Extending the stream Framework

Michael Hahsler  Matthew Bolaños
Southern Methodist University  Research Now

John Forrest
Microsoft Corporation

Abstract

This document describes how to add new data stream sources DSD and data stream
tasks DST to the stream framework.

Keywords: data streams, data mining, clustering.

1. Extending the stream framework

Since stream mining is a relatively young field and many advances are expected in the near
future, the object oriented framework in stream is developed with easy extensibility in mind.
Implementations for data streams (DSD) and data stream mining tasks (DST) can be easily
added by implementing a small number of core functions. The actual implementation can
be written in either R, Java, C/C++ or any other programming language which can be in-
terfaced by R. In the following we discuss how to extend stream with new DSD and DST
implementations.

1.1. Adding a new data stream source (DSD)

DSD objects can be a management layer on top of a real data stream, a wrapper for data
stored in memory or on disk, or a generator which simulates a data stream with know prop-
eties for controlled experiments. Figure 1 shows the relationship (inheritance hierarchy) of
the DSD classes as a UML class diagram (Fowler 2003). All DSD classes extend the abstract
base class DSD. There are currently two types of DSD implementations, classes which imple-
ment R-based data streams (DSD_R) and MOA-based stream generators (DSD_MOA) provided
in streamMOA. Note that abstract classes define interfaces and only implement common
functionality. Only implementation classes can be used to create objects (instances). This
mechanism is not enforced by S3, but is implemented in stream by providing for all abstract
classes constructor functions which create an error.

The class hierarchy in Figure 1 is implemented using the S3 class system (Chambers and Hastie
1992). Class membership and the inheritance hierarchy is represented by a vector of class
names stored as the object’s class attribute. For example, an object of class DSD_Gaussians
will have the class attribute vector c("DSD_Gaussians", "DSD_R", "DSD") indicating that
the object is an R implementation of DSD. This allows the framework to implement all
common functionality as functions at the level of \texttt{DSD} and \texttt{DSD\_R} and only a minimal set of functions is required to implement a new data stream source. Note that the class attribute has to contain a vector of all parent classes in the class diagram in bottom-up order.

For a new DSD implementation only the following two functions need to be implemented:

1. A creator function (with a name starting with the prefix \texttt{DSD\_}) and
2. the \texttt{get\_points()} method.

The creator function creates an object of the appropriate DSD subclass. Typically this S3 object contains a list of all parameters, an open R connection and/or an environment or a reference class for storing state information (e.g., the current position in the stream). Standard parameters are \texttt{d} and \texttt{k} for the number of dimensions of the created data and the true number of clusters, respectively. In addition an element called "\texttt{description}" should be provided. This element is used by \texttt{print()}. The implemented \texttt{get\_points()} needs to dispatch for the class and create as the output a data frame containing the new data points as rows. Also, if the ground truth (true cluster assignment as an integer vector; noise is represented by \texttt{NA}) is available, then this can be attached to the data frame as an attribute called "\texttt{assignment}".

For a very simple example, we show here the implementation of \texttt{DSD\_UniformNoise} available in the package’s source code in file \texttt{DSD\_UniformNoise.R}. This generator creates noise points uniformly distributed in a \texttt{d}-dimensional hypercube with a given range.

\begin{verbatim}
R> library("stream")

R> DSD\_UniformNoise <- function(d = 2, range = NULL) {
+   if(is.null(range)) range <- matrix(c(0, 1), ncol = 2, nrow = d,
+       byrow = TRUE)
+   structure(list(description = "Uniform Noise Data Stream", d = d,
+       k = NA\_integer\_., range = range),
+       class = c("DSD\_UniformNoise", "DSD\_R", "DSD"))
+ }
R> get\_points.DSD\_UniformNoise <- function(x, n = 1,

\end{verbatim}
The constructor only stores the description, the dimensionality and the range of the data. For this data generator \( k \), the number of true clusters, is not applicable. Since all data is random, there is also no need to store a state. The \texttt{get_points()} implementation creates \( n \) random points and if assignments are needed attaches a vector with the appropriate number of NA\$s indicating that the data points are all noise.

