Spatio-Temporal Wildfire Prediction using Multi-Modal Data

Chen Xu, Yao Xie
Daniel A. Zuniga Vazquez, Rui Yao, and Feng Qiu

INFORMS 2023, Session TD69 -
Data Mining for Power System Resilience and Reliability

Layout

- Motivation
- Problem setup
- Methods
- Theoretical guarantee
- Numerical results
- Summary

Motivation

- Wildfire incidents have severe consequences in power systems and the economy in general.
- For instance, utility companies have to schedule utility shutdown for high wildfire risk regions.

Motivation

- Wildfire incidents have severe consequences in power systems and the economy in general.
- For instance, utility companies have to schedule utility shutdown for high wildfire risk regions.
- However, existing metrics for measuring fire risks (e.g., burning index, fire load index) are static features lacking adaptivity.

Motivation

- Wildfire incidents have severe consequences in power systems and the economy in general.
- For instance, utility companies have to schedule utility shutdown for high wildfire risk regions.
- However, existing metrics for measuring fire risks (e.g., burning index, fire load index) are static features lacking adaptivity.
- Multi-modal features are available: weather variables, infrastructure information, etc.

Motivation

However, there are technical challenges in effective modeling.

- Events happen at specific locations but asynchronously (e.g. influence changes over time and should not be captured by fixed metrics).

Motivation

However, there are technical challenges in effective modeling.

- Events happen at specific locations but asynchronously (e.g. influence changes over time and should not be captured by fixed metrics).
- Data are limited, with only one-class wildfire information available.

Motivation

However, there are technical challenges in effective modeling.

- Events happen at specific locations but asynchronously (e.g. influence changes over time and should not be captured by fixed metrics).
- Data are limited, with only one-class wildfire information available.
- To predict wildfire severity, classification methods can also been used, whose uncertainty analyses remain unexplored.

Problem setup

- Each observed wildfire datum

$$x_i = (t_i, u_i, m_i), t_i \in [0, T], u_i \in \mathbb{R}^2, m_i \in \mathbb{R}^d$$

is a tuple consisting of time, location, and marks.

Problem setup

- Each observed wildfire datum

$$x_i = (t_i, u_i, m_i), t_i \in [0, T], u_i \in \mathbb{R}^2, m_i \in \mathbb{R}^d$$

  is a tuple consisting of time, location, and marks.
- In particular, $m_i = (z_i, m_i')$ contains both static marks z_i and dynamic marks m_i'.
- Example of static marks z_i: Fire zone tier, vegetation type, road condition.
- Example of dynamic marks m_i': Seasonality, weather, fire probability.

Problem setup

- Each observed wildfire datum

$$x_i = (t_i, u_i, m_i), t_i \in [0, T], u_i \in \mathbb{R}^2, m_i \in \mathbb{R}^d$$

  is a tuple consisting of time, location, and marks.
- In particular, $m_i = (z_i, m_i')$ contains both static marks z_i and dynamic marks m_i'.
- Example of static marks z_i: Fire zone tier, vegetation type, road condition.
- Example of dynamic marks m_i': Seasonality, weather, fire probability.
- For each x_i, we may also know its severity $y_i \in \{0, \ldots, M\}$.

Problem setup

We have two tasks in this work:

1. Model the probability of $x(u, t)$ at location u and $t > T$.
   Solution: Marked spatio-temporal Hawkes process.

Problem setup

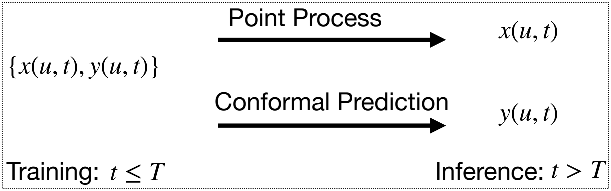
We have two tasks in this work:

1. Model the probability of $x(u,t)$ at location u and $t > T$.
   Solution: Marked spatio-temporal Hawkes process.
2. Predict the severity y, assuming a fire event x will happen.
   Solution: Machine learning classifiers, whose predictions $\hat{y}$ are calibrated to provide uncertainty sets for true event types.

