Supplementary Material: Learning Self-Growth Maps for Fast and Accurate Imbalanced Streaming Data Clustering

Yiqun Zhang, *Senior Member, IEEE*, Pengkai Wang, Sen Feng, Zexi Tan, Xiaopeng Luo, Yuzhu Ji, *Member, IEEE*, Rong Zou, and Yiu-Ming Cheung, *Fellow, IEEE*

A Parameter Sensitivity Evaluation

In the proposed SOHI framework, there are two parameters, namely the initial number of neurons Q and the neighborhood count κ. The varying values of these parameters may affect the clustering performance in various ways. Here is an explanation of these parameters.

Regarding the parameter Q used in the initialization of the sub-networks in SO, it directly determines the initial number of neurons, that is, the number of initial sub-networks. The value of Q is not set too large, as our goal is to allow the network to grow automatically to an appropriate size. On the other hand, a too small value of Q can lead to a scarcity of sub-networks, thereby causing issues with bridge nodes, especially when Q is only 3, SO will degenerate to a situation with only one network, directly affecting the quality of the SOHI framework.

Regarding the parameter κ used in the approximate κ-neighbors algorithm in HI, it calculates the global separability value for a specific cluster. When κ is too large, it may lead to increased computational overhead. Conversely, when κ is too small, it may render the global separability value meaningless.

We conduct experiments on eleven datasets by varying the values of Q and κ, recording the obtained ARI, NMI and DBI scores, as shown in Fig. 1 to Fig. 22. Q values are traversed from 3 to 30, while κ values range from 1 to 20. Specifically, κ performs well and stabilizes between 7 to 15. Different values of Q produce greater changes in accuracy than κ. Larger values of Q generally imply better results; the effects tend to converge once Q exceeds 15. Generally speaking, the SOHI framework is less sensitive to parameter variations.