Now the new stream type can already be used.

\begin{verbatim}
R> stream <- DSD_UniformNoise()
R> stream

Uniform Noise Data Stream
Class: DSD_UniformNoise, DSD_R, DSD
With NA clusters in 2 dimensions

R> plot(stream, main = description(stream))
\end{verbatim}

The resulting plot is shown in Figure 2.

Several more complicated examples are available in the package's source code directory in files starting with \texttt{DSD_}.
1.2. Adding a new data stream tasks (DST)

DST refers to any data mining task that can be applied to data streams. The design is flexible enough for future extensions including even currently unknown tasks. Figure 3 shows the class hierarchy for DST. It is important to note that the DST base class is shown merely for conceptual purpose and is not directly visible in the code. The reason is that the actual implementations of data stream operators (DSO), clustering (DSC), classification (DSClass) or frequent pattern mining (DSFPM) are typically quite different and the benefit of sharing methods would be minimal.

DST classes implement mutable objects which can be changed without creating a copy. This is more efficient, since otherwise a new copy of all data structures used by the algorithm would be created for processing each data point. Mutable objects can be implemented in R using environments or the recently introduced reference class construct (see package methods by the R Core Team (2014)). Alternatively, pointers to external data structures in Java or C/C++ can be used to create mutable objects.

To add a new data stream mining tasks (e.g., frequent pattern mining), a new package with a subclass hierarchy similar to the hierarchy in Figure 3 for data stream clustering (DSC) can be easily added. This new package can take full advantage of the already existing infrastructure in stream. An example is the package streamMOA Hahsler and Bolanos (2015), which can be used as a model to develop a new package. We plan to provide more add-on packages to stream for frequent pattern mining and data stream classification in the near future.

In the following we discuss how to interface an existing algorithm with stream. We concentrate again on clustering, but interfacing algorithms for other types of tasks is similar. To interface
an existing clustering algorithm with stream,

1. a creator function (typically named after the algorithm and starting with DSC_) which created the clustering object,

2. an implementation of the actual cluster algorithm, and

3. accessors for the clustering

are needed. The implementation depends on the interface that is used. Currently an R interface is available as DSC_R and a MOA interface is implemented in DSC_MOA (in streamMOA). The implementation for DSC_MOA takes care of all MOA-based clustering algorithms and we will concentrate here on the R interface.

For the R interface, the clustering class needs to contain the elements "description" and "RObj". The description needs to contain a character string describing the algorithm. RObj is expected to be a reference class object and contain the following methods:

1. cluster(newdata, ...), where newdata is a data frame with new data points.

2. For micro-clusters: get_microclusters(...) and get_microweights(...)

3. For macro-clusters: get_macroclusters(...), get_macroweights and microToMacro(micro, ...) which does micro- to macro-cluster matching.

Note that these are methods for reference classes and do not contain the called object in the parameter list. Neither of these methods are called directly by the user. Figure ?? (on page ??) shows that the function update() is used to cluster data points, and get_centers() and get_weights() are used to obtain the clustering. These user facing functions call internally the methods in RObj via the R interface in class DSC_R.

For a comprehensive example of a clustering algorithm implemented in R, we refer the reader to DSC_DStream (in file DSC_DStream.R) in the package’s R directory.

References


Affiliation:
Michael Hahsler
Engineering Management, Information, and Systems
Lyle School of Engineering
Southern Methodist University
P.O. Box 750122
Dallas, TX 75275-0122
E-mail: mhahsler@lyle.smu.edu
URL: http://lyle.smu.edu/~mhahsler
Matthew Bolaños
Research Now
5800 Tennyson Pkwy # 600
Plano, TX 75024 E-mail: mbolanos@curiouscrane.com
John Forrest
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-7329
E-mail: jforrest@microsoft.com