Time Series Classification with Shapelets

(draft lecture notes in Advanced Data Mining)

Piotr Lipiński

Definitions

- **Time Series Data:** Consider a set of I time series X_i, for $i = 1, 2, \ldots, I$, where each time series $X_i = (x_t^{(i)})$ is a sequence

$$x_1^{(i)}, x_2^{(i)}, \ldots, x_Q^{(i)}$$

  of Q observations in the successive time instants $t = 1, 2, \ldots, Q$, where Q is the length of the time series (the same for all time series considered).
- **Time Series Labels:** Each time series X_i is labelled with a target value $y_i \in \{1, 2, \ldots, C\}$, where C is the number of classes.

Definitions

- **Shapelets:** Consider a set of K shapelets S_k, for $k = 1, 2, \ldots, K$, where each shapelet $S_k = (s_l^{(k)})$ is a sequence

$$s_1^{(k)}, s_2^{(k)}, \ldots, s_L^{(k)}$$

of L elements $l = 1, 2, \ldots, L$, where L is the length of the shapelet (the same for all shapelets considered).

Definitions

- **Sliding Window Segment:** A sliding window segment of length L starting at time j of the time series X_i is the following sub-sequence of the time series

$$x_j^{(i)}, x_{j+1}^{(i)}, \ldots, x_{j+L-1}^{(i)}.$$

Definitions

- **Shapelet-Time Series Distance:** The distance $M_{i,k}$ between the shapelet S_k and the time series X_i is the minimum distance between the shapelet and each segment of the time series, i.e.

$$M_{i,k} = \min_{j=1,2,\ldots,J} \frac{1}{L} \sum_{l=1}^{L} |x_{j+l-1}^{(i)} - s_l^{(k)}|^2,$$

where $J = Q - L + 1$.

Definitions

- **Shapelet-based Time Series Representation:** For a given set of shapelets, each time series X_i can be encoded in a form of a K-dimensional vector

$$\boldsymbol{m}_i = (M_{i,1}, M_{i,2}, \ldots, M_{i,K}) \in \mathbb{R}^K.$$

- **Shapelet-based Time Series Classification:** For a given set of shapelets, each time series may be encoded in the shapelet-based representation and a regular classification approach may be used.

General Approach

- **IDEA:** Try to define the set of shapelets in such a way that it leads to an efficient classification of the time series in the shapelet-based representation.
- **REMARK 1:** For the sake of simplicity, consider the binary classification problem (i.e. $C = 2$ and $y_i \in \{0, 1\}$).
- **REMARK 2:** Classification will be based on logistic regression classification.

Learning Model

- According to the regression approach, the class label of each time series X_i, should be predicted by

$$\hat{y} = w_0 + \sum_{k=1}^{K} w_k M_{i,k},$$

where $w_0, w_1, \ldots, w_K$ are the regression parameters.

- As the minimum function in $M_{i,k}$ is not differentiable, it would be replaced with the **soft-minimum function** resulting in

$$\hat{M}_{i,k} = \frac{\sum_{j=1}^{J} D_{i,k,j} e^{\alpha D_{i,k,j}}}{\sum_{j=1}^{J} e^{\alpha D_{i,k,j}}} \approx \min_{j=1,2,\ldots,J} D_{i,k,j} = M_{i,k},$$

where

$$D_{i,k,j} = \frac{1}{L} \sum_{l=1}^{L} |x_{j+l-1}^{(i)} - s_l^{(k)}|^2.$$

Learning Model

- Finally

$$\hat{y} = w_0 + \sum_{k=1}^{K} w_k \hat{M}_{i,k},$$

and the regression parameters $w_0, w_1, \ldots, w_K$ as well as the shapelets $S_1, S_2, \ldots, S_K$ (required to evaluate $\hat{M}_{i,k}$) will be defined in the optimization process of an objective function $\mathcal{F}(\mathcal{S}, \boldsymbol{w})$ (proposed in the next slides), i.e.

$$\mathcal{S}, \boldsymbol{w} = \arg\min_{\mathcal{S}, \boldsymbol{w}} \mathcal{F}(\mathcal{S}, \boldsymbol{w}),$$

where $\mathcal{S}$ denotes the set of shapelets $S_1, S_2, \ldots, S_K$ and $\boldsymbol{w} = (w_0, w_1, \ldots, w_K)$ denotes the regression parameters.

