMF4 and WBL4, with details of the used curve models presented in Table 1.

Reference	Formula
MMF4	$(ab + cn^d)/(b + n^d)$
WBL4	$c - b\exp(-an^d)$

Table 1: Parametric Learning Curve Models

3.2 Data Prepossessing

In data preprocessing, particularly for learning curve data, it's essential to identify and eliminate bad fits to ensure the accuracy and reliability of the analysis. These bad fits can arise from various sources, such as measurement errors, data entry mistakes, or anomalies in the data collection process. We followed the same approach as [1] to define bad fits. Removing these inaccuracies is crucial because they can skew the results, leading to incorrect conclusions and potentially misleading insights into the learning process. The overall preprocessing approach in our research is aligned with that outlined in [1].

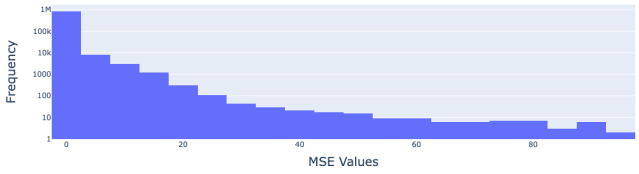


Figure 1: Histogram: Distribution of Mean Squared Error (MSE) Values

Moreover, we incorporated an additional filtering criterion based on the distribution characteristics of the MSE values, see Figure 1. We decided to remove extreme outliers after observing a distinct right skew in the MSE distribution, with the mean `(0.1610)` considerably higher than the median `(0.0008)`. Specifically, we eliminated data points with MSE values exceeding 0.0070, representing the 75th percentile. This decision was guided by the aim to exclude abnormally high errors that could obscure the underlying patterns in the learning curve data.

The fitting paramaters, central to our clustering analysis, underwent diverse preprocessing treatments across different experiments. In some instances, we standardized these values using a StandardScaler [12]. This step was crucial for algorithms like K-Means, which depend on distance metrics and are sensitive to the scale of data. Conversely, other experiments purposefully omitted this standardization to assess the impact of the original scale of fitting parameters on clustering outcomes. By adopting these varied approaches, we aimed to gain comprehensive insights into how different preprocessing techniques influence the pattern detection process in learning curve data.

3.3 K-Means

Our research investigates the similarity of learning curves by analyzing their fit to two different curve fitting models, employing k-means clustering for the respective fitting parameters. K-means was initially chosen to facilitate comparison with the work of another researcher in the same lab. However, due to the scope of the projects, this comparison was not feasible in the end. A notable drawback of K-means is its sensitivity to the initial points. To mitigate this, we executed the K-means algorithm hundred times per experiment to ensure the reliability and consistency of the cluster assignments and addressing the randomness in the algorithm's initialization. We then took the average of these 100 clusters as the final result.

To determine the optimal number of clusters (K), we adopted two strategies. The first strategy involved techniques like the silhouette score [7] and the elbow method [13] to select a K value balancing data segmentation effectively. The second strategy predetermined the K value based on specific research requirements and hypotheses. All K-Means implementations were conducted using the Scikit-Learn library [14].

3.4 K Value

To determine the K value for our clustering, we explored three approaches: the silhouette score, the elbow method, and experimentation with various K values.

The first approach involved using the silhouette score to find the optimal K value. In Experiment 1, the silhouette score peaked at K = 2 for MMF4 (see Figure 2), and WBL4 (see Figure 3). However, since the score differences between K values were marginal, these results were not considered entirely reliable. The elbow method yielded similar outcomes and led us to base the K value on our hypothesis.

We initially considered K = 20, hypothesizing that each cluster might represent an individual learner. This assumption proved incorrect, yet it revealed another crucial aspect: most of the data points are concentrated in a single cluster.

