Minimax Classification under Concept Drift with Multidimensional Adaptation and Performance Guarantees

Basque Center for Applied Mathematics-BCAM

Verónica Álvarez, valvarez@bcamath.org

Santiago Mazuelas, smazuelas@bcamath.org

Jose A. Lozano, jlozano@bcamath.org

Supervised classification under concept drift

$$
\begin{array}{cccccccc}
\mathrm{p}_1 & & \mathrm{p}_2 & \cdots & \mathrm{p}_t & & \mathrm{p}_{t+1} & \cdots \\
\wr & & \wr & & \wr & & \wr & \\
(x_1, y_1), & & (x_2, y_2), & \ldots, & (x_t, y_t), & & (x_{t+1}, y_{t+1}), & \ldots \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & \\
\mathrm{h}_1 & \rightarrow & \mathrm{h}_2 & \ldots & \mathrm{h}_t & \rightarrow & \mathrm{h}_{t+1} & \ldots
\end{array}
$$

Supervised classification under concept drift

$$
\begin{array}{ccccccc}
\mathrm{p}_1 & & \mathrm{p}_2 & \cdots & \mathrm{p}_t & & \mathrm{p}_{t+1} & \cdots \\
\wr & & \wr & & \wr & & \wr & \\
(x_1, y_1), & & (x_2, y_2), & \ldots, & (x_t, y_t), & & (x_{t+1}, y_{t+1}), & \ldots \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & \\
\mathrm{h}_1 & \rightarrow & \mathrm{h}_2 & \ldots & \mathrm{h}_t & \rightarrow & \mathrm{h}_{t+1} & \ldots
\end{array}
$$

For instance,

$$\mathrm{h}_{t+1} = \mathrm{h}_t$$

Supervised classification under concept drift

$$
\begin{array}{cccccccccc}
\mathrm{p}_1 & & \mathrm{p}_2 & \cdots & \mathrm{p}_t & & \mathrm{p}_{t+1} & \cdots \\
\wr & & \wr & & \wr & & \wr & \\
(x_1, y_1), & & (x_2, y_2), & \ldots, & (x_t, y_t), & & (x_{t+1}, y_{t+1}), & \ldots \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & \\
\mathrm{h}_1 & \rightarrow & \mathrm{h}_2 & \ldots & \mathrm{h}_t & \rightarrow & \mathrm{h}_{t+1} & \ldots
\end{array}
$$

