“How much Data is Enough?” Learning Curves for Machine Learning

Can patterns be identified amongst learning curves after the application of the K-Means algorithm using point and statistical vectors?

Background

Learning curves can be used to indicate the “performance of trained models versus the training set size” [1]. The current state of research on learning curves is that it is not yet well-understood.

Clustering refers to the grouping of items based off a similarity metric. A curve can be transformed into a vector using its generalised features.

Methodology

All curve data available is in the Learning Curve Database (LCDB) with 20 working learners on 250 datasets.

Metrics: Point Vector and Statistical Vector

Point Vector: Consists of each point of the learning curve of each dataset and interpolated to same length.

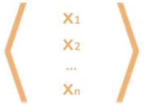

Figure 1: Points Vector

Statistical Vector: Consists of mean, standard deviation, skew, and kurtosis. Scaling conducted using *sklearn MinMaxScaler* [2] to ensure not one scale dominates the clustering.

$\bar{x}$
σ
S
K

Figure 2: Statistical Vector

K-Means Algorithm: This is an algorithm that generates groups based off the provided metric. Optimised K found through Silhouette Score method.

Results

2D Principal Component Analysis (PCA) to visualise distribution.

Point Vector (PV): The Silhouette Score gave an optimised K = 2.

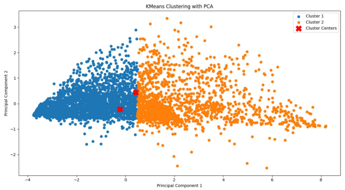

Figure 3: PCA 2D Plot PV

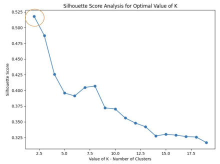

Figure 4: Silhouette Score PV

Statistical Vector (SV): The Silhouette Score gave an optimised K = 2.

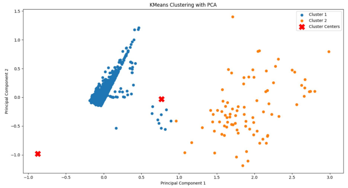

Figure 5: PCA 2D Plot SV

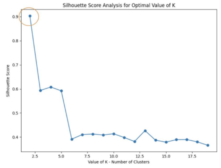

Figure 6: Silhouette Score SV

Discussion

Analysis PV: The PCA plot for the point vectors indicated that 2702 points in C0 and 1165 points in C1. 6 of the learners are within the range of 50-60% of datasets in C0, meaning fairly equal distribution.

An apparent relation among ensemble learning techniques surfaced in the learner analysis.

Analysis SV: The PCA plot for the statistical vectors indicated that 4169 points in C0 and 110 points in C1. All learners are in range 90-100% for C0.

Clear isolated patterns could not be determined after clustering the statistical vectors.

Conclusion

Findings: Exact equivalence relations could not be established through K-means clustering.

Future Work:

- Anchor points in interpolation of point vector
- Use of model-based clustering algorithm
- Investigate overlapping clusters formed in PV

References

[1] Viering, T. (2023). *"How Much Data is Enough?" Learning Curves for Machine Learning.* Project Forum. https://projectforum.tudelft.nl/course_editions/74/generic_projects/4899

[2] Scikit-learn (2023). *sklearn.preprocessing.MinMaxScaler.* (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Author Pravesha Ramsundersingh **Contact** p.s.p.ramsundersingh@student.tudelft.nl

TUDelft

Supervisors Tom Viering, Taylan Turan **Responsible Professor** Tom Viering