Problem setup

We have two tasks in this work:

1. Model the probability of $x(u, t)$ at location u and $t > T$.
   Solution: Marked spatio-temporal Hawkes process.
2. Predict the severity y, assuming a fire event x will happen.
   Solution: Machine learning classifiers, whose predictions $\hat{y}$ are calibrated to provide uncertainty sets for true event types.



Method–Marked Spatio-temporal Hawkes process

- First, we focus on the task of predicting instantaenous wildfire probability of $x(u, t)$.

Method–Marked Spatio-temporal Hawkes process

- First, we focus on the task of predicting instantaenous wildfire probability of $x(u, t)$.
- Denote $\mathcal{H}_t$ as the history of neighboring and historical events:
- The Hawkes point process models the conditional intensity

$$\lambda(t, u, m \mid \mathcal{H}_t) := \lim_{\Delta t, \Delta u \to 0} \frac{\mathbb{E}\left[N\left([t, t + \Delta t) \cdot B(u, \Delta u) \cdot B(m, \Delta m) \mid \mathcal{H}_t\right]\right.}{\Delta t \cdot B(u, \Delta u) \cdot B(m, \Delta m)}$$

Method–Marked Spatio-temporal Hawkes process

- First, we focus on the task of predicting instantaenous wildfire probability of $x(u,t)$.
- Denote $\mathcal{H}_t$ as the history of neighboring and historical events:
- The Hawkes point process models the conditional intensity

$$\lambda\left(t,u,m \mid \mathcal{H}_t\right) := \lim_{\Delta t,\Delta u\to 0} \frac{\mathbb{E}\left[N\left([t,t+\Delta t\right)\cdot B(u,\Delta u)\cdot B(m,\Delta m) \mid \mathcal{H}_t\right]}{\Delta t\cdot B(u,\Delta u)\cdot B(m,\Delta m)}$$

- Numerous works on parametrizing the intensity function λ (Hawkes, 1971; Daley and Vere-Jones, 2003; Bauwens and Hautsch, 2009; Zhu et al. 2021; Dong et al. 2022).
- See (Reinhart 2018) for a survey of Hawkes process models.

Method–Marked Spatio-temporal Hawkes process

- In this work, we assume the intensity has the form

$$\lambda(t, u, m \mid \mathcal{H}_t) = \underbrace{\left(f_u + \sum_{j:t_j<t} f_u(u_j, u) \cdot f_t(t_j, t)\right)}_{\lambda_g(t,u):=} \cdot f_m(m), \tag{1}$$

Method–Marked Spatio-temporal Hawkes process

- In this work, we assume the intensity has the form

$$\lambda(t, u, m \mid \mathcal{H}_t) = \underbrace{\left(f_u + \sum_{j:t_j<t} f_u(u_j, u) \cdot f_t(t_j, t) \right)}_{\lambda_g(t,u):=} \cdot f_m(m), \tag{1}$$

- $f_u = \mu_u$ is the baseline rate, assuming location is discretized.

Method–Marked Spatio-temporal Hawkes process

- In this work, we assume the intensity has the form

$$\lambda(t, u, m \mid \mathcal{H}_t) = \underbrace{\left(f_u + \sum_{j: t_j < t} f_u(u_j, u) \cdot f_t(t_j, t) \right)}_{\lambda_g(t,u):=} \cdot f_m(m), \tag{1}$$

- $f_u = \mu_u$ is the baseline rate, assuming location is discretized.
- $f_u(u_j, u) = \alpha_{u_j, u}$ captures interactions among locations.