Learning Model

- In order to compare the predicted class label $\hat{y}_i$ with the target class label y_i, we consider **the logistic regression loss function**

$$\mathcal{L}(y, \hat{y}) = -y \log(\sigma(\hat{y})) - (1 - y) \log(1 - \sigma(\hat{y})),$$

  where

$$\sigma(y) = \frac{1}{(1 + e^{-y})}$$

  is **the logistic sigmoid function**.

- In order to evaluate the regression parameters $\boldsymbol{w}$ and the set of shapelets $\mathcal{S}$, we consider the objective function

$$\mathcal{F}(\mathcal{S}, \boldsymbol{w}) = \sum_{i=1}^{I} \mathcal{L}(y_i, \hat{y}_i) + \lambda_w |\boldsymbol{w}|^2,$$

  where λ_w is a regularization parameter.

Stochastic Gradient Descent Approach

- Considering one data sample, the time series X_i, its contribution to the objective function $\mathcal{F}$ may be approximated by

$$\mathcal{F}_i = \mathcal{L}(y_i, \hat{y}_i) + \frac{\lambda_w}{I} \sum_{k=1}^{K} w_k^2.$$

Learning Time Series Shapelets

for iteration $= 1, 2, \dots,$ max-iter **do**

  for $i = 1, 2, \dots, I$ **do**

    for $k = 1, 2, \dots, K$ **do**

      $w_k \leftarrow w_k - \eta \frac{\partial \mathcal{F}_i}{\partial w_k}$

      for $l = 1, 2, \dots, L$ **do**

        $s_l^{(k)} \leftarrow s_l^{(k)} - \eta \frac{\partial \mathcal{F}_i}{\partial s_l^{(k)}}$

      end for

    end for

    $w_0 \leftarrow w_0 - \eta \frac{\partial \mathcal{F}_i}{\partial w_0}$

  end for

end for

Stochastic Gradient Descent Approach

- Considering one element $s_l^{(k)}$ of one shapelet S_k, its gradient is

$$\frac{\partial \mathcal{F}_i}{\partial s_l^{(k)}} = \frac{\partial \mathcal{L}(y_i, \hat{y}_i)}{\partial s_l^{(k)}} = \frac{\partial \mathcal{L}(y_i, \hat{y}_i)}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial \hat{M}_{i,k}} \sum_{j=1}^{J} \frac{\partial \hat{M}_{i,k}}{\partial D_{i,k,j}} \frac{\partial D_{i,k,j}}{\partial s_l^{(k)}},$$

where

Stochastic Gradient Descent Approach

$$\frac{\partial \mathcal{L}(y_i, \hat{y}_i)}{\partial \hat{y}_i} = -(y_i - \sigma(\hat{y}_i)),$$

$$\frac{\partial \hat{y}_i}{\partial \hat{M}_{i,k}} = w_k,$$

$$\frac{\partial \hat{M}_{i,k}}{\partial D_{i,k,j}} = \frac{e^{\alpha D_{i,k,j}(1+\alpha(D_{i,k,j}-\hat{M}_{i,k}))}}{\sum_{j=1}^{J} e^{\alpha D_{i,k,j}}},$$

and

$$\frac{\partial D_{i,k,j}}{\partial s_l^{(k)}} = \frac{2}{L}(s_l^{(k)} - x_{j+l-1}^{(i)}).$$

Stochastic Gradient Descent Approach

- Considering a regression parameter w_k, for $k > 0$, its gradient is

$$\frac{\partial \mathcal{F}_i}{\partial w_k} = -(y_i - \sigma(\hat{y}_i))\hat{M}_{i,k} + \frac{2\lambda_w}{I} w_k,$$

- and, for $k = 0$, its gradient is

$$\frac{\partial \mathcal{F}_i}{\partial w_0} = -(y_i - \sigma(\hat{y}_i)).$$

Learning Time Series Shapelets

for iteration $= 1, 2, \ldots,$ max-iter **do**
  for $i = 1, 2, \ldots, I$ **do**
    for $k = 1, 2, \ldots, K$ **do**
      $w_k \leftarrow w_k - \eta \frac{\partial \mathcal{F}_i}{\partial w_k}$
      for $l = 1, 2, \ldots, L$ **do**
        $s_l^{(k)} \leftarrow s_l^{(k)} - \eta \frac{\partial \mathcal{F}_i}{\partial s_l^{(k)}}$
      end for
    end for
    $w_0 \leftarrow w_0 - \eta \frac{\partial \mathcal{F}_i}{\partial w_0}$
  end for
end for

References

J. Grabocka, N. Schilling, M. Wistuba, and L. Schmidt-Thieme.
Learning time-series shapelets.
In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 392–401, New York, NY, USA, 2014. Association for Computing Machinery.