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

Enes Arda Ozgur
E.A.Ozgur@student.tudelft.nl

TU Delft

Supervisors: Dr. Tom Viering and Taylan Turan

1 - Background

- **Learning Curves** illustrate the relationship between the performance of learning algorithms and the increasing volume of training data.
- Learning curves are **diverse**, and **no universal model** has been established [1].
- Various factors can **influence** the shape of Learning curves [2].
- Learning curves are fitted into **20 parametric models**, assuming similar curve models behave [1].
- **Lack of research** on clustering these fitting parameters.

2 - 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?"

Hypothesis: curve model, learner and dataset types affects the clustering output and distinct patterns can be detected.

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

3 - Experiment

3 experiment setup:

- Curve Model Analysis Across Datasets and Learners
- Dataset Analysis Across Learners
- Interactions of Learners with Different Datasets

4 - Results

1

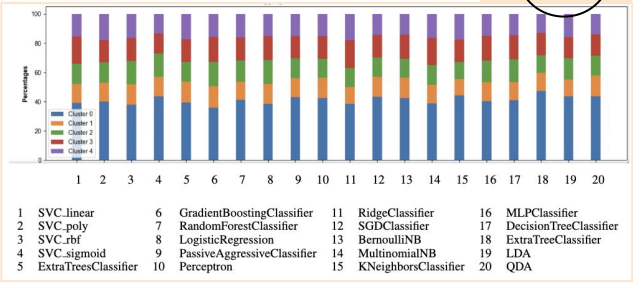
Cluster	Learner	Percentage
0	SVC_linear	5.74%
1	Perceptron	100.00%
2	BernoulliNB	100.00%

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%

2

Learner	MMF4 Count	WBL4 Count
QuadraticDiscriminantAnalysis	57	49
Perceptron	0	30
RandomForestClassifier	10	5

3



5 - Conclusion

- Most data points for both MMF4 and WBL4 models **reside in a single, diverse, and dominant cluster.**
- Other clusters can be represented by individual learners.
- Some learners, like **Quadratic Discriminant Analysis**, have distinguishable characteristics and can be detected regardless of datasets' characteristics.
- 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**.

6 - Limitations & Future Work

- **Exclusive focus** on the MMF4 and WBL4 models.
- Broaden its scope to include all **20 curve models.**
- Delve deeper into the **dominant cluster.**
- For the **same datasets in both MMF4 and WBL4, certain learners predominantly do not fall into the dominant cluster.** Further examination of these datasets.
- Given the contrasting results in Experiment 3, **additional testing across various curve models** is necessary