Bayesian feature discovery for predictive maintenance

Amir Dib[†], Charles Truong[†], Laurent Oudre[†], Mathilde Mougeot[†], Nicolas Vayatis[†], Heloïse Nonne[‡].

[†]Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, 91190, Gif-sur-Yvette, France
[‡]ITNOVEM, SNCF, 93120, Saint-Denis, France

European Signal Processing Conference (EUSIPCO 2021).

Predictive maintenance

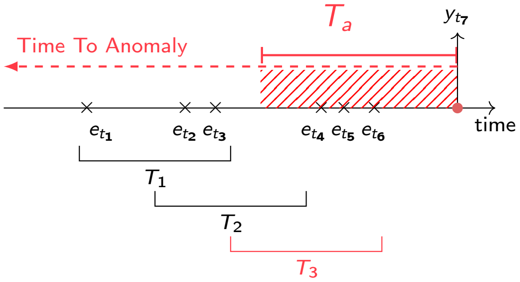

Figure: Temporal aggregation of log-events $(e_{t_1}, \dots, e_{t_6})$ over sliding windows (T_1, T_2, T_3). In red, events that occur in the period T_a before y_{t_7} are considered anomalous and labeled $l = 1$. The aggregation produces the itemsets $x_1 = \{e_{t_1}, e_{t_2}, e_{t_3}\}, x_2 = \{e_{t_2}, e_{t_3}\}, x_3 = \{e_{t_4}, e_{t_5}, e_{t_6}\}$ and the labels $l_1 = 0$, $l_2 = 0$ and $l_3 = 1$. The goal is to correctly predict the labels l_i from the itemsets x_i.

Background FIM

- Let $E = ed$ the base dictionary of events and $\mathcal{E} = \mathcal{P}(E)$ the collection of all 2^d possible patterns on E.

Background FIM

- Let $E = ed$ the base dictionary of events and $\mathcal{E} = \mathcal{P}(E)$ the collection of all 2^d possible patterns on E.
- A database of pattern from a random process valued in $\mathcal{E}$ is composed of ordered set of event from E and an associated label, such that $\mathcal{D} = \{(x_i, l_i)_{i=1}^n\}$ of elements of $\mathcal{E} \times \{0, 1\}$

Sequence	Label	Events
T_1	1	$\{e_1, e_2\}$
T_2	0	$\{e_1, e_2, e_4\}$
T_3	1	$\{e_1, e_2, e_3, e_4\}$
T_4	0	$\{e_1, e_3\}$
T_5	0	$\{e_2, e_3, e_4\}$ ...
...		

Background FIM

- Let $E = ed$ the base dictionary of events and $\mathcal{E} = \mathcal{P}(E)$ the collection of all 2^d possible patterns on E.
- A database of pattern from a random process valued in $\mathcal{E}$ is composed of ordered set of event from E and an associated label, such that $\mathcal{D} = \{(x_i, l_i)_{i=1}^n\}$ of elements of $\mathcal{E} \times \{0, 1\}$

Sequence	Label	Events
T_1	1	$\{e_1, e_2\}$
T_2	0	$\{e_1, e_2, e_4\}$
T_3	1	$\{e_1, e_2, e_3, e_4\}$
T_4	0	$\{e_1, e_3\}$
T_5	0	$\{e_2, e_3, e_4\}$ ...
...		

- Question: For any pattern in $x \in \mathcal{P}(E)$, what is the statistical difference of frequency in each class.

Discriminative pattern

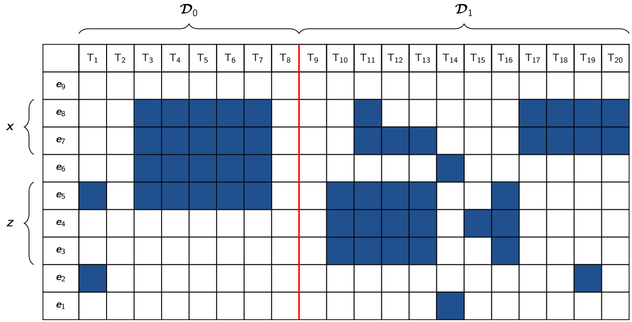

Figure: An example data set of events $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$. Row corresponds to items in $E = e9$ and columns to $n = 20$ samples. A blue colored area indicates that the item is present in the sample column considered. In this data set, the pattern $x = \{e_7, e_8\}$ in $\mathcal{E}$ seems to be nondiscriminative since $s_0(x) = s_1(x)$. On the contrary, the pattern $z = \{e_3, e_4, e_5\}$ appears to be specific to the positive class $l = 1$.

Discriminative pattern

- Discriminative pattern mining is an important problem with various application in many area;
- The fundamental difficult resides in the computation of frequency that requires to enumerate an exponential number of object. The problem is tipically NP-hard;
- All traditional approaches such as SPuManTE rely on a Mining step from a common frequent itemset miner on each class followed by a frequency based test [3].