Figure 2: Silhouette Scores for MMF4

Figure 3: Silhouette Scores for WBL4

We then incrementally increased the K value for each curve model, starting from 2. Initially, for K =2, the result showed one diverse cluster containing most of the data, while the second cluster was represented by a single learner. We continued to increase the K value one by one, observing that each new cluster tended to correspond to an individual learner. We kept incrementing the K value until this pattern no longer held true.

As a result, we settled on K = 5 for MMF4 and K = 3 for WBL4. The detailed implications of these choices for our experiments are discussed later. For all the experiments in this research, we will consistently use K = 5 for MMF4 and K = 3 for WBL4.

4 Experiments

Our study encompassed three distinct experimental setups, each designed to explore different aspects of learning curve clustering.

4.1 Experiment 1: Curve Model Analysis Across Datasets and Learners

Our first setup involves all datasets, all learners, and a single curve model. This setup is aimed at understanding the generalizability and comparative analysis across a wide spectrum of datasets and learners under a uniform curve model. The intent is to identify overarching patterns, commonalities, and divergences in learning curve behaviors, contributing to a holistic understanding of the learning process in various learning scenarios.

4.2 Experiment 2: Dataset Analysis Across Learners

Experiment 2 is, in its essence, a repetition of Experiment 1 with the difference that we now consider each dataset at a time instead of merging all into one, larger dataset as done in Experiment 1. This approach allows us to compare and contrast the behaviors and patterns of different learners within the same dataset and curve model. The goal is to identify unique clustering patterns and trends that are specific to different learners, thereby offering insights into the learners' performance variations under identical dataset and curve model conditions.

4.3 Experiment 3: Interactions of Learners with Different Datasets

In the third and final setup, we broadened our horizon to encompass all datasets while maintaining a focus on a single learner and a single curve model. This setup is designed to explore how one learner and curve model interact with all datasets. The purpose is to uncover the variability in learning curve patterns due to dataset differences, providing a broader perspective on the learner's adaptability and performance across varied data scenarios.

5 Results and Discussion

This section details the outcomes and analyses of the previously mentioned experiments.

5.1 Experiment 1: Curve Model Analysis Across Datasets and Learners

The results from Experiment 1 reveal intriguing distinctions and similarities between the MMF4 and WBL4 curve models. In both MMF4 and WBL4 models, the clustering patterns exhibit striking similarities. Most notably, all clusters apart from 0, in both models are characterized by an absolute dominance of a single learner type, as seen in Tables 2 and 3. This is evidenced by the Perceptron and BernoulliNB learners for WBL4 and SGDClassifier, SVC sigmoid, SVC poly and DecisionTreeClassifier for MMF4, each accounting for 100% of their respective clusters. Such uniformity suggests that these clusters are highly specific to the learning styles of these individual learners.

Cluster	Learner	Percentage
0	PassiveAggressiveClassifier	5.76%
1	SGDClassifier	100.00%
2	SVC_sigmoid	100.00%
3	SVC_poly	100.00%
4	DecisionTreeClassifier	100.00%

Table 2: Most Dominant Learner and Percentage per Cluster - MMF4

Conversely, Cluster 0 in both models presents a notable contrast. Unlike other clusters, it is not dominated by any single learner type. In MMF4, the PassiveAggressiveClassifier represents only 5.76% of Cluster 0, while in WBL4, the SVC_linear learner comprises just 5.74%. These low percentages point towards a significant diversity within Cluster 0 (dominant cluster), indicating the presence of multiple learner types that are not as distinctly separable as in other clusters. The similarity in this diversity between MMF4 and WBL4 is particularly interesting, as it implies that despite the differences in the curve models, they both identify a similar level of heterogeneity in certain learning scenarios.