Method–Marked Spatio-temporal Hawkes process

- In this work, we assume the intensity has the form

$$\lambda\left(t, u, m \mid \mathcal{H}_t\right)=\underbrace{\left(f_u+\sum_{j: t_j<t} f_u\left(u_j, u\right) \cdot f_t\left(t_j, t\right)\right)}_{\lambda_g(t, u):=} \cdot f_m(m), \tag{1}$$

- $f_u=\mu_u$ is the baseline rate, assuming location is discretized.
- $f_u\left(u_j, u\right)=\alpha_{u_j, u}$ captures interactions among locations.
- $f_t\left(t_j, t\right)=\beta \exp(-\beta(t-t_j))$ denotes temporal influence from past events.

Method–Marked Spatio-temporal Hawkes process

- In this work, we assume the intensity has the form

$$\lambda\left(t, u, m \mid \mathcal{H}_t\right) = \underbrace{\left(f_u + \sum_{j: t_j < t} f_u\left(u_j, u\right) \cdot f_t\left(t_j, t\right)\right)}_{\lambda_g(t,u) :=} \cdot f_m\left(m\right), \tag{1}$$

- $f_u = \mu_u$ is the baseline rate, assuming location is discretized.
- $f_u\left(u_j, u\right) = \alpha_{u_j, u}$ captures interactions among locations.
- $f_t\left(t_j, t\right) = \beta \exp(-\beta(t - t_j))$ denotes temporal influence from past events.
- $f_m\left(m_j'\right)$ measures contribution from marks. We consider (1) $f_m\left(m\right) = \gamma^T m$ (`LinearSTHawkes`) (2) f_m as a pre-trained feature extractor (`NonLinearSTHawkes`).

Method–Marked Spatio-temporal Hawkes process

- In this work, we assume the intensity has the form

$$\lambda(t, u, m \mid \mathcal{H}_t) = \underbrace{\left(f_u + \sum_{j: t_j < t} f_u(u_j, u) \cdot f_t(t_j, t) \right)}_{\lambda_g(t,u):=} \cdot f_m(m), \tag{1}$$

- $f_u = \mu_u$ is the baseline rate, assuming location is discretized.
- $f_u(u_j, u) = \alpha_{u_j, u}$ captures interactions among locations.
- $f_t(t_j, t) = \beta \exp(-\beta(t - t_j))$ denotes temporal influence from past events.
- $f_m(m'_j)$ measures contribution from marks. We consider
  (1) $f_m(m) = \gamma^T m$ (`LinearSTHawkes`)
  (2) f_m as a pre-trained feature extractor (`NonLinearSTHawkes`).
- The goal is to estimate parameters θ given past $\{x_i\}$.

Estimation–Marked Spatio-temporal Hawkes process

- The full log-likelihood given n observations:

$$\ell(\theta) = \sum_{i=1}^{n} \log\left(\lambda\left(t_i, u_i, m_i\right)\right) - \int_0^T \int_U \lambda_g(t, u) \mathrm{d}u \mathrm{d}t, \tag{2}$$

whereby parameters are solved via maximum likelihood estimation.

Estimation–Marked Spatio-temporal Hawkes process

- The full log-likelihood given n observations:

$$\ell(\theta) = \sum_{i=1}^{n} \log\left(\lambda\left(t_i, u_i, m_i\right)\right) - \int_0^T \int_U \lambda_g(t, u) \mathrm{d}u \mathrm{d}t, \qquad (2)$$

  whereby parameters are solved via maximum likelihood estimation.

- Constraints can also be included (e.g., sparsity in interaction, contribution from marks, etc.).

Estimation–Marked Spatio-temporal Hawkes process

- Full log-likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log\left(\lambda\left(t_i, u_i, m_i\right)\right) - \int_0^T \int_U \lambda_g(t, u) \mathrm{d}u \mathrm{d}t, \qquad (3)$$

- In general, (3) is non-convex in θ and MLE can be computationally costly.