BFP Algorithm

In the contrary, our approach is based on fitting a bayesian model on the process of sequences and a bayes ratio [2]. There is many advantages of this approach:

- Inference of the bayesian model can be performed by classifcal EM algorithm;
- No minimum user-treshold is required for the mining step [1];
- It is fast to evaluate any discriminative score since the frequency can be evaluated in closed-form;
- We can easily obtain confidence interval on the discriminative score by sampling from the joint distribution.

Pattern model

Let $\mathrm{X} = \boldsymbol{x}n$ be an i.i.d.sample and suppose the underlying model is a BMM with K components. For $k \in \{1, \ldots, K\}$, the k-ith sampling distribution $p_k(\boldsymbol{x_i}|\boldsymbol{\theta_k})$ depends has parameter $\boldsymbol{\theta_k} = (\theta_{kj})_{j=1}^{d}$. Denoting λ_k the probability of sampling from the k-th component with $\sum_{k=1}^{K} \lambda_k = 1$, the global sampling distribution writes

$$p_(\boldsymbol{x_i}|\Theta, \boldsymbol{\lambda}) = \sum_{h=1}^{K} \lambda_k p_k(\boldsymbol{x_i}|\boldsymbol{\theta_k}), \tag{1}$$

where $\Theta = (\boldsymbol{\theta_k})_{k=1}^{K}$ and $\boldsymbol{\lambda} = (\lambda_k)_{k=1}^{K})$.

Pattern model

Knowing the mixture component parameter $\boldsymbol{\lambda}$, the component indicator $\boldsymbol{w_i} = (w_{i1}, \ldots, w_{iK})$ for the sample i is thus distributed as Multin($\boldsymbol{\lambda}$). Finally, the joint distribution is derived as

$$p(\mathrm{X}, \mathrm{W}|\Theta, \boldsymbol{\lambda}) = p(\mathrm{W}|\boldsymbol{\lambda})p(\mathrm{X}|\mathrm{W}, \Theta) \tag{2}$$

$$= \sum_{k=1}^{K} \lambda_k \prod_{i=1}^{n} p_k(\boldsymbol{x_i}|\boldsymbol{\theta_k})^{w_{ik}}. \tag{3}$$

Pattern model

Knowing the mixture component parameter $\boldsymbol{\lambda}$, the component indicator $\boldsymbol{w_i} = (w_{i1}, \dots, w_{iK})$ for the sample i is thus distributed as Multin($\boldsymbol{\lambda}$). Finally, the joint distribution is derived as

$$p(\mathrm{X}, \mathrm{W} | \Theta, \boldsymbol{\lambda}) = p(\mathrm{W}|\boldsymbol{\lambda}) p(\mathrm{X}|\mathrm{W}, \Theta) \tag{2}$$

$$= \sum_{k=1}^{K} \lambda_k \prod_{i=1}^{n} p_k(\boldsymbol{x_i} | \boldsymbol{\theta_k})^{w_{ik}}. \tag{3}$$

$$\begin{aligned} \boldsymbol{\lambda} | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}), \\ \boldsymbol{w_i} | \boldsymbol{\lambda} &\sim \text{Multin}(\boldsymbol{\lambda}), \\ \theta_{kj} | \boldsymbol{\beta}, \boldsymbol{\gamma} &\sim \text{Beta}(\boldsymbol{\beta}, \boldsymbol{\gamma}), \\ x_{ij} | \theta_{kj} &\sim \text{Bernoulli}(\theta_{kj}). \end{aligned} \tag{4}$$

The BFP algorithm

BFP algorithm consists mainly of three steps:

- Given $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$, fit the bernoulli mixture model on each subset to find the set of optimal parameter $\Gamma_i = (\Theta_i, \boldsymbol{\lambda_i}, K)$ associated with label i.

The BFP algorithm

BFP algorithm consists mainly of three steps:

- Given $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$, fit the bernoulli mixture model on each subset to find the set of optimal parameter $\Gamma_i = (\Theta_i, \boldsymbol{\lambda_i}, K)$ associated with label i.
- For a pattern $x \in \mathcal{E}$ compute the ratio

$$r(x) = \frac{p(\mathcal{M}_1 \mid x)}{p(\mathcal{M}_0 \mid x)} \tag{5}$$

$$= \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)} \times \frac{p(x \mid \Gamma_1)}{p(x \mid \Gamma_0)}. \tag{6}$$

The BFP algorithm

BFP algorithm consists mainly of three steps:

- Given $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$, fit the bernoulli mixture model on each subset to find the set of optimal parameter $\Gamma_i = (\Theta_i, \boldsymbol{\lambda_i}, K)$ associated with label i.
- For a pattern $x \in \mathcal{E}$ compute the ratio

$$r(x) = \frac{p(\mathcal{M}_1 \mid x)}{p(\mathcal{M}_0 \mid x)} \quad (5)$$

$$= \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)} \times \frac{p(x \mid \Gamma_1)}{p(x \mid \Gamma_0)}. \quad (6)$$

- The best discriminative pattern are then appended as a variable in the feature space on which any classifier can be trained.

