Improved terrestrial GPP estimation using Multisource Data

Hemant Banke (MD2107)

Under the guidance of Dr. B. Uma Shankar

Statistics and Mathematics unit
Indian Statistical Institute, Kolkata
West Bengal - 700 108, India

March 22, 2025

Overview

1. **Drifter Algorithm**

2. **Direct Loss Estimation**

Setup

- Training Data : $D_{tr} = \{(i, x_i, y_i)\}_{i=1}^{n_{tr}}$
- Test Data : $D_{te} = \{(i, x'_i, y'_i)\}_{i=1}^{n_{te}}$
- Segments of the data are defined by tuples $s = (a, b)$ where a and b are the endpoints of the segment such that $a \leq b$. $D_{|s} = \{(i, x'_i, y'_i) | a \leq i \leq b\}$
- For a regression function $f : R^m \rightarrow R$ trained using D_{tr} , The generalization error of f on the data set $D = \{(i, x'_i, y'_i)\}_{i=1}^{n}$ is defined as

$$RMSE(f, D) = \Big(\sum_{i=1}^{n} [f(x'_i) - y'_i]^2 / n\Big)^{1/2}$$

Main Problem discussed in Paper

- Given a regression function f trained using the dataset D_{tr}, and a threshold σ, predict whether the generalization error E of f on the testing data D satisfies $E \geq \sigma$ when only the true dependent variable $y'_i, i \in [n]$, is unknown.

Main Idea

- Virtual Concept Drift is change in the distribution of the covariates $p(x)$. Without the ground truth we can only detect virtual concept drift.
- We need a distance measure $d(x)$ that measures how “far” a vector x is from the data D_{tr} that was used to train the regressor. Small values of d(x) mean that we are close to the training data and the regressor function should be reliable, while a large value of d(x) means that we have moved away from the training data, after which the regression estimate may be inaccurate.
- $d(x)$: Concept drift indicator variable

Main Idea

Train different regression functions, say f and f', on different subsets of the training data. We then define the distance measure to be the difference between the predictions of these two functions.

$$d(x) = [f(x) - f'(x)]^2$$

This distance measure has the suitable property that if some attributes are independent of the dependent variable, then they will not affect the behavior of the regression functions and, hence, the distance measure d is not sensitive to them.

Proof for OLS model

Consider a linear model with usual notations,

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

$\beta \in R^m$, $X_{n \times m}$ is the data matrix and ϵ_i are independent random variables with zero mean and variance of σ_y^2.
Consider 2 OLS models trained :

- $\hat{y} = f(x) = \hat{\beta}^T x$ trained on dataset of size n
- $\hat{y'} = f'(x) = \hat{\beta'}^T x$ trained on independently sampled dataset of size n'.

Proof for OLS model

Theorem The expected mean squared error $E[(f(x){-}y)^2]$ is monotonically related to the expectation of the squared difference between the two regressors f and f', i.e., $E[(f(x){-}f'(x))^2]$, by the following equation to a leading order in n^{-1} and n'^{-1}:

$$E[(f(x){-}y)^2] = (1 + n/n')^{-1} E[(f(x){-}f'(x))^2] + \sigma_y^2$$

Proof for OLS model

Proof

WLG we can assume that the distribution from which the covariates have been sampled has been centered so that all terms except the intercept have an expectation of zero, or $x_{i1} = 1$ and $E[x_{ij}] = 0$ for all $j \neq 1$. We can further assume that axes of covariates have been rotated so that they are uncorrelated and satisfy $E[x_{ij}x_{ik}] = \sigma^2_{jk}\delta_{jk}$, where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise.