Estimation–Marked Spatio-temporal Hawkes process

- Full log-likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log\left(\lambda\left(t_i, u_i, m_i\right)\right) - \int_0^T \int_U \lambda_g(t,u)\mathrm{d}u\mathrm{d}t, \qquad (3)$$

- In general, (3) is non-convex in θ and MLE can be computationally costly.
- However, our previous parametrization of $\lambda(t,u,m|\mathcal{H}_t)$ ensures $\ell(\theta)$ is convex in all but $\beta \in \mathbb{R}_+$.

Estimation–Marked Spatio-temporal Hawkes process

- Full log-likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log\left(\lambda\left(t_i, u_i, m_i\right)\right) - \int_0^T \int_U \lambda_g(t, u) \mathrm{d}u \mathrm{d}t, \qquad (3)$$

- In general, (3) is non-convex in θ and MLE can be computationally costly.
- However, our previous parametrization of $\lambda(t, u, m|\mathcal{H}_t)$ ensures $\ell(\theta)$ is convex in all but $\beta \in \mathbb{R}_+$.
- Thus, estimates can be found via simple iterative procedures with *performance guarantees*.

Estimation–Marked Spatio-temporal Hawkes process

- Full log-likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log\left(\lambda\left(t_i, u_i, m_i\right)\right) - \int_0^T \int_U \lambda_g(t, u) \mathrm{d}u \mathrm{d}t, \qquad (3)$$

- In general, (3) is non-convex in θ and MLE can be computationally costly.
- However, our previous parametrization of $\lambda(t, u, m|\mathcal{H}_t)$ ensures $\ell(\theta)$ is convex in all but $\beta \in \mathbb{R}_+$.
- Thus, estimates can be found via simple iterative procedures with *performance guarantees*.
- Once parameters are estimated, they are substituted in (1) for prediction.

Method–Conformal Prediction for Dependent Data

- Now, we switch gear to predict severity y and quantify prediction uncertainty.

Method–Conformal Prediction for Dependent Data

- Now, we switch gear to predict severity y and quantify prediction uncertainty.
- Typical classification setting, where $x_i \in \mathbb{R}^{1+2+d}$ (time, location, feature dimension) is used to predict its severity $y_i \in [M]$.
- Classifier $\hat{f}$ can be any generic (machine learning) models.

Method–Conformal Prediction for Dependent Data

- Now, we switch gear to predict severity y and quantify prediction uncertainty.
- Typical classification setting, where $x_i \in \mathbb{R}^{1+2+d}$ (time, location, feature dimension) is used to predict its severity $y_i \in [M]$.
- Classifier $\hat{f}$ can be any generic (machine learning) models.
- **In particular**, we want to produce uncertainty sets $C(x_t, \alpha) \subset [M]$, so that marginal coverage holds:

$$\mathbb{P}(Y_t \in C(x_t, \alpha)) \geq 1 - \alpha. \tag{4}$$

Importantly, we want $C(x_t, \alpha)$ to be distribution-free.

Method–Conformal Prediction for Dependent Data

- Our solution is conformal prediction (Shafer and Vovk 2008), which intuitively includes in $C(x_t, \alpha)$ all possible types that *conform* to past observed types.

Method–Conformal Prediction for Dependent Data

- Our solution is conformal prediction (Shafer and Vovk 2008), which intuitively includes in $C(x_t, \alpha)$ all possible types that *conform* to past observed types.
- **Benefits** (1) distribution-free (2) suitable for general prediction model (3) satisfy coverage guarantee.

Method–Conformal Prediction for Dependent Data

- Our solution is conformal prediction (Shafer and Vovk 2008), which intuitively includes in $C(x_t, \alpha)$ all possible types that *conform* to past observed types.
- **Benefits** (1) distribution-free (2) suitable for general prediction model (3) satisfy coverage guarantee.
- **Limitation**: Data must be exchangeable (e.g., i.i.d.).