Cluster	Learner	Percentage
0	SVC_linear	5.74%
1	Perceptron	100.00%
2	BernoulliNB	100.00%

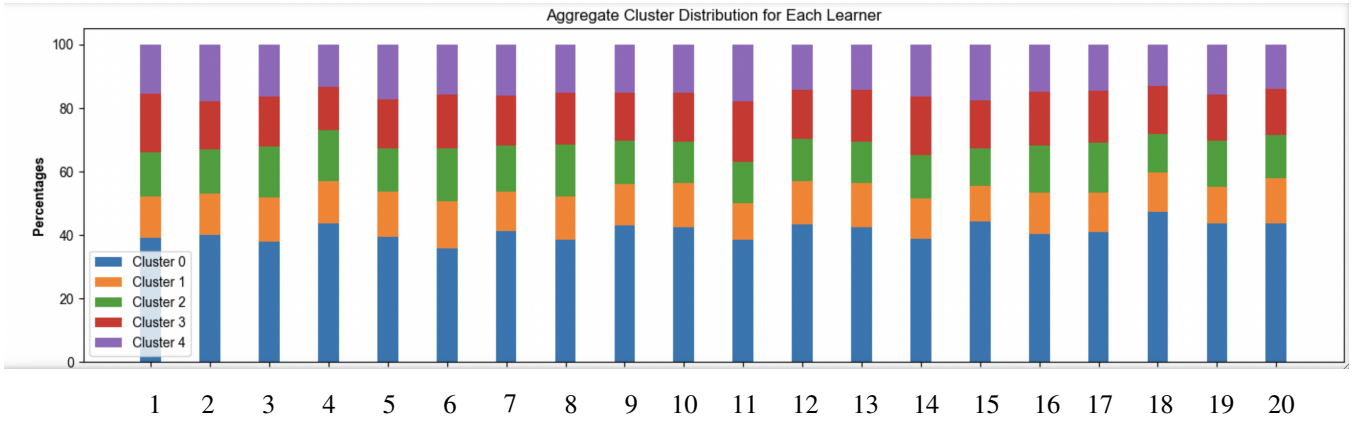
Table 3: Most Dominant Learner and Percentage per Cluster - WBL4

5.2 Experiment 2: Dataset Analysis Across Learners

To understand the unique clustering patterns and trends specific to different learners within a single dataset, we gather statistical data on whether data points of each learner reside in the dominant cluster or not for each individual dataset.

Experiment 2 yielded results similar to Experiment 1. The MMF4 and WBL4 models are distinguished by a clear dominance of individual learner types. Table 4 shows the relative count in which a learner predominantly does not fall into the dominant cluster. For example, the count for SVC_linear is 22, indicating that most data points for SVC_linear did not reside in the dominant cluster in 22 out of the 246 datasets.

The most striking finding was that the QuadraticDiscriminantAnalysis learner is the leading one whose data points do not predominantly reside in the dominant cluster, with the respective count of 57 in MMF4 and 49 in WBL4. This consistent pattern across both models suggests that this learner tends



1	SVC_linear	6	GradientBoostingClassifier	11	RidgeClassifier	16	MLPClassifier
2	SVC_poly	7	RandomForestClassifier	12	SGDClassifier	17	DecisionTreeClassifier
3	SVC_rbf	8	LogisticRegression	13	BernoulliNB	18	ExtraTreeClassifier
4	SVC_sigmoid	9	PassiveAggressiveClassifier	14	MultinomialNB	19	LDA
5	ExtraTreesClassifier	10	Perceptron	15	KNeighborsClassifier	20	QDA

Figure 4: All learners exhibit similar interactions with datasets when analyzed using the MMF4 curve model.

Learner	MMF4 Count	WBL4 Count
SVC_linear	22	16
SVC_poly	32	18
SVC_rbf	20	18
SVC_sigmoid	8	38
LinearDiscriminantAnalysis	12	22
QuadraticDiscriminantAnalysis	57	49
ExtraTreesClassifier	10	10
GradientBoostingClassifier	13	20
RandomForestClassifier	10	5
LogisticRegression	16	13
PassiveAggressiveClassifier	36	25
Perceptron	0	30
RidgeClassifier	18	20
SGDClassifier	36	31
BernoulliNB	19	20
MultinomialNB	11	11
KNeighborsClassifier	28	16
MLPClassifier	22	12
DecisionTreeClassifier	15	14
ExtraTreeClassifier	28	16

Table 4: The number of datasets in which a learner predominantly does not fall into the dominant cluster

to identify unique data characteristics not commonly captured by the dominant trends, regardless of dataset characteristics.

