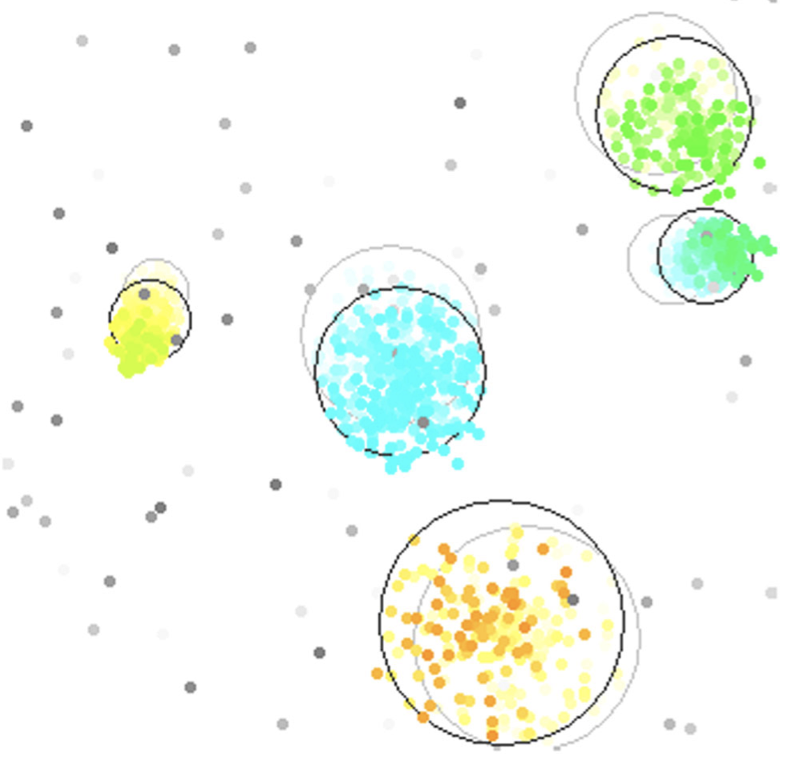
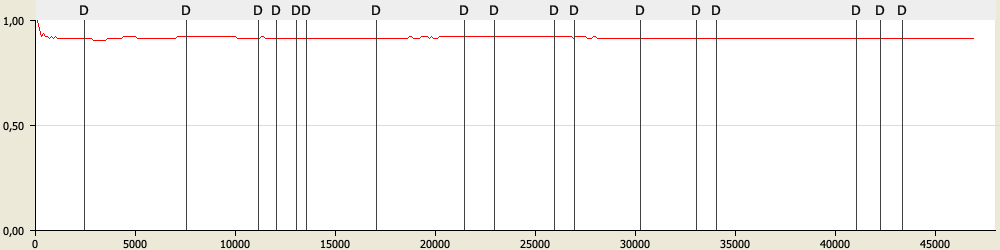
We used MOA, the most popular open source framework for data stream mining related to the WEKA project.

It includes a collection of machine learning algorithms for clustering and techniques for concept drift detection.



Using the Basic Concept Drift Performance Evaluator, we can see that 17 statistical changes are detected considering the stream of our 50.000 instances.



We decided to compare three different Streaming clustering algorithms in our analysis, withKmeans, StreamKM and BICO.