For instance,

$$\mathrm{h}_{t+1} = \mathrm{h}_t - k \nabla \ell(\mathrm{h}_t, (x_t, y_t))$$

Supervised classification under concept drift

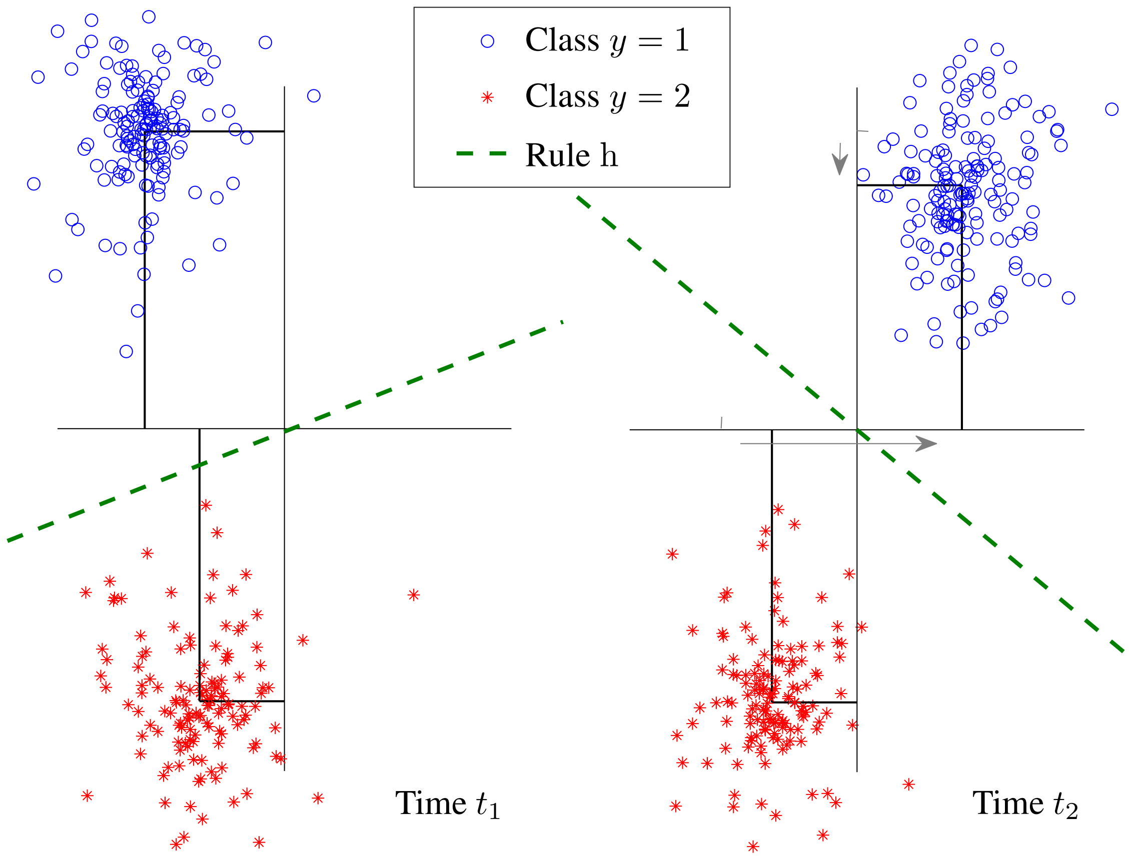
$$\begin{array}{ccccccccc}
\mathrm{p}_1 & & \mathrm{p}_2 & \cdots & \mathrm{p}_t & & \mathrm{p}_{t+1} & \cdots \\
\wr & & \wr & & \wr & & \wr & \\
(x_1, y_1), & & (x_2, y_2), & \ldots, & (x_t, y_t), & & (x_{t+1}, y_{t+1}), & \ldots \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & \\
\mathrm{h}_1 & \rightarrow & \mathrm{h}_2 & \ldots & \mathrm{h}_t & \rightarrow & \mathrm{h}_{t+1} & \ldots
\end{array}$$

For instance,

$$\mathrm{h}_{t+1} = \mathrm{h}_t - \boxed{k} \nabla \ell(\mathrm{h}_t, (x_t, y_t))$$

$k \in \mathbb{R}$ accounts for a global rate of change

Supervised classification under concept drift



$$\mathrm{h}_{t+1} = \mathrm{h}_t - \boxed{k} \nabla \ell(\mathrm{h}_t, (x_t, y_t))$$

$k \in \mathbb{R}$ accounts for a global rate of change

Knowledge gaps

Account for a scalar rate of change: learning rate or forgetting factor

Provide bounds in terms of non-computable quantities: discrepancies between consecutive distributions

Key contributions

Account for multidimensional time changes: model the evolution of each statistical characteristic of instance-label pairs

Provide computable tight bounds for: error probabilities and accumulated mistakes

Methodology of Adaptive Minimax Risk Classifiers

Multidimensional adaptation

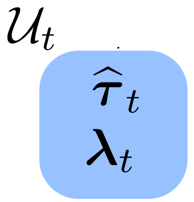
$$\boldsymbol{\tau}_t = \mathbb{E}_{\mathrm{p_t}}\{\Phi(x,y)\}$$

Feature mapping

$$\Phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^m$$

Uncertainty set

$$\mathcal{U}_t = \{\mathrm{p} \in \Delta(\mathcal{X} \times \mathcal{Y}) : |\mathbb{E}_{\mathrm{p}}\{\Phi(x,y)\} - \widehat{\boldsymbol{\tau}}_t| \preceq \boldsymbol{\lambda}_t\}$$