For a dataset of size n, the OLS estimate of β, denoted by $\hat{\beta}$, is a random variable that obeys a distribution with a mean of β and a covariance given by $n^{-1}\Sigma$

$$Var(\hat{\beta}) = \sigma^2_y(X^TX)^{-1} = \sigma^2_y(n\, diag\{1, \sigma^2_{22}, ..., \sigma^2_{mm}\})^{-1} = n^{-1}\Sigma$$

where $\Sigma = \sigma^2_y diag\{1, \sigma^{-2}_{22}, ..., \sigma^{-2}_{mm}\}$

Proof for OLS model

Hence,

$$E[(f(x)-y)^2] = Var(x^T\hat{\beta}-y) = x^T(n^{-1}\Sigma)x + \sigma_y^2$$

and

$$E[(f(x)-f'(x))^2] = Var(x^T\hat{\beta} - x^T\hat{\beta}') = x^T(n^{-1}\Sigma)x + x^T(n'^{-1}\Sigma)x = x^T(n^{-1} + n'^{-1}\Sigma)x$$

Replacing $x^T\Sigma x$ in first equation we get ,

$$E[(f(x)-y)^2] = (1 + n/n')^{-1}E[(f(x)-f'(x))^2] + \sigma_y^2$$

Hence proved.

Our claim is therefore that the difference between the estimates of regressors trained on different subsets of the data in the point x defines a distance function which can be evaluated even when the ground truth is unknown. If a data point x is close to the data points used to train the regressors the distance should be small. On the other hand, if the data point is far away from the data used to train the regressors, the predictions of the regressors diverge and the distance and also the prediction error will be larger.

Drifter Algorithm

Training phase

- Train the main model f on D_{tr}.
- Consider l subsequences $(s_1, s_2, ..., s_l)$ in $[n_{tr}]$. These are called ”segments” of training data. The segments are chosen such that the data in each segment corresponds to only one ”concept”. Segments can be overlapping and of different lengths to make the process more robust. Segmentation consisting of equally-sized segments of length l_{tr} with 50% overlap is quite robust.
- Train l segment models, i.e. model f_i is trained on $D_{|s_i}$, $i = 1(1)l$. The segment models be of different/weaker form than the main model f.

Drifter Algorithm

Testing phase

- For each of the segment model, find RMSE on D_{te}

$$z_i = RMSE(f, f_i, D_{te}) = \Big(\sum_{j=1}^{n_{te}} [f(x_j') - f_i(x_j')]^2 / n_{te}\Big)^{1/2}$$

- This gives us l values z_i estimating the generalization error, and we then choose the n_{ind} 'th smallest value as the value for the concept drift indicator variable d.
- For overlapping segmentation scheme, $n_{ind} = 2$ can be used as it is reasonable to assume that at least two of the segment models should have small values for z_i's, if the testing data has no concept drift, while a single small value for z_i could still occur by chance.

Drifter Algorithm

Drift detection threshold

- δ : threshold for the concept drift indicator variable d that estimates the threshold σ for the generalization error.
- Split the training dataset into $[n_{tr}/n_{te}]$ (non-overlapping) segments of the same length as the testing data.
- Compute the concept drift indicator value d_i for each segment s_i in the training data $D_{tr|s_i}$.
- Choose a concept drift detection threshold δ by,

$$\delta = mean(d_i) + c \times sd(d_i)$$

where c is constant multiplier.

Direct Loss Estimation

Data is divided into 3 categories :

- Training Set : (X_{train}, Y_{train})
- Reference Set : (X_{ref}, Y_{ref})
- Testing/Analysis Set : (X_{test})

Two models are trained : **Child** and **Nanny** models

Direct Loss Estimation

Training Phase

- Train the Child model on Training set

$$X_{train} \rightarrow Child \rightarrow Y_{train}$$

- Predict the Child model on Reference set

$$X_{ref} \rightarrow Child \rightarrow \hat{Y}_{ref}$$

- Train the Nanny model as

$$(X_{ref}, \hat{Y}_{ref} \rightarrow Nanny \rightarrow Loss(Y_{ref}, \hat{Y}_{ref})$$

Two models are trained : **Child** and **Nanny** models

Direct Loss Estimation

Testing Phase

- Predict the Child model on Testing set

$$X_{test} \rightarrow Child \rightarrow \hat{Y_{test}}$$

- To get Loss estimate, predict Nanny model as

$$(X_{test}, \hat{Y_{test}} \rightarrow Nanny \rightarrow \hat{Loss}(Y_{test}, \hat{Y_{test}})$$

Two models are trained : **Child** and **Nanny** models

The End