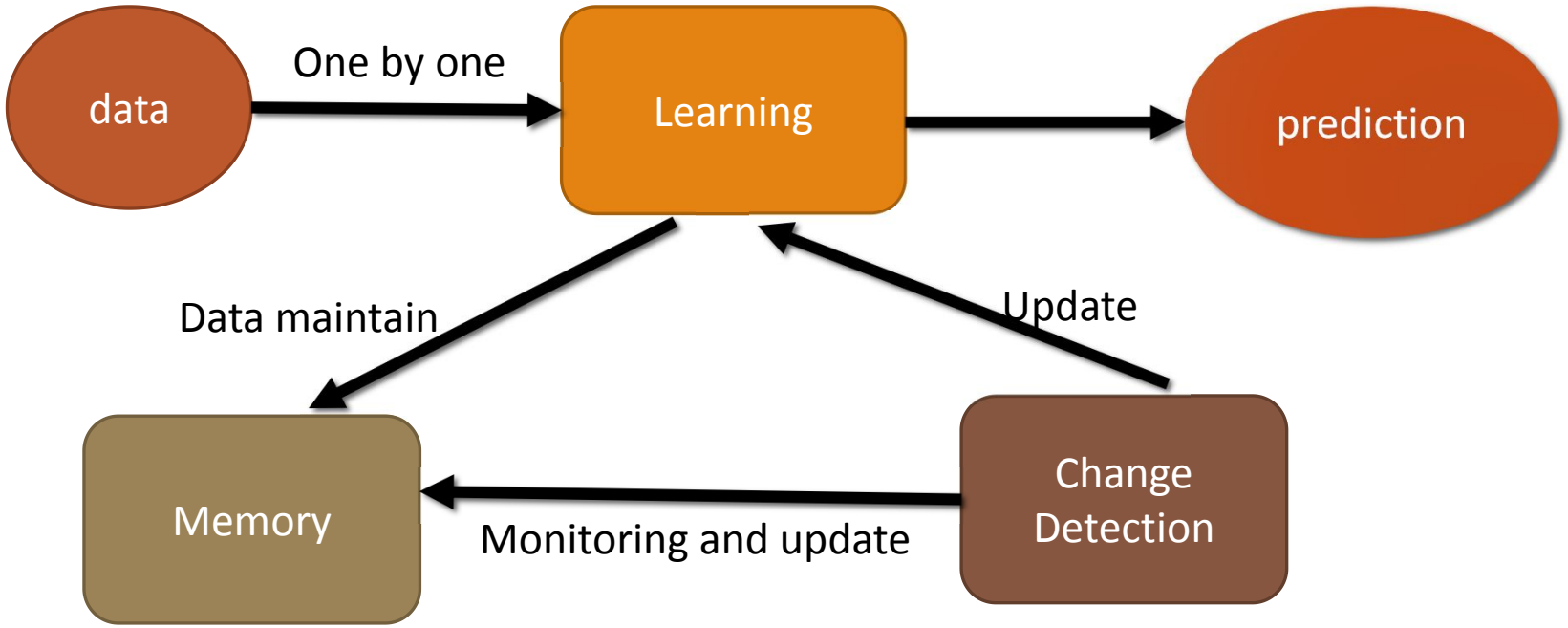
Data stream Learning

Outline

1. Framework of data stream learning
2. Memory
3. Change Detection
4. Learning

Framework of data stream Learning



Memory

Key problem is how to **maintain** and **forget** some of data.

a. Single example
b. sliding window(decay function)
c. Adaptive Filter data

Generally speaking, there is a popular assumption that more recent data, more representative.

Online SVM

The object function of svm is

$$\min(\frac{\lambda}{2}||w||^2 + \sum_{i=1}^{n} \max(1 - y_i(w * x_i + b), 0))$$

if $1 - y_i(w * x_i + b) > 0$ **update** w and b accroding following foruma(SGD):

$$w = (1 - \lambda)w + \eta y_i x_i,\ b = b + y_i$$

Single example

However, this kind of method do not have explicit forgetting mechanisms. Adaptation happens only as the old concepts are diluted due to the new incoming data.so they can’t detect abrupt concept drift in time.