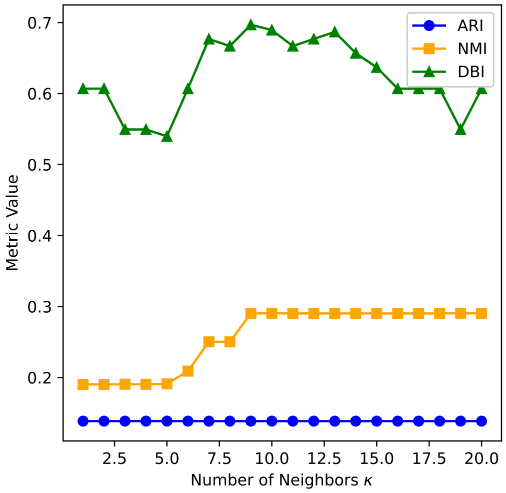

Figure 1: Clustering performance on AB dataset w.r.t. different values of κ

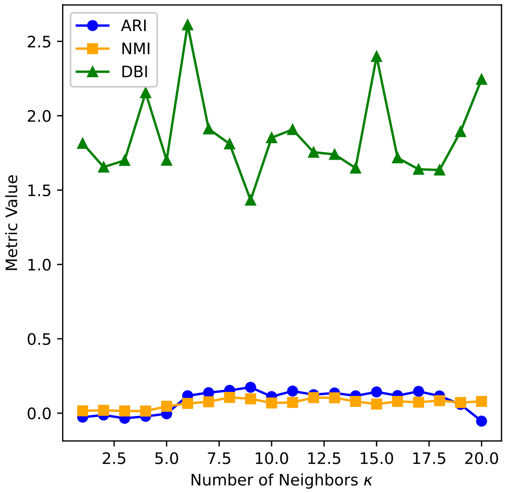

Figure 2: Clustering performance on CAE dataset w.r.t. different values of κ

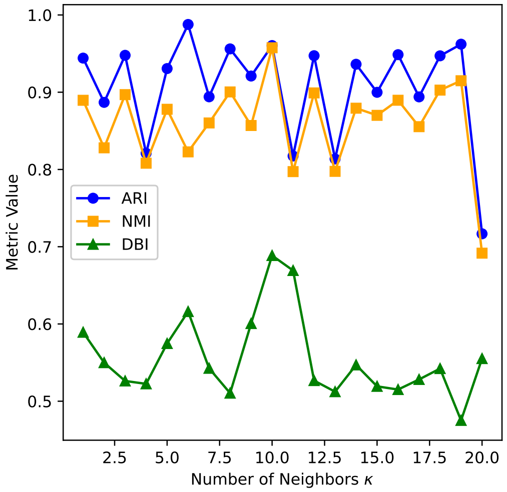

Figure 3: Clustering performance on GAS dataset w.r.t. different values of κ

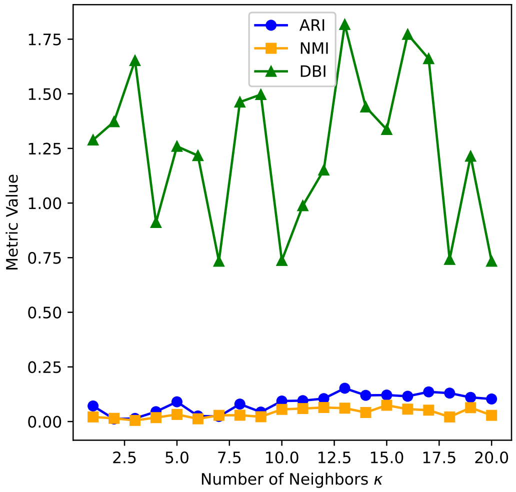

Figure 4: Clustering performance on HM dataset w.r.t. different values of κ

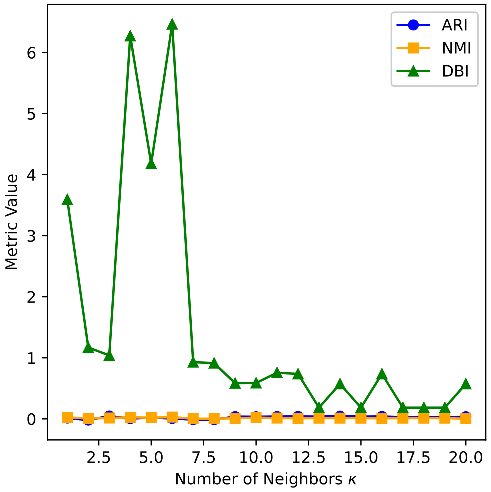

Figure 5: Clustering performance on HFCR dataset w.r.t. different values of κ

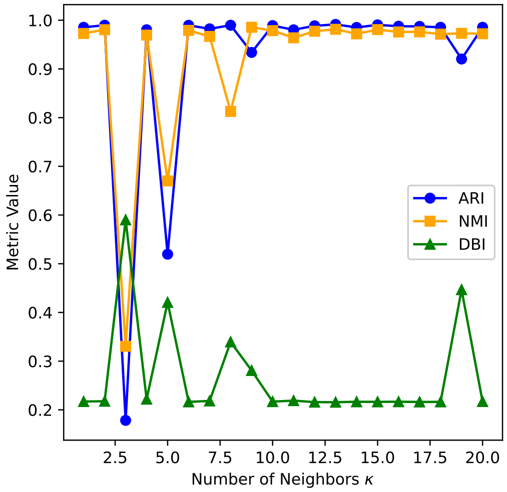