Learners such as PassiveAggressiveClassifier (36), SGDClassifier (36), KNeighborsClassifier (28), and ExtraTreeClassifier (28) show similar count for MMF4. This commonality indicates that these learners may share a similar aspect of their algorithms that is less frequently represented in the dominant cluster patterns.

Interestingly, the RandomForestClassifier and Perceptron have most of their data points residing in the dominant cluster. This suggests that both learners do not easily get affected by different types of datasets, indicating a level of robustness and adaptability to various data characteristics.

5.3 Experiment 3: Interactions of Learners with Different Datasets

The aggregate cluster distribution data from both models presents intriguing patterns and insights.

In MMF4, most learners predominantly fall into Cluster 0, as seen with tree.ExtraTreeClassifier (47.23%). This trend suggests a common cluster configuration responsive to a variety of learners, see Figure 4. Interestingly, all learners exhibit similar, but not identical. This indicates a certain uniformity in how different learners interact with datasets within the MMF4 model.

WBL4's analysis reveals a highly skewed distribution, with almost all learners predominantly in Cluster 0 (ranging from 99.61% to 99.94%). This skewness suggests a model that generalizes across different learners, capturing a common pattern that most learners adhere to. The minimal representation in Clusters 1 and 2 (ranging from 0% to 0.39%) across all learners in WBL4 indicates that these clusters might capture rare or unique learner behaviors, that are not as frequently observed.

To wrap up, both MMF4 and WBL4 show a dominant Cluster 0; however, the degree of dominance is more explicit in WBL4. While MMF4 demonstrates some level of diversity in cluster distribution, WBL4 tends to generalize learners into a single dominant cluster. This distinction suggests that MMF4 might be more sensitive to capturing diverse learner behaviors. Although MMF4 and WBL4 have different outcomes, the learners within the same curve models are similar. This understanding is key to advancing our comprehension of how different learning algorithms interact with varying datasets through the lenses of different curve models.

6 Conclusion and Future Work

In this study, we explored the potential of k-means clustering in deciphering patterns within a learning curve database, utilizing curve model fitting parameters. Our approach, through three distinct experimental setups, yielded noteworthy insights into learning curve behaviors and characteristics.

Experiment 1 revealed that most data points for both MMF4 and WBL4 models reside in a single, diverse, and dominant cluster. Furthermore, with an appropriately selected K value, other clusters can be represented by individual learners. This suggests that although many learners exhibit similar characteristics, there are some that possess unique traits. However, the study's limitation lies in its exclusive focus on the MMF4 and WBL4 models. Future research should broaden its scope to include all 20 curve models available in the dataset, providing a more holistic view of learning curve dynamics. Additionally, there's a need to delve deeper into the dominant cluster to understand a wider spectrum of learning behaviors.

Experiment 2 revealed that some learners, like QuadraticDiscriminantAnalysis, have distinguishable characteristics and can be detected regardless of datasets' characteristics. Oppositely, some learners, like RandomForestClassifier and Perceptron do not have distinguishable characteristics and follow a common pattern. Experiment 2 can be richened by exploring the QuadraticDiscriminantAnalysis learner through other curve models. Furthermore, for the same datasets in both MMF4 and WBL4, certain learners predominantly do not fall into the dominant cluster. Further examination of these datasets is necessary to comprehend the underlying reasons.

In Experiment 3, we delved into how individual learners adapt and perform. The findings from the MMF4 model suggested a trend towards a common cluster configuration responsive to multiple learners. In contrast, the WBL4 model demonstrated a skewed distribution, predominantly clustering learners into a single group. This difference emphasises the importance of selecting the appropriate curve model for learning curve analysis, as different models like MMF4 and WBL4 offer varied insights into learner behaviors and patterns. Furthermore, given the contrasting results in Experiment 3, additional testing across various curve models is necessary to validate these findings.