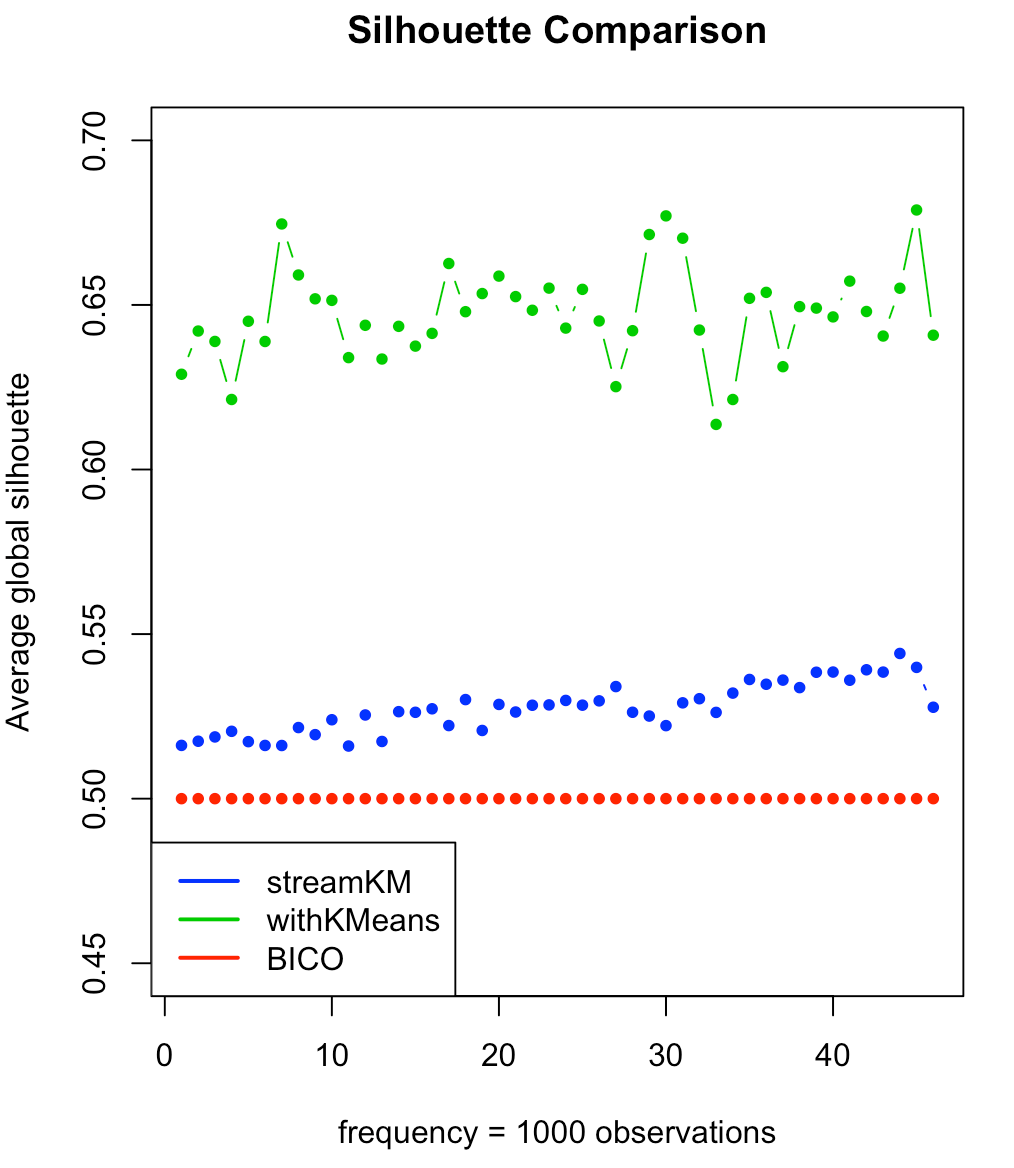
StreamKM++: It computes a small weighted sample of the data stream and it uses the k-means++ algorithm as a randomized seeding technique to choose the first values for the clusters. To compute the small sample, it employs coreset constructions using a coreset tree for speed up.

BICO (an acronym for “BIRCH meets coresets for k-means clustering”) is due to Fichtenberger et al. [101]. It combines the data structure of BIRCH with the theoretical concept of coresets for clustering. Coresets are explained in detail in section 9.6. BIRCH decides heuristically how to group the points into subclusters. The goal of BICO is to find a reduced set that is not only small, but also offers guarantees of approximating the original point set.

WIthKmeans: Determine the closest kernel, check whether instance fits into closestKernel. If the instance fits, it is put into the kernel otherwise it applies a free-memory strategy to insert a new kernel (trying to forget oldest kernels).

Then, it merges the two closest kernels,

Choosing StreamKM (sizeCoreset = 10.000, numClusters = 5, length = 100.000), WithKMeans (horizon = 1000, maNumKernels 100, kernelRadiFactor = 2, k = 5) and BICO (Cluster = 5, MaxClusterFeatures = 1000, Projections = 10), we obtain the following average global silhouette comparison with a frequency of 1000 observations.



From this comparison we can determine that withKMeans is the best streaming clustering algorithm.

Immagine che contiene testo

Descrizione generata automaticamente

Immagine che contiene testo

Descrizione generata automaticamente

Immagine che contiene testo

Descrizione generata automaticamente