Method–Conformal Prediction for Dependent Data

- Our solution is conformal prediction (Shafer and Vovk 2008), which intuitively includes in $C(x_t, \alpha)$ all possible types that *conform* to past observed types.
- **Benefits** (1) distribution-free (2) suitable for general prediction model (3) satisfy coverage guarantee.
- **Limitation**: Data must be exchangeable (e.g., i.i.d.).
- **Remedy**: Inspired by our recent work on CP for time-series regression (Xu and Xie 2021), we design methods that achieve coverage approximately with theoretical guarantees (Xu and Xie 2022).

Method–Conformal Prediction for Dependent Data

The proposed CP method involves four steps

Method–Conformal Prediction for Dependent Data

The proposed CP method involves four steps

1. Train classifiers $\hat{f}$ as leave-one-out ensemble estimators to maximize prediction accuracy.

Method–Conformal Prediction for Dependent Data

The proposed CP method involves four steps

1. Train classifiers $\hat{f}$ as leave-one-out ensemble estimators to maximize prediction accuracy.
2. Define non-conformity scores $\tau(\hat{f}, x, y)$ dependent on the classifier, borrowing ideas in (Angelopoulos et al. 2021).

Method–Conformal Prediction for Dependent Data

The proposed CP method involves four steps

1. Train classifiers $\hat{f}$ as leave-one-out ensemble estimators to maximize prediction accuracy.
2. Define non-conformity scores $\tau(\hat{f}, x, y)$ dependent on the classifier, borrowing ideas in (Angelopoulos et al. 2021).
3. Compute τ on (x_i, y_i) with leave-one-out $\hat{f}$'s to obtain $\{\hat{\tau}_i\}$.

Method–Conformal Prediction for Dependent Data

The proposed CP method involves four steps

1. Train classifiers $\hat{f}$ as leave-one-out ensemble estimators to maximize prediction accuracy.
2. Define non-conformity scores $\tau(\hat{f}, x, y)$ dependent on the classifier, borrowing ideas in (Angelopoulos et al. 2021).
3. Compute τ on (x_i, y_i) with leave-one-out $\hat{f}$'s to obtain $\{\hat{\tau}_i\}$.
4. Given x_t, return the $1-\alpha$ prediction set as

$$\widehat{C}(x_t, \alpha) := \{c \in [C] : \sum_{j=t-n}^{t-1} \mathbf{1}(\hat{\tau}_j \leq \tau(\hat{f}, x_t, c))/n < 1-\alpha\}.$$

Theory: Hawkes process estimation

We borrow the technique from (Juditsky and Nemirovski 2019) to prove the convergence guarantee of $\hat{\theta}$ as an estimator of $\theta^* \in \Theta$.

Theorem ((Informal) Parameter estimation guarantee)

Suppose Θ is compact and the gradient of log-likelihood objective is bounded, then the estimator $\hat{\theta}$ obeys the bound

$$\|\hat{\theta} - \theta^*\|_2^2 = \mathcal{O}\left(\frac{1}{J^2} + \frac{1}{k+1}\right), \tag{5}$$

where J is the number of grid search of β over $[\beta_0, \beta_1]$ and k is the number of projected gradient descent steps of $\theta - \{\beta\}$ per iterative training given β.

Theory: Conformal prediction

We extend the analyses in (Xu and Xie 2021) to prove the coverage and set size guarantees.

- Let $\tau(f, x, y)$ be the true non-conformity score dependent on $f = y|x$.
- Let $\delta_n^2 = \sum_{j=1}^{n} (\tau_i - \hat{\tau}_i)^2/n$.

Theorem (Coverage guarantee)

Suppose the set of $\{\tau_i\}$ is i.i.d. For any training set of size n and significance level $\alpha \in (0, 1)$:

$$|\mathbb{P}(Y_t \notin \widehat{C}(x_t, \alpha)) - \alpha| \leq \mathcal{O}(\sqrt{\log(n)/n} + \delta_n^{2/3}). \tag{6}$$

Theory: Conformal prediction

We extend the analyses in (Xu and Xie 2021) to prove the coverage and set size guarantees.