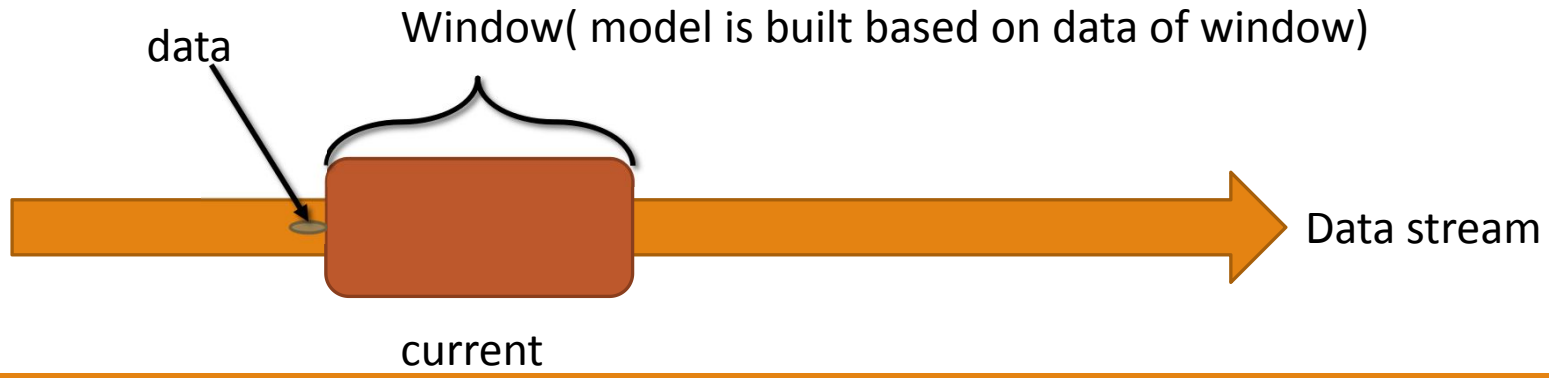
Window based

maintain a predictive model consistent with a set of recent examples.

The model is updated following two processes:

1. a learning process (update the model based on the new data)
2. a forgetting process(discard data that is moving out of the window)

The key challenge is to select an appropriate window size.



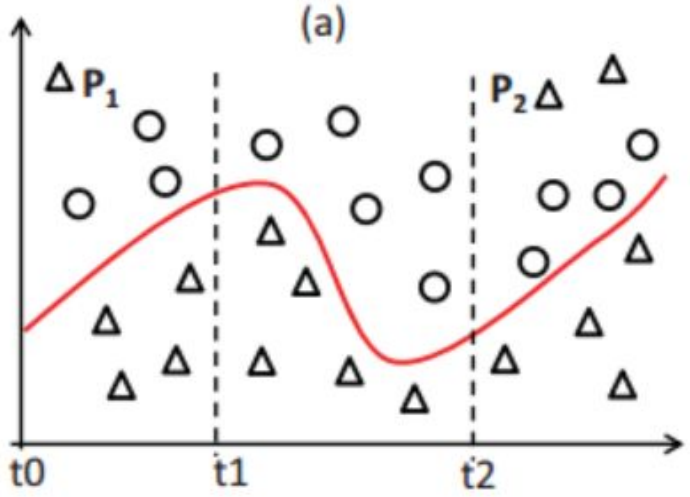
Adaptive Filter data

How to maintain data which is representative for current concept? Lazy learning is a good choice.

1. syncStream
2. Instance based learning stream（IBLStream）

Change Detection

The change detection component refers to the techniques and mechanisms for explicit drift and change detection. It characterizes and quantifies concept drift by identifying change points or small time intervals during which changes occur.