Methodology of Adaptive Minimax Risk Classifiers

Multidimensional adaptation

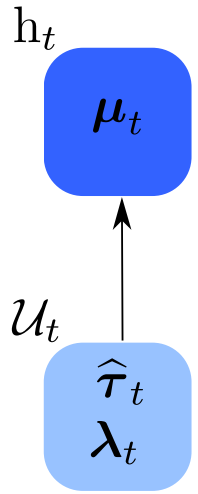
$$\boldsymbol{\tau}_t = \mathbb{E}_{\mathrm{p_t}}\{\Phi(x,y)\}$$

Feature mapping

$$\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^m$$

Uncertainty set

$$\mathcal{U}_t = \{\mathrm{p} \in \Delta(\mathcal{X}\times\mathcal{Y}) : |\mathbb{E}_{\mathrm{p}}\{\Phi(x,y)\} - \widehat{\boldsymbol{\tau}}_t| \preceq \boldsymbol{\lambda}_t\}$$



Learning

$$\min_{\mathrm{h}\in\mathrm{T}(\mathcal{X},\mathcal{Y})} \max_{\mathrm{p}\in\mathcal{U}_t} \ell(\mathrm{h},\mathrm{p})$$

$$= \min_{\boldsymbol{\mu}} 1 - \boldsymbol{\tau}_t^{\mathrm{T}}\boldsymbol{\mu} + \varphi(\boldsymbol{\mu}) + \boldsymbol{\lambda}_t^{\mathrm{T}}|\boldsymbol{\mu}|$$

Prediction

$$\hat{y} \in \arg\max_{y\in\mathcal{Y}} \Phi(x,y)^{\mathrm{T}}\boldsymbol{\mu}^*$$

Methodology of Adaptive Minimax Risk Classifiers

Multidimensional adaptation

$$\boldsymbol{\tau}_t = \mathbb{E}_{\mathrm{p_t}}\{\Phi(x,y)\}$$

Feature mapping

$$\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^m$$

Uncertainty set

$$\mathcal{U}_t = \{\mathrm{p} \in \Delta(\mathcal{X} \times \mathcal{Y}) : |\mathbb{E}_{\mathrm{p}}\{\Phi(x,y)\} - \widehat{\boldsymbol{\tau}}_t| \preceq \boldsymbol{\lambda}_t\}$$

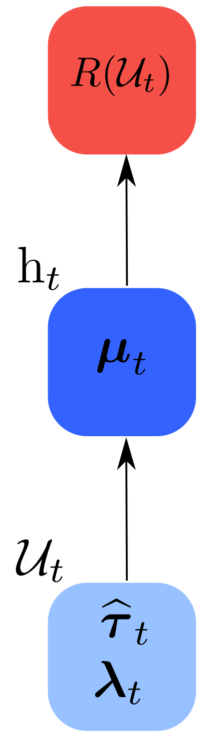
Learning

$$\min_{\mathrm{h} \in \mathrm{T}(\mathcal{X},\mathcal{Y})} \max_{\mathrm{p} \in \mathcal{U}_t} \ell(\mathrm{h},\mathrm{p})$$

$$= \min_{\boldsymbol{\mu}} 1 - \boldsymbol{\tau}_t^{\mathrm{T}}\boldsymbol{\mu} + \varphi(\boldsymbol{\mu}) + \boldsymbol{\lambda}_t^{\mathrm{T}}|\boldsymbol{\mu}| = R(\mathcal{U}_t)$$

Prediction

$$\hat{y} \in \arg\max_{y \in \mathcal{Y}} \Phi(x,y)^{\mathrm{T}}\boldsymbol{\mu}^*$$