- Let $\tau(f, x, y)$ be the true non-conformity score dependent on $f = y|x$.
- Let $\delta_n^2 = \sum_{j=1}^{n}(\tau_i - \hat{\tau}_i)^2/n$.

Theorem (Coverage guarantee)

Suppose the set of $\{\tau_i\}$ is i.i.d. For any training set of size n and significance level $\alpha \in (0,1)$:

$$|\mathbb{P}(Y_t \notin \widehat{C}(x_t, \alpha)) - \alpha| \leq \mathcal{O}(\sqrt{\log(n)/n} + \delta_n^{2/3}). \tag{6}$$

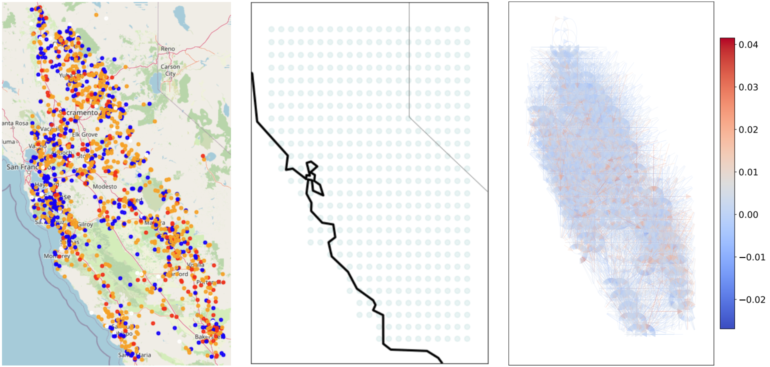
Theorem (Set size guarantee)

Under additional regularity condition, we have that there exists n' large enough so that for all $n \geq n'$

$$\widehat{C}(x_t, \alpha) \Delta C(x_t, \alpha) \leq 1 \tag{7}$$

Experiment–Marked Spatio-temporal Hawkes process

- We model the occurrence of California wildfire, which is known to have spatio-temporal dependencies.
- Figure (c) quantifies interaction strength through $\sum_i \alpha_{ij}$.



(a) Raw data **(b)** Grid discretization **(c)** Interaction α_{ij}

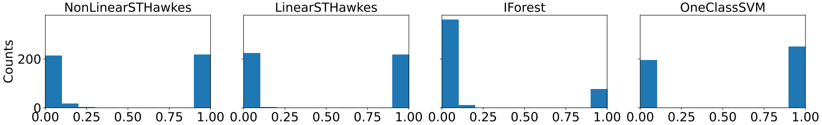
Experiment–Marked Spatio-temporal Hawkes process

- (`LinearSTHawkes`) Estimated parameters of static and dynamic marks, measuring their contributions.

	Three Largest Estimates			Three Smallest Estimates		
Static mark estimate	0.301	0.231	0.184	0.046	0.024	0.008
Static mark name	Fire Tier1	Fire Tier2	Fire Tier3	PHYS=Developed-Roads	PHYS=Conifer	PHYS=Developed
Dynamic mark estimate	0.57	0.472	0.46	0.217	0.117	0.02
Dynamic mark name	Summer	Temperature	Relative Humidity	LFP	Spring	Winter

Experiment–Marked Spatio-temporal Hawkes process

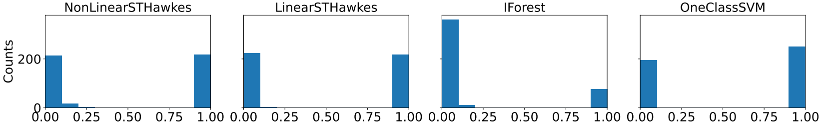
- Prediction by adapting *dynamic hedging* (Raginsky et al. 2012).
- Comparison against widely-used one-class classifiers.
- An inherently challenging task: 365 daily predictions $\times$ 453 locations for prediction with $\sim$500 true incidents.



