Additional Materials for “Clustering in the Presence of Concept Drift”

Richard Hugh Moulton, Herna L. Viktor, Nathalie Japkowicz and João Gama

1 Introduction

This document contains additional materials for the paper “Clustering in the Presence of Concept Drift,” which has been accepted for presentation at the 2018 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases as well as inclusion in the Springer LNCS conference proceedings. Formal citation for this publication will be provided when available.

2 Data Stream Clustering Algorithm Parameters

The MOA 17.06 [3] implementation of each data stream clustering algorithm (DSCA) was used, with the exceptions of ClusTree [6] and D-Stream [5]. Slightly modified versions of these algorithms were used; further details and links to the publicly available code are given in the respective sections.

We chose the parameters to be consistent across all experiments. In all cases, algorithms with a parameter k were given the number of classes at the beginning of the stream.

2.1 CluStream

The implementation CluStream with K-means [2] was used.

Symbol	Description	Value
h	time-horizon	100
k	the number of clusters to find	*
q	the maximum number of micro-clusters	100
t	parameter controlling the maximum boundary of micro-clusters	2

Table 1: Parameters for CluStream

2.2 ClusTree

A slightly modified version of ClusTree [6] was used to permit the proper generation of macroclusters.[1]

Symbol	Description	Value
h	horizon, which determines the decay rate λ	100
H	the maximal height of the tree	5
-	whether to implement a breadth-first strategy instead of a depth-first strategy	FALSE
-	which offline clustering algorithm to use	Silhouette k-means

Table 2: Parameters for ClusTree

2.3 D-Stream

A slightly modified version of D-Stream [5] was used to permit the user to specify the width of grid cells for numerical attributes.[2]

Symbol	Description	Value
λ	decay factor	0.998
C_m	parameter controlling the lower threshold for considering a grid cell to be dense	3
C_l	parameter controlling the upper threshold for considering a grid cell to be sparse	0.8
β	parameter controlling the length of time to wait before assessing whether a sparse grid cell is sporadic	0.3
-	Grid width (Synthetic real-valued data streams)	0.035
-	Grid width (ADFA-LD data streams)	0.5

Table 3: Parameters for D-Stream

[1] available https://doi.org/10.5281/zenodo.1216189

[2] available https://doi.org/10.5281/zenodo.1213802

2.4 DenStream

The implementation of DenStream [4] with DBSCAN was used.

Symbol	Description	Value
H	horizon	100
ϵ	the upper threshold for the radius of microclusters	0.065
β	determine the threshold of outlier relative to c-micro-clusters	0.2
μ	the lower threshold for the weight of microclusters	1
-	number of points to use for initialization	100
-	offline multiplier for ϵ	2
λ	decay factor	0.25
-	number of incoming points per time unit	10

Table 4: Parameters for DenStream

2.5 StreamKM++

StreamKM++ [1] was given the length of the data stream rounded up to the nearest 100, denoted by **.

Symbol	Description	Value
m	the size of the coreset	100
k	the number of clusters to find	*
n	the number of points in the data stream	**
-	random seed	1

Table 5: Parameters for StreamKM++

3 Additional Results

These are the results for cases that were not included in the main paper, for reasons of similarity with included results. Figures 1, 2 and 3 show the additional results for experiments A, B, and C respectively.

Figure 4 shows the additional results for the ADFA-LD data streams. The behaviour of the DSCAs for these data streams was considered very similar to their behaviour in the two reported cases.

