FACIL – An Approach for Classifying Data Streams by Decision Rules and Border Examples

Francisco J. Ferrer–Troyano, Jesús S. Aguilar–Ruiz, and José C. Riquelme

Department of Computer Science, University of Seville
Avenida Reina Mercedes s/n, 41012 Seville, Spain
{ferrer, riquelme}@lsi.us.es

Abstract. This paper describes FACIL, a classifier based on decision rules and border examples that avoids unnecessary revisions when virtual drifts are present in data. Rules in FACIL are both pure - consistent - and impure - inconsistent -. Pure rules classify new test examples by covering and impure rules classify them by distance as the nearest neighbor algorithm. In addition, the system provides an implicit forgetting heuristic so that positive and negative examples are removed from a rule when they are not near one another.

1 Introduction

Formally, a data stream is an ordered sequence of data items $< \ldots e_{i-1} \prec e_i \succ e_{i+1} \ldots >$ read in increasing order of the indices i. In practice, a data stream is an unbounded sequence of items liable to both noise and concept drift, and received at a so high rate that each one can be read at most once by a real time application [2]. Thus, data streams contexts compel to learning systems to give approximate answers using small and constant time per example [3]. Recent works on data streams classification has been mainly addressed by two different approaches: decision trees [1, 3, 4] and ensemble methods [5, 8, 9].

Domingos & Hulten's VFDT and CVFDT systems [3] build a decision tree based on Hoeffding bounds, which guarantee constant time and memory per example and an output model asymptotically nearly identical to that given by a batch conventional learner from enough examples. Since VFDT and CVFDT are evaluated for data streams with symbolic attributes, Jin & Agrawal propose in [4] a numerical interval pruning approach to reduce the processing time for numerical attributes, without loss in accuracy. Gama et al.'s VFDTc system [1] extends the VFDT properties in two directions: the ability to deal with numerical attributes and the ability to apply nave Bayes classifiers in tree leaves.

Ensemble batch learning algorithms such as Boosting and Bagging have proven to be highly effective from disk–resident data sets. These techniques perform repeated resampling of the training set, making them a priori inappropriate in a data streams environment. Despite what might be expected, novel ensemble methods are increasingly gaining attention because of they have proved to offer an improvement in prediction accuracy. In general, every incremental ensemble approach uses some criteria to dynamically delete, reactivate, or create