DDM

The theory guarantees that while the class distribution of the examples is **stationary** , the error rate of the learning algorithm will **decrease** when i increases**. A significant increase in the error of the algorithm**, suggest a change in the class distribution, and whether is a significant increase is based on following formula.

$$p_i + s_i \geq p_{\min} + 3 * s_{\min}$$

the error-rate is the probability of observe False, p_i
, with standard deviation given by $s_i = sqrt(\mathrm{p}_i(1 - p_i) / \mathrm{i})$

Adaptive Windowing（ADWIN）

The idea is simple: whenever two “large enough” subwindows of W exhibit “distinct enough” averages, one can conclude that the corresponding expected values are different,and the older portion of the window is dropped.

begin
  Initialize Window W;
  foreach $(t) > 0$ **do**
    $W \leftarrow W \cup \{x_t\}$ (i.e., add x_t to the head of W);
    repeat
      Drop elements from W
    until $|\hat{\mu}_{W_0} - \hat{\mu}_{W_1}| < \epsilon_{cut}$ *holds for every split of* W *into* $W = W_0 \cup W_1$;
  end
  Output $\hat{\mu}_W$
end

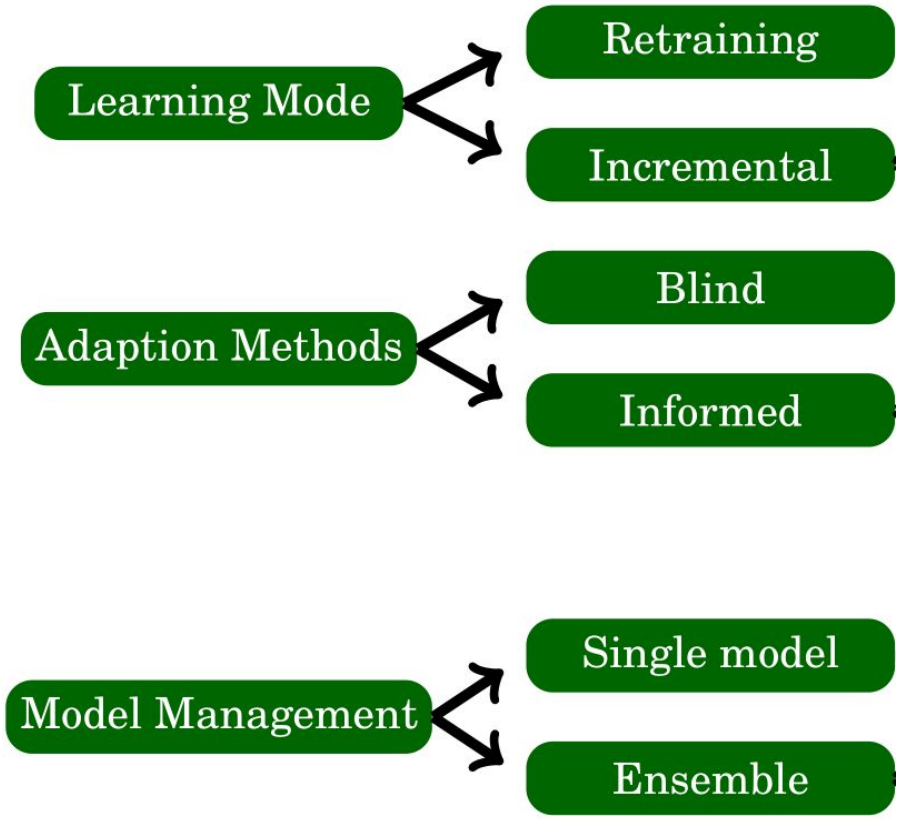
Denstream

It maintains p-micro-clusters and o-micro-clusters which include sufficient information for clustering and use weight and decay function change p-micro-clusters and o-micro-clusters dynamically.

Classification vs cluster on data stream

What is difference between classification and cluster on about dynamic characteristic data stream ?