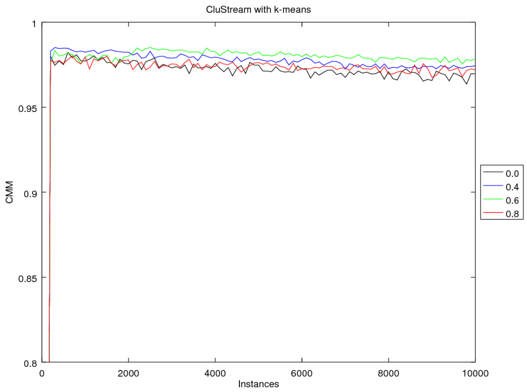
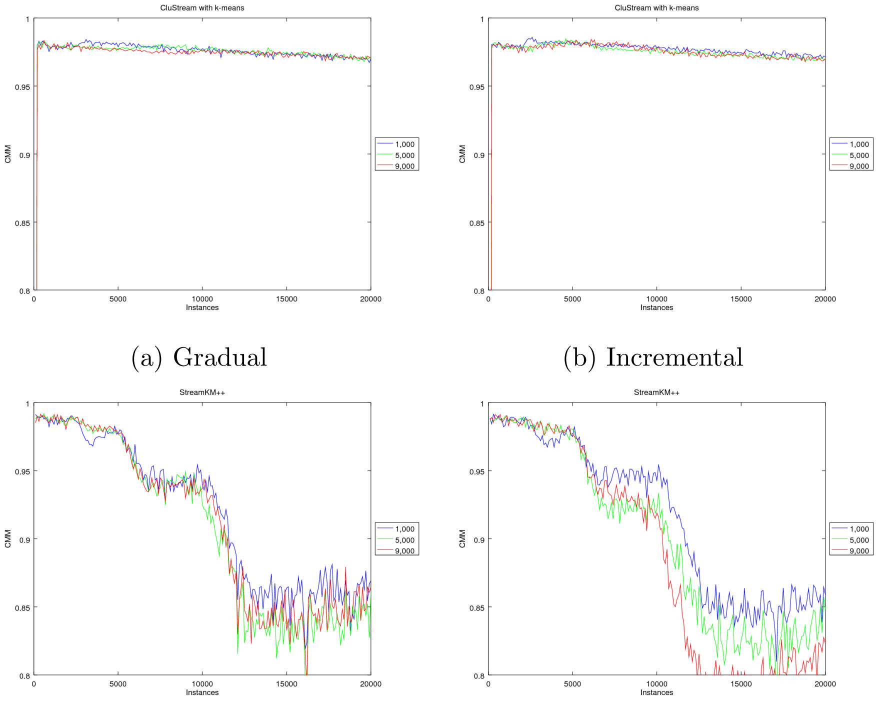


Figure 1: Additional results for Experiment A (Abrupt Concept Drift)



(a) Gradual

(b) Incremental

(c) Gradual

(d) Incremental

Figure 2: Additional results for Experiment B (Gradual and Incremental Concept Drift)

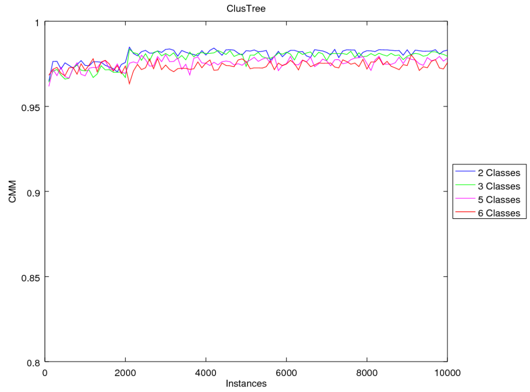


Figure 3: Additional results for Experiment C (Concept Evolution)

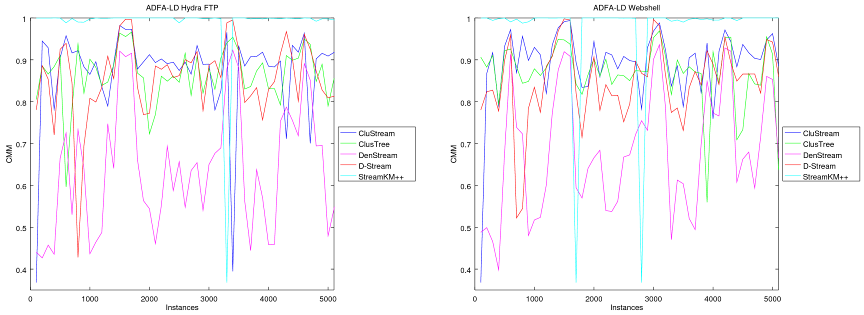


Figure 4: Additional results for the ADFA-LD data streams

References

[1] Marcel R. Ackermann, Marcus Märtens, Christoph Raupach, Kamil Swierkot, Christiane Lammersen, and Christian Sohler. StreamKM++: A Clustering Algorithm for Data Streams. *ACM Journal of Experimental Algorithmics*, 17(2):30, jul 2012.

[2] Charu C. Aggarwal, Jiawei Han, Jianyong Wang, and Philip S. Yu. A Framework for Clustering Evolving Data Streams. In *29th Very Large Database Conference*, page 12, Berlin, 2003.

[3] Albert Bifet, Geoff Holmes, Richard Kirkby, and Bernhard Pfahringer. Moa: Massive online analysis. *Journal of Machine Learning Research*, 11(May):1601–1604, 2010.

[4] Feng Cao, Martin Ester, Weining Qian, and Aoying Zhou. Density-Based Clustering over an Evolving Data Stream with Noise. In Joydeep Ghosh, Diane Lambert, David Skillicorn, and Jaideep Srivastava, editors, *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 328–339, Bethesda, 2006.

[5] Yixin Chen and Li Tu. Density-Based Clustering for Real-Time Stream Data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, San Jose, USA, 2007.

[6] Philipp Kranen, Ira Assent, Corinna Baldauf, and Thomas Seidl. The ClusTree: indexing micro-clusters for anytime stream mining. *Knowledge and Information Systems*, 29(2):249–272, nov 2011.