Abstract

The objective of this work is to develop a predictive model when data batches are collected in a sequential manner. With streaming data, information is constantly being updated and a major statistical challenge for these types of data is that the underlying distribution and the true input-output dependency might change over time, a phenomenon known as concept drift. The concept drift phenomenon makes the learning process complicated because a predictive model constructed on the past data is no longer consistent with new examples. In order to effectively track concept drift, we propose new novel model-combining methods using constrained and penalized regression that possesses a grouping property.

The new learning methods enable us to select data batches as a group that are relevant to the current one, reduce the effects of irrelevant batches, and adaptively reflect the degree of concept drift emerging in data streams. We study theoretical properties of the proposed methods and finite sample performance using simulated and real examples. The analytical and empirical results indicate that the proposed methods can effectively adapt to various types of concept drift and show superior performance over existing methods.

Concept Drift

- Consider a sequential regression problem in a data stream which is a series of data batches entering over time continuously.
- Data batch has observations, i.e. input-output pairs, and is used for constructing a learning model for predicting the future observations.
- Main concern: concept drift
- Any changes of circumstance in a learning process
- Examples: subprime mortgage crisis, spam classification, and Airplane Delay

Figure: Abrupt changes in airplane delay minutes

Setting

- Let \( \{D_n, \ n = 1, 2, \ldots, M\} \) be a sequence of data batches which consists of input-output pairs.
- Suppose observations in \( D_n \) are random samples from unknown distributions \( F_n(x, y) \).
- \( x \in \mathbb{R}^p \) is a predictor vector.
- \( y \in \mathbb{R} \) is a response.
- \( x \) and \( y \) arise from a statistical model \( y = f(x) + \epsilon \) where \( \epsilon \) is a random error satisfying \( E(\epsilon) = 0 \) and is independent of \( x \).
- The underlying distribution can be different for each batch.
- The relationship between response and predictors can change over time.
- Prediction obtained from the outputs of base models and the most recent data batch \( D_M = \{(x_1, y_1), \ldots, (x_n, y_n)\} \) that best reflects the target concept to be tracked.
- For each \( x_i, i = 1, \ldots, n \), denote the outputs of base models by \( f_i(x) = (f_i(x_1), \ldots, f_i(x_n), \hat{f}_i(x)) \).
- \( \hat{f}_i(x) \) represents a leave-one-out estimate of \( f_i(x) \) to avoid over-fitting.
- Data batches \( D_1, \ldots, D_M \).

Input-output pairs

\[
\begin{array}{c|c|c}
\text{Base Models} & D_1 & D_M \\
\hline
f_1(x) & & \\
\vdots & & \\
\hat{f}_i(x) & & \\
\hline
\end{array}
\]

Base Models

\[
\begin{array}{c|c|c}
f_1(x) & \hat{f}_1(x) & \hat{f}_M(x) \\
\hline
f_1(x) & \hat{f}_1(x) & \hat{f}_M(x) \\
\vdots & \vdots & \vdots \\
\hat{f}_i(x) & \hat{f}_i(x) & \hat{f}_i(x) \\
\hline
\end{array}
\]

Important Result

The proposed combiner holds grouping property by assigning the same weights to identical concepts and produces better prediction when there is concept drift by assigning higher weights on relevant batches.

Definition (1)

(A measure of concept drift) For \( w^* = (1/M, \ldots, 1/M)^T \) and \( w \) obtained by a combiner, we define the degree of concept drift as the angle between the two vectors given by

\[
\eta(w^*, w) = \cos^{-1}\left(\frac{\langle w^*, w \rangle}{\|w^*\|\|w\|}\right)
\]

where \( \langle \cdot, \cdot \rangle \) denotes the inner product of two vectors in \( \mathbb{R}^M \).
- \( \eta(w^*, w) \) is small when there is no or gradual concept drift.
- \( \eta(w^*, w) \) gets larger when there is considerable concept drift.

Theorem (1)

When concept drift is not present, the minimizer of \( \eta \) is given as \( w = (1/M, \ldots, 1/M)^T \) for sufficiently large \( \lambda \).

Theorem (2)

For \( w \) obtained from \( \eta \), \( \eta(w^*, w) \) in Definition 1 is monotonically decreasing as \( \lambda \) increases.

Methodology

- the design matrix and the response for the model aggregation

\[
X = \begin{bmatrix}
\hat{f}_1(x_1) & \hat{f}_2(x_1) & \cdots & \hat{f}_M(x_1) \\
\hat{f}_1(x_2) & \hat{f}_2(x_2) & \cdots & \hat{f}_M(x_2) \\
\vdots & \vdots & \ddots & \vdots \\
\hat{f}_1(x_n) & \hat{f}_2(x_n) & \cdots & \hat{f}_M(x_n)
\end{bmatrix}, \quad y = \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix}
\]

- Final predictive model has the form of \( f(x) = \sum_{i=1}^{M} w_i \hat{f}_i(x) \).
- Goal: optimally determine the weights \( w = (w_1, \ldots, w_M) \) using \( f(X) \).
- Proposed method: A constrained penalized aggregation method minimizes the following objective function with respect to \( w = (w_1, \ldots, w_M) \):

\[
R^D(w) = \frac{1}{n} \sum_{i=1}^{n} \left[ y_i - \sum_{i=1}^{M} w_i \hat{f}_i(x_i) \right]^2 + \lambda \|Dw\|_1^2,
\]

such that \( \sum_{i=1}^{M} w_i = 1 \) where \( \lambda \geq 0 \).

Weights

\[
\begin{array}{c|c|c|c|c|c|c|c|c|c}
\text{Weight Value} & w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 \\
\hline
\text{Weight} & 0.0 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1.0 \\
\hline
\end{array}
\]

Simulation

- 20 training sequential data batches \( D_1, \ldots, D_M \).
- Each \( D_i \) contains 200 observations independently from \( y = \sum_{i=1}^{M} \alpha_i X_i + \epsilon \) is a response variable. \( (X_1, \ldots, X_M) \) are independent predictors from Uniform(0,1). The noise variable \( \epsilon \) is independently from standard normal distribution.
- The concept drift is incorporated by changing the value of \( \alpha = (\alpha_1, \ldots, \alpha_M) \) from a certain time \( T = 0, 3, 5, 7, 9, 11, 13, 15, 17, 19 \).
- For example, if \( T = 10 \), each \( D_i \) for \( i = 1, \ldots, 9 \) is generated with \( \alpha = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5) \) and the remaining data batches are with \( \alpha = (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0) \).
- Test data batch: same data generation process as the last batch \( D_M \) with \( n = 1000 \) independent observations.
- Tuning parameter \( \lambda \) is selected by 10-fold cross-validation from \( \log_{10}(0.5 \times \lambda) = \epsilon \) where \( \epsilon \) varies from \(-3 \) to \( 3 \) by \( 0.2 \) selected based on the lowest RMSE.
- Compared with methods:
  - MC.WAvg: Weighted average method (Wang et al. 2003)
  - MC.Rice1: Ridge regression method (Yeong et al. 2010)

Simulation Results

(a) Angle

(b) RMSE

Figure: The left panel (a) shows the angles with standard errors for four methods at each position of concept drift and the right panel. The angle increases as the concept drift point gets closer to current time point (b) shows the RMSE at each position of concept drift.

References