Methodology of Adaptive Minimax Risk Classifiers

Multidimensional adaptation

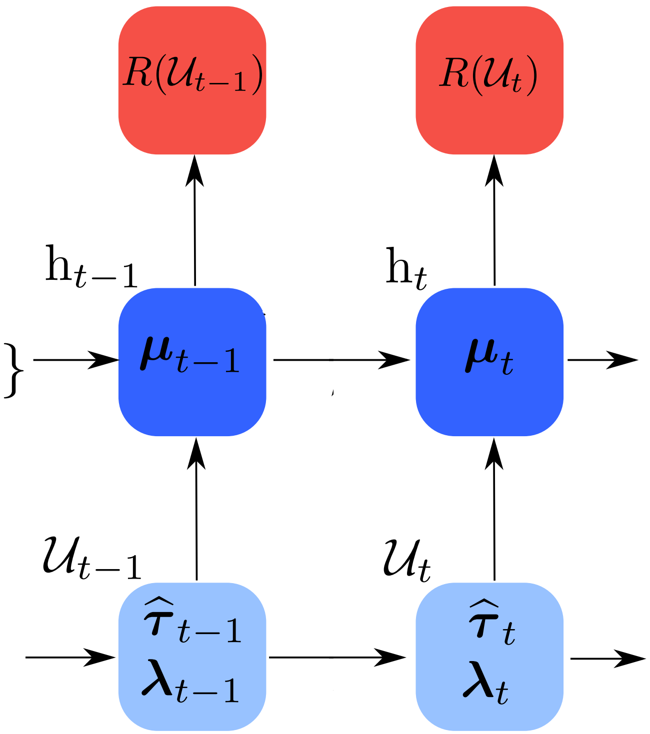
$$\boldsymbol{\tau}_t = \mathbb{E}_{\mathrm{p_t}}\{\Phi(x,y)\}$$

Feature mapping

$$\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^m$$

Uncertainty set

$$\mathcal{U}_t = \{\mathrm{p} \in \Delta(\mathcal{X} \times \mathcal{Y}) : |\mathbb{E}_{\mathrm{p}}\{\Phi(x,y)\} - \widehat{\boldsymbol{\tau}}_t| \preceq \boldsymbol{\lambda}_t\}$$



Learning

$$\min_{\mathrm{h} \in \mathrm{T}(\mathcal{X},\mathcal{Y})} \max_{\mathrm{p} \in \mathcal{U}_t} \ell(\mathrm{h}, \mathrm{p})$$

$$= \min_{\boldsymbol{\mu}} 1 - \boldsymbol{\tau}_t^{\mathrm{T}} \boldsymbol{\mu} + \varphi(\boldsymbol{\mu}) + \boldsymbol{\lambda}_t^{\mathrm{T}} |\boldsymbol{\mu}| = R(\mathcal{U}_t)$$

Prediction

$$\hat{y} \in \arg\max_{y \in \mathcal{Y}} \Phi(x,y)^{\mathrm{T}} \boldsymbol{\mu}^*$$

Multidimensional Adaptation

Tracking underlying distribution: obtain $\widehat{\boldsymbol{\tau}}_t, \boldsymbol{\lambda}_t$

Dynamical system: model the evolution of each component of $\boldsymbol{\tau}_t$

$$\boldsymbol{\eta}_{t,i} = \mathbf{H}_t \boldsymbol{\eta}_{t-1,i} + \mathbf{w}_{t,i}$$
$$\Phi_i(x_t, y_t) = \tau_{t,i} + v_{t,i}$$

with $\boldsymbol{\eta}_{t,i} = \left[\tau_{t,i}, \tau'_{t,i}, \tau''_{t,i}, \ldots, \tau^{k)}_{t,i}\right]^{\mathrm{T}}$

Multidimensional Adaptation

Tracking underlying distribution: obtain $\widehat{\boldsymbol{\tau}}_t, \boldsymbol{\lambda}_t$

Dynamical system: model the evolution of each component of $\boldsymbol{\tau}_t$

$$\boldsymbol{\eta}_{t,i} = \mathbf{H}_t \boldsymbol{\eta}_{t-1,i} + \mathbf{w}_{t,i}$$
$$\Phi_i(x_t, y_t) = \tau_{t,i} + v_{t,i}$$

with $\boldsymbol{\eta}_{t,i} = \left[\tau_{t,i}, \tau'_{t,i}, \tau''_{t,i}, \ldots, \tau^{k)}_{t,i}\right]^{\mathrm{T}}$