Table: Test Accuracy, Recall and AUC 10× cross-validated for bpfd, pf and bc classifiers (with grid-search hyperparameter tuning) for benchmark datasets.

	X Gradient Boosting			Random Forest			Light Gradient-Boosting Machine			Categorical Boosting			Linear Regression			k-Nearest Neighbors		
	BC	PF	bpfd	BC	PF	bpfd	BC	PF	bpfd	BC	PF	bpfd	BC	PF	bpfd	BC	PF	bpfd
ijcnn1																		
AUC	0.728	0.769	**0.927**	0.726	0.767	**0.913**	0.732	0.769	**0.926**	0.727	0.768	**0.927**	0.714	0.732	**0.899**	0.614	0.643	**0.841**
Accuracy	0.906	0.907	**0.929**	0.906	0.907	**0.928**	0.906	0.907	**0.929**	0.906	0.907	**0.93**	0.905	0.905	**0.918**	0.89	0.897	**0.922**
Recall	0.0398	0.0465	**0.403**	0.0411	0.0479	**0.416**	0.0238	0.0372	**0.401**	0.0413	0.0474	**0.407**	0	0.0002	**0.245**	0.106	0.105	**0.419**
F1	0.0742	0.0862	**0.519**	0.0762	0.0885	**0.523**	0.0455	0.0702	**0.516**	0.0765	0.0877	**0.523**	0	0.0003	**0.362**	0.154	0.16	**0.505**
cod-rna																		
AUC	0.776	0.496	**0.815**	0.776	0.496	**0.815**	0.776	0.496	**0.815**	0.776	0.496	**0.815**	0.765	0.495	**0.813**	0.706	0.5	**0.764**
Accuracy	0.718	0.667	**0.775**	0.718	0.667	**0.775**	0.717	0.667	**0.775**	0.718	0.667	**0.775**	0.713	0.667	**0.774**	0.688	0.591	**0.739**
Recall	**0.588**	0	0.383	**0.585**	0	0.386	**0.592**	0	0.384	**0.588**	0	0.384	**0.512**	0	0.364	0.483	0.231	**0.516**
F1	**0.581**	0	0.532	**0.58**	0	0.534	**0.583**	0	0.532	**0.581**	0	0.532	**0.544**	0	0.518	0.503	0.263	**0.568**
a9a																		
AUC	0.89	**0.896**	0.88	0.863	0.869	**0.875**	0.894	0.9	**0.903**	0.894	0.9	**0.904**	0.893	0.902	**0.902**	0.837	0.848	**0.85**
Accuracy	0.841	0.844	**0.846**	0.825	0.826	**0.829**	0.844	0.846	**0.849**	0.844	0.847	**0.848**	0.841	**0.849**	0.847	0.817	**0.826**	0.824
Recall	0.597	0.604	**0.615**	0.564	**0.582**	0.578	0.606	0.613	**0.626**	0.595	0.606	**0.611**	0.581	**0.611**	0.604	0.566	0.584	**0.589**
F1	0.643	0.649	**0.658**	0.607	0.616	**0.619**	0.651	0.656	**0.666**	0.646	0.654	**0.66**	0.637	**0.659**	0.655	0.597	0.616	**0.617**
Doors																		
AUC	0.707	0.691	**0.736**	0.713	0.707	**0.753**	0.706	0.697	**0.739**	0.722	0.715	**0.749**	0.635	0.629	**0.637**	0.557	**0.574**	0.574
Accuracy	0.643	0.629	**0.679**	0.655	0.645	**0.686**	0.647	0.637	**0.681**	0.663	0.657	**0.684**	**0.6**	0.592	0.597	0.546	**0.551**	0.551
Recall	0.614	0.608	**0.642**	0.594	0.585	**0.608**	0.595	0.577	**0.619**	0.569	0.56	**0.592**	0.652	**0.674**	0.648	**0.545**	0.526	0.526
F1	0.632	0.62	**0.667**	0.632	0.622	**0.659**	0.627	0.613	**0.66**	0.627	0.619	**0.652**	0.62	**0.623**	0.617	**0.545**	0.539	0.539

Discussion

Advantages

- Approach is fast to infer and evaluate;
- Allow to easily obtain confidence bound;
- Can use expert-knowledge in the prior setting.

Possible improvement

- We could improve the model by using a non parametric approach for the bernoulli mixture model using bread stick approach to replace the choice of K;
- Even though efficient, the EM algorithm could be replace with variational inference approach in order to speed up the inference phase;
- Other discriminative score could be more suited given the use case at hand.

Bibliography

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. "Mining Association Rules between Sets of Items in Large Databases". In: *In: Proceedings of the 1993 Acm Sigmod International Conference on Management of Data, Washington Dc (Usa*. 1993, pp. 207–216.

[2] Andrew Gelman et al. *Bayesian Data Analysis, Third Edition*. en. CRC Press, Nov. 2013.

[3] Leonardo Pellegrina, Matteo Riondato, and Fabio Vandin. "SPuManTE: Significant Pattern Mining with Unconditional Testing". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 1528–1538.