Learning


Learning Mode
Retraining
Incremental
Adaption Methods
Blind
Informed
Model Management
Single model
Ensemble

VFDT

Just a incremental Decision Tree, but the key motivation is when to split to children node. The simplest idea is the more data, the better.

Our goal is to ensure that, with high probability, the attribute chosen using n examples is the same that would be chosen using infinite example.

VFDT

$$\Delta\overline{G} = \overline{G}(X_a) - \overline{G}(X_b) \geq 0$$

$G(X_i)$ be the heuristic measure sued to choose test attributes (CART ,Gini)

Hoeffding bound !!!!!!!!!!!!!!!!!!

It guarantees that is the correct choice with probability 1 – δ if n examples have been seen at this node and $\Delta\overline{G} > \varepsilon^2$

CVFDT

CVFDT maintains a window of training examples; whenever a new example is read it is added to the statistics at all the nodes in the tree that it passes through, the last example in the window is forgotten from every node where it had previously had an effect, and the validity of all statistical tests are checked.

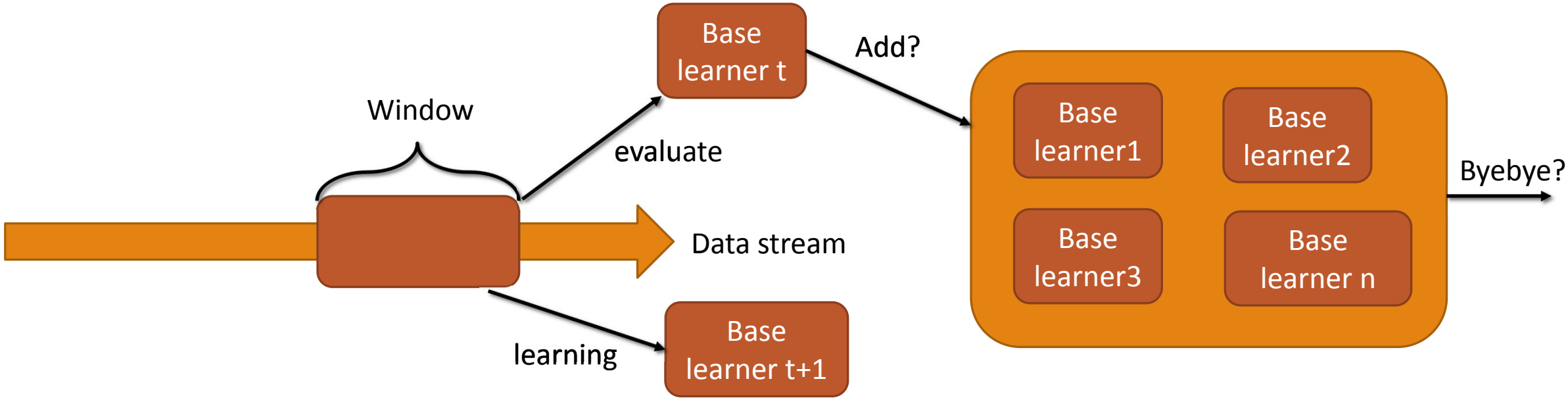
Ensemble classifier for data stream

Ensemble learning for data stream maintains in memory an ensemble of multiple models that make a combined prediction. There is three topic about ensemble classifier for data stream.

1. How do we get base learner trained
2. How combine these base learners
3. How are new learners added and inefficient ones removed

Streaming ensemble algorithm(SEA)

It trains a separate classifier on each sequential batch of training examples. A trained classifier is added to a fixed-size ensemble, while the worst performing classifier is discarded. The final prediction is made using a simple majority voting.



Dynamic Weighted Majority(DWM)

Maintains as its concept description an ensemble of learning algorithms, each referred to as an expert and each **with an associated weight**. Given an instance, the performance element polls the experts, each returning a prediction for the instance. Using these predictions and expert weights, DWM returns as the global prediction the class label with the highest accumulated weight.