Figure 6: Clustering performance on IDS2 dataset w.r.t. different values of κ

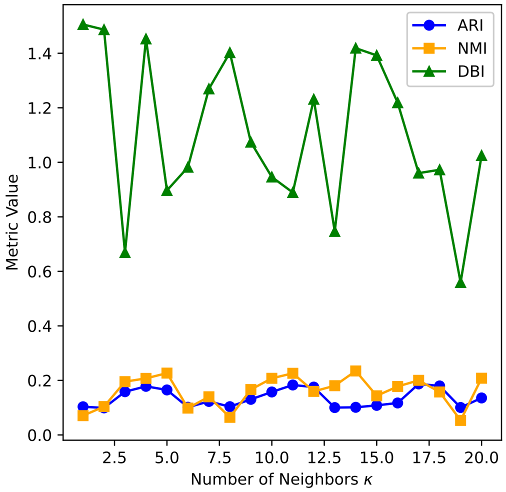

Figure 7: Clustering performance on MD dataset w.r.t. different values of κ

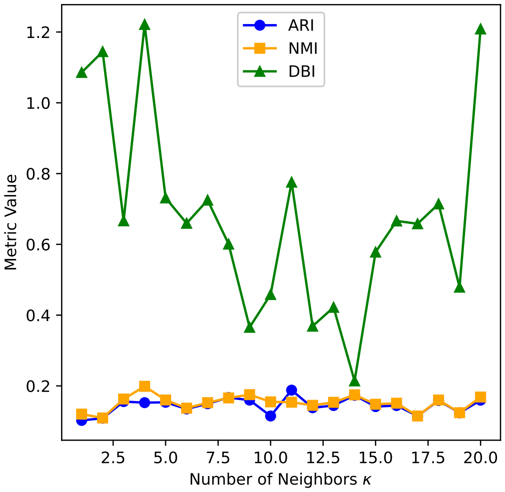

Figure 8: Clustering performance on PB dataset w.r.t. different values of κ

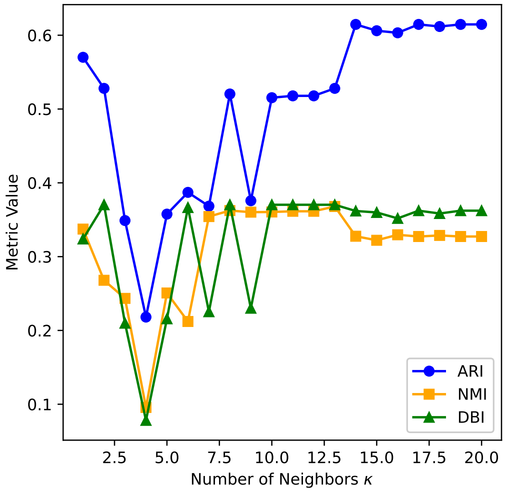

Figure 9: Clustering performance on RA dataset w.r.t. different values of κ

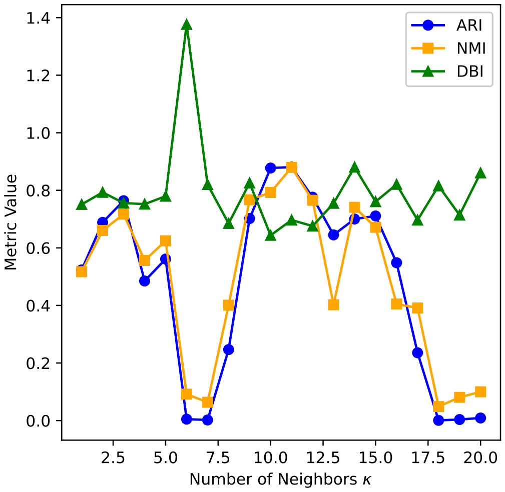

Figure 10: Clustering performance on SD dataset w.r.t. different values of κ

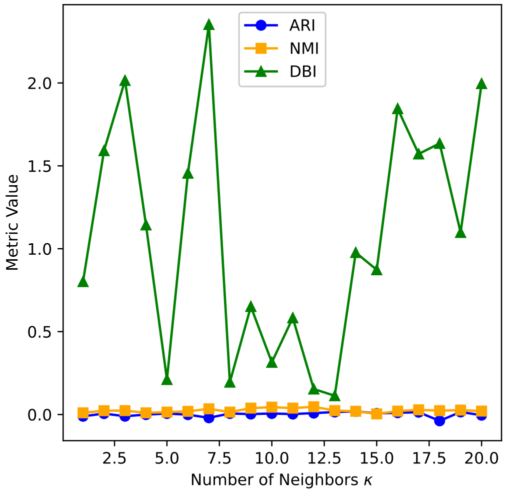

Figure 11: Clustering performance on WC dataset w.r.t. different values of κ