Unbiased linear estimator of $\boldsymbol{\eta}_{t,i}$ with minimum MSE

$$\widehat{\boldsymbol{\eta}}_{t,i} = \mathbf{H}_t \widehat{\boldsymbol{\eta}}_{t-1,i} - \mathbf{k}_{t,i}(\widehat{\tau}_{t-1,i} - \Phi_i(x_{t-1}, y_{t-1}))$$

Multidimensional Adaptation

Tracking underlying distribution: obtain $\widehat{\boldsymbol{\tau}}_t, \boldsymbol{\lambda}_t$

Dynamical system: model the evolution of each component of $\boldsymbol{\tau}_t$

$$\boldsymbol{\eta}_{t,i} = \mathbf{H}_t \boldsymbol{\eta}_{t-1,i} + \mathbf{w}_{t,i}$$
$$\Phi_i(x_t, y_t) = \tau_{t,i} + v_{t,i}$$

with $\boldsymbol{\eta}_{t,i} = \left[\tau_{t,i}, \tau'_{t,i}, \tau''_{t,i}, \ldots, \tau^{k)}_{t,i}\right]^{\mathrm{T}}$

Unbiased linear estimator of $\boldsymbol{\eta}_{t,i}$ with minimum MSE

$$\widehat{\boldsymbol{\eta}}_{t,i} = \mathbf{H}_t \widehat{\boldsymbol{\eta}}_{t-1,i} - \boxed{\mathbf{k}_{t,i}}(\widehat{\tau}_{t-1,i} - \Phi_i(x_{t-1}, y_{t-1}))$$

$\mathbf{k}_{t,1}, \mathbf{k}_{t,2}, \ldots, \mathbf{k}_{t,m}$ accounts for multidimensional time changes

Multidimensional Adaptation

Tracking underlying distribution: obtain $\widehat{\boldsymbol{\tau}}_t, \boldsymbol{\lambda}_t$

Dynamical system: model the evolution of each component of $\boldsymbol{\tau}_t$

$$\boldsymbol{\eta}_{t,i} = \mathbf{H}_t \boldsymbol{\eta}_{t-1,i} + \mathbf{w}_{t,i}$$
$$\Phi_i(x_t, y_t) = \tau_{t,i} + v_{t,i}$$

with $\boldsymbol{\eta}_{t,i} = \left[\tau_{t,i}, \tau'_{t,i}, \tau''_{t,i}, \ldots, \tau^{k)}_{t,i}\right]^{\mathrm{T}}$

Unbiased linear estimator of $\boldsymbol{\eta}_{t,i}$ with minimum MSE

$$\widehat{\boldsymbol{\eta}}_{t,i} = \mathbf{H}_t \widehat{\boldsymbol{\eta}}_{t-1,i} - \boxed{\mathbf{k}_{t,i}}(\widehat{\tau}_{t-1,i} - \Phi_i(x_{t-1}, y_{t-1}))$$

$\mathbf{k}_{t,1}, \mathbf{k}_{t,2}, \ldots, \mathbf{k}_{t,m}$ accounts for multidimensional time changes