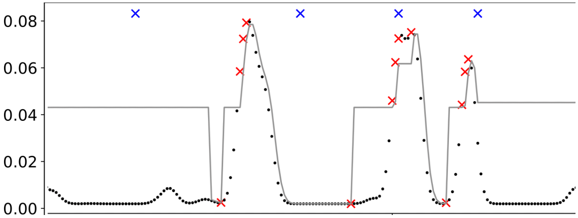
(a) F_1 score comparison of proposed method (left two) with baselines

Experiment–Marked Spatio-temporal Hawkes process

- Prediction by adapting *dynamic hedging* (Raginsky et al. 2012).
- Comparison against widely-used one-class classifiers.
- An inherently challenging task: 365 daily predictions $\times$ 453 locations for prediction with $\sim$500 true incidents.



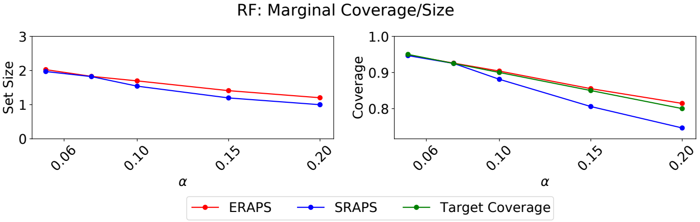
(a) F_1 score comparison of proposed method (left two) with baselines



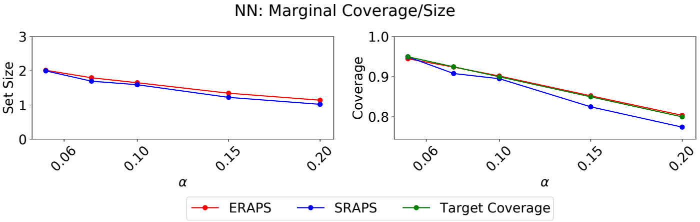
(b) `NonLinearSTHawkes` real-time prediction

Experiment–Conformal Prediction for Dependent Data

- Our method is named ERAPS.



(a) Random forest classifier



(b) Neural network classifier

Summary

- Model the conditional intensity of correlated observations with a flexible marked Spatio-temporal Hawkes process.
- Design the process to yield convex likelihood with efficient solving procedures.
- Provide uncertainty quantification for machine learning classifiers using recent advances in conformal prediction.

Reference: (Xu, Xie, et al. 2023) "Spatio-Temporal Wildfire Prediction Using Multi-Modal Data". *IEEE Journal on Selected Areas in Information Theory*, pp. 302–313.

References I

Angelopoulos, Anastasios Nikolas et al. (2021). "Uncertainty Sets for Image Classifiers using Conformal Prediction". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=eNdiU_DbM9.

Juditsky, Anatoli B and Arkadii S Nemirovski (2019). "Signal recovery by stochastic optimization". In: *Automation and Remote Control* 80, pp. 1878–1893.

Raginsky, Maxim et al. (Aug. 2012). "Sequential anomaly detection in the presence of noise and limited feedback". In: *IEEE Transactions on Information Theory* 58.8, pp. 5544–5562.

Reinhart, Alex (2018). "A review of self-exciting spatio-temporal point processes and their applications". In: *Statistical Science* 33.3, pp. 299–318.

Shafer, Glenn and Vladimir Vovk (2008). "A Tutorial on Conformal Prediction.". In: *Journal of Machine Learning Research* 9.3.

References II

Xu, Chen and Yao Xie (2021). "Conformal prediction interval for dynamic time-series". In: *International Conference on Machine Learning*. PMLR, pp. 11559–11569.

– (2022). "Conformal prediction set for time-series". In: *arXiv preprint arXiv:2206.07851*.

Xu, Chen, Yao Xie, et al. (2023). "Spatio-Temporal Wildfire Prediction Using Multi-Modal Data". In: *IEEE Journal on Selected Areas in Information Theory* 4, pp. 302–313. DOI: 10.1109/JSAIT.2023.3276054.