7 Responsible Research

Reproducibility is very important in research. It helps researchers check how strong their results are and lets others confirm these results. To make sure others can reproduce our work, we explained the methodology and experiment setups in a detailed way.

ChatGPT was utilized during the research and report-writing phases of this study. Examples and used prompts can be found in the Appendix.

References

[1] T. Viering and M. Loog, "The Shape of Learning Curves: A Review," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 6, pp. 7799-7819, 1 June 2023, doi: 10.1109/TPAMI.2022.3220744.

[2] A. K. Jain, R. P. W. Duin and Jianchang Mao, "Statistical pattern recognition: a review," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4-37, Jan. 2000, doi: 10.1109/34.824819.

[3] Michel Jose Anzanello, Flavio Sanson Fogliatto, "Learning curve models and applications: Literature review and research directions" in International Journal of Industrial Ergonomics, vol 41, issue 5, pp 573-583, 2011, doi: 10.1016/j.ergon.2011.05.001.

[4] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: A big data-ai integration perspective," pp. 1328–1347, 4 2021.

[5] Meek, Christopher, Thiesson, Bo Heckerman, David. (2002). The Learning-Curve Sampling Method Applied to Model-Based Clustering. Journal of Machine Learning Research. 2. 397-418. 10.1162/153244302760200678.

[6] J. Brownlee, "Optimization for Machine Learning," Machine Learning Mastery, 2021. [Online]. Available: https://books.google.nl/books?id=tW1HEAAAQBAJ.

[7] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, NSW, Australia, 2020, pp. 747-748, doi: 10.1109/DSAA49011.2020.00096.

[8] Roca, Thomas, (2014), SCATTER3D: Stata module to create 3D scatter plots for the web, using HTML5 3D feature and java api developped by CanvasXpress.

[9] Jacobs, J. (1885). H. Ebbinghaus, Ueber das Gedächtnis. Mind 10:454.

[10] Lewandowsky, Farrell, S. (2011). Computational modeling in cognition: Principles and practice. Thousand Oaks: Sage.

[11] Murre, J.M.J. S-shaped learning curves. Psychon Bull Rev 21, 344–356 (2014). https://doi.org/10.3758/s13423-013-0522-0

[12] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi and V. Padma, "Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 729-735, doi: 10.1109/ICSSIT48917.2020.9214160.

[13] Rena Nainggolan et al 2019 J. Phys.: Conf. Ser. 1361 012015

[14] Ahmed M, Seraj R, Islam SMS. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics. 2020; 9(8):1295. https://doi.org/10.3390/electronics9081295

A Appendix - ChatGPT

In this section, we detail the various prompts used with ChatGPT during different stages of the project.

During the Research Phase: ChatGPT was employed to assist in identifying coding errors, offering suggestions to enhance coding skills and overall code quality. It was also utilized to learn different methods of visualizing clustering results. Examples of prompts used include:

- ”I am encountering this error in my code, what could be causing it?”
- ”How can I effectively visualize my clustering results?”

While Writing the Report: The primary role of GPT in the report-writing phase was to check for grammatical errors, provide advice on paragraph structuring, combine sentences coherently, suggest transitions, identify potential mistakes, assist in brainstorming necessary content, and find suitable words. Prompts in this phase included:

- ”Can you perform a grammar check on this paragraph?”
- ”How can I merge these two sentences effectively?”
- ”Could you suggest a transition between these paragraphs?”
- ”Please analyze this paragraph. What elements might be missing?”
- ”What are the key components to include in an abstract?”
- ”What’s an alternative word for this term?”
- ”How do I create a table in LaTeX on Overleaf?”
- ”Can you provide guidance on writing a concise yet informative conclusion?”