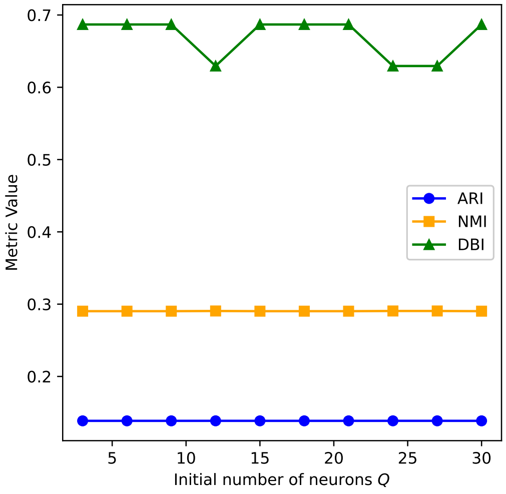

Figure 12: Clustering performance on AB dataset w.r.t. different values of Q

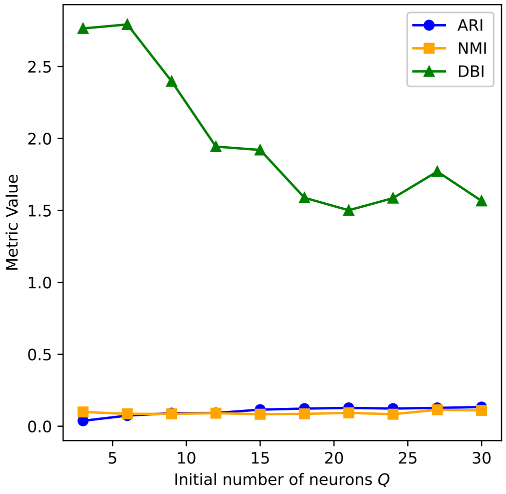

Figure 13: Clustering performance on CAE dataset w.r.t. different values of Q

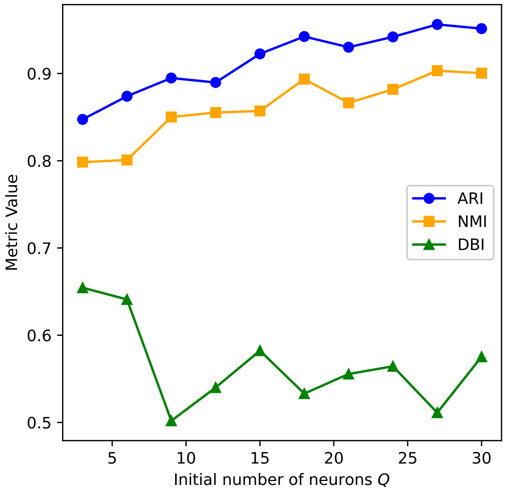

Figure 14: Clustering performance on GAS dataset w.r.t. different values of Q

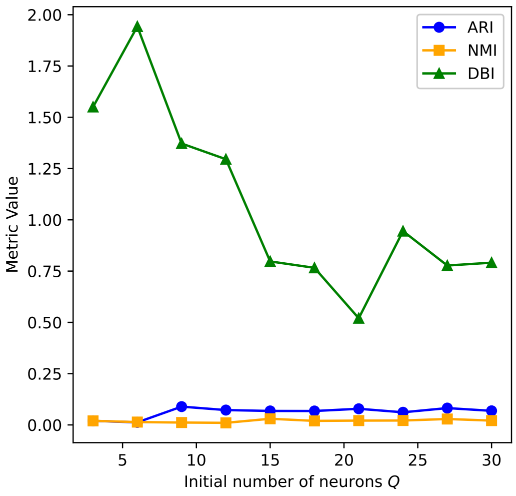

Figure 15: Clustering performance on HM dataset w.r.t. different values of Q

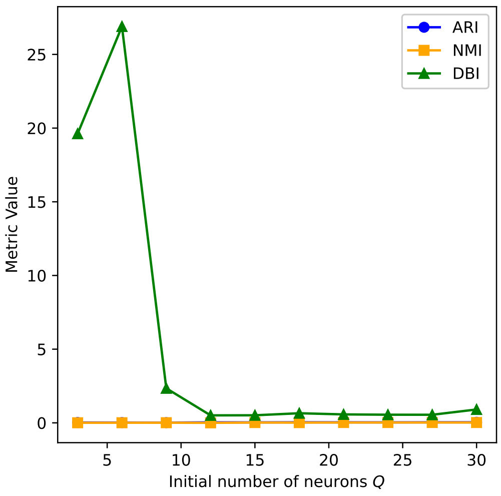

Figure 16: Clustering performance on HFCR dataset w.r.t. different values of Q

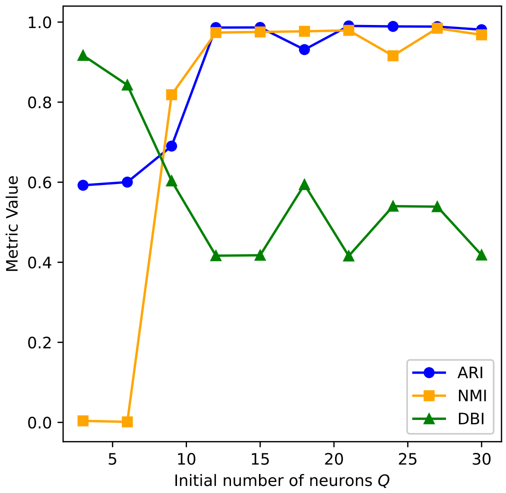