Performance guarantees

Error probability

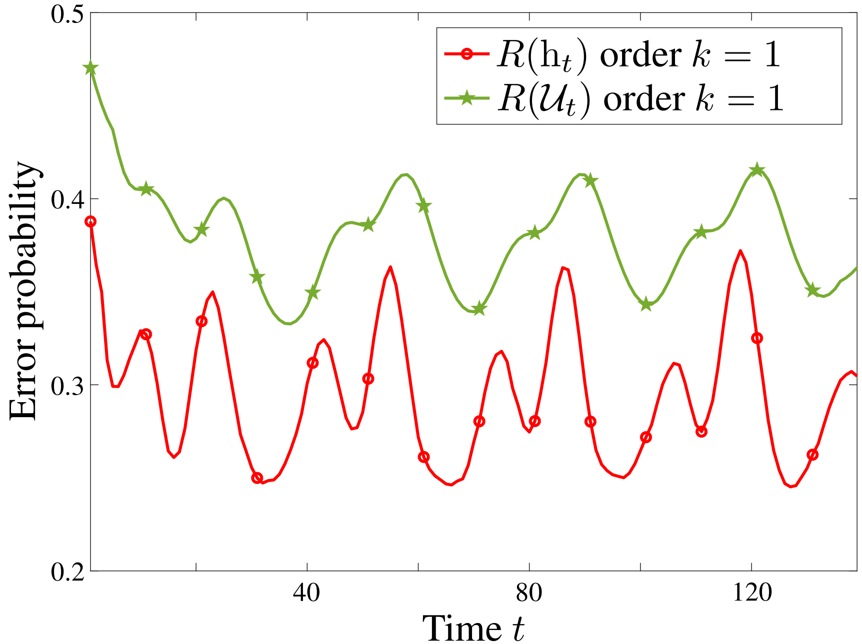
$$R(\mathrm{h}_t) \leq R(\mathcal{U}_t)$$

Accumulated mistakes

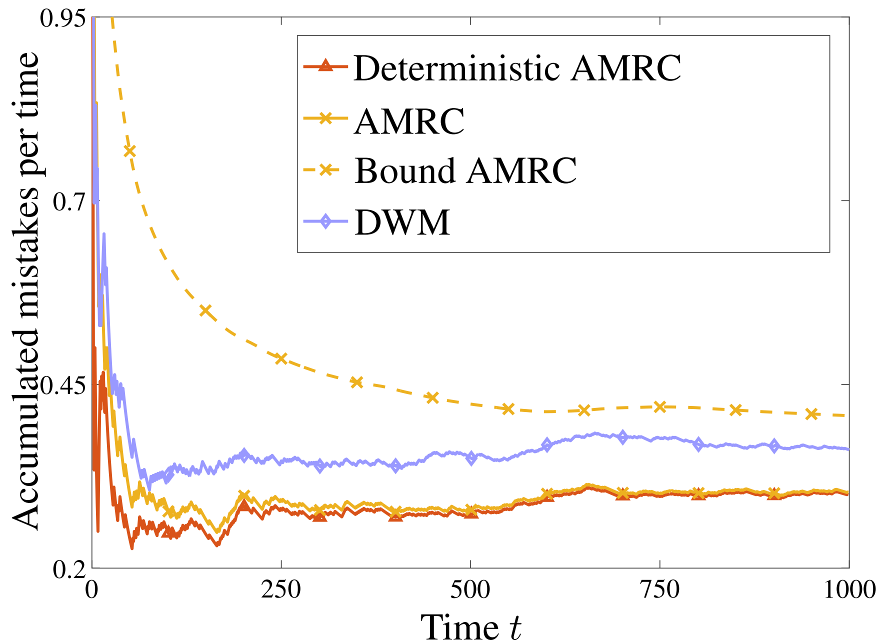
$$\sum_{t=1}^{T} \mathbb{I}\left\{\hat{y}_t \neq y_t\right\} \leq \sum_{t=1}^{T} R(\mathcal{U}_t) + \sqrt{2T \log \frac{1}{\delta}}$$

Experimental results

Instantaneous bounds for error probabilities



Bounds for accumulated mistakes



Algorithm	Weather	Elec2	German	Chess	Usenet1	Email	Poker
DWM	**30.0**	36.3	41.2	35.2	36.3	39.5	22.0
Projectron	30.7	36.8	40.0	35.1	49.1	48.2	22.6
Unid. AMRC	31.3	40.1	30.3	38.3	46.3	48.4	39.4
AMRC	32.3	35.8	30.3	**27.7**	35.7	43.7	**21.9**
Det. AMRC	**30.0**	**33.9**	**30.0**	33.4	**32.0**	**33.9**	**21.9**

Paper

Code