Figure 17: Clustering performance on IDS2 dataset w.r.t. different values of Q

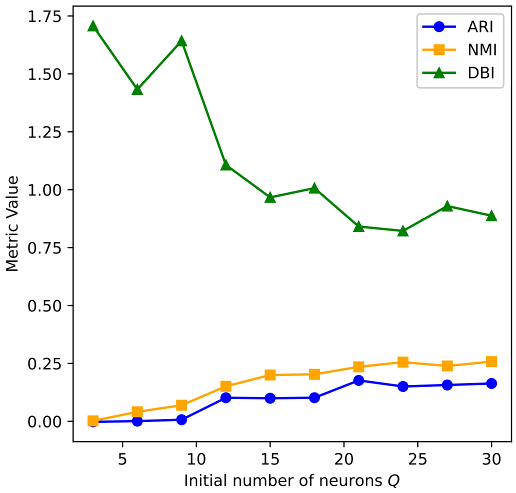

Figure 18: Clustering performance on MD dataset w.r.t. different values of Q

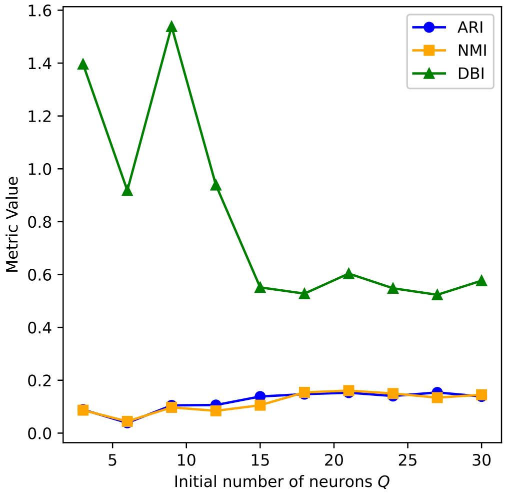

Figure 19: Clustering performance on PB dataset w.r.t. different values of Q

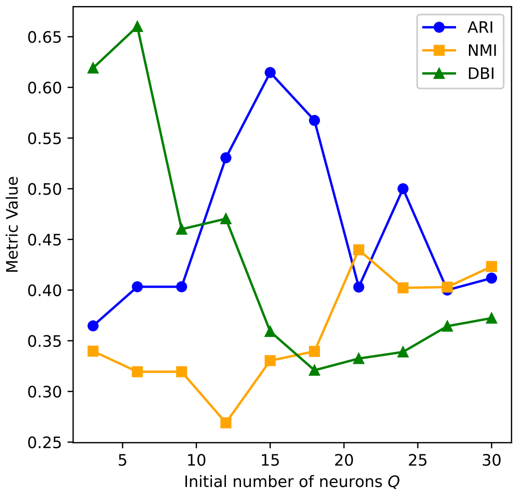

Figure 20: Clustering performance on RA dataset w.r.t. different values of Q

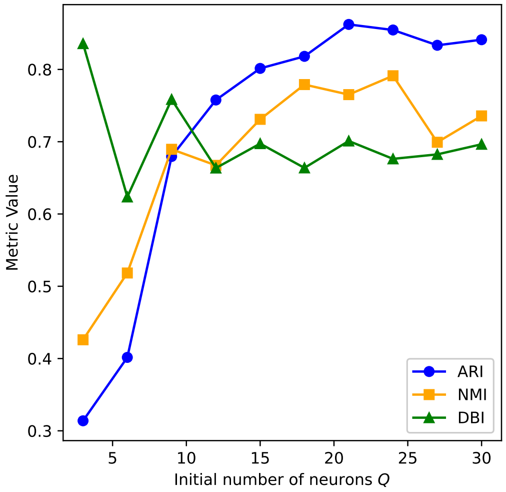

Figure 21: Clustering performance on SD dataset w.r.t. different values of Q

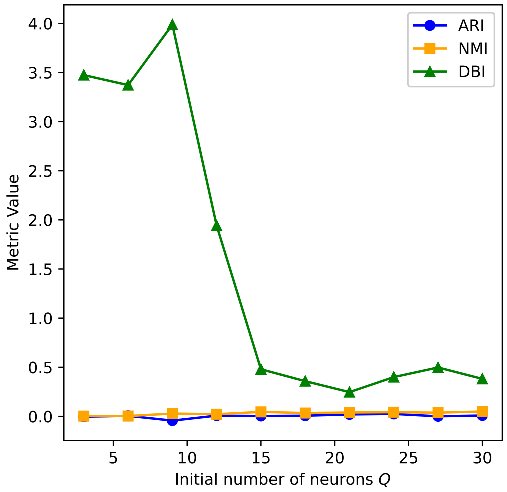

Figure 22: Clustering performance on WC dataset w.r.t. different values of Q