AutoMS:
Automatic Model Selection for Novelty Detection with Error Rate Control

Yifan Zhang (Nankai University),
Haiyan Jiang (MBZUAI & Baidu Research),
Haojie Ren (Shanghai Jiao Tong University),
Changliang Zou (Nankai University),
Dejing Dou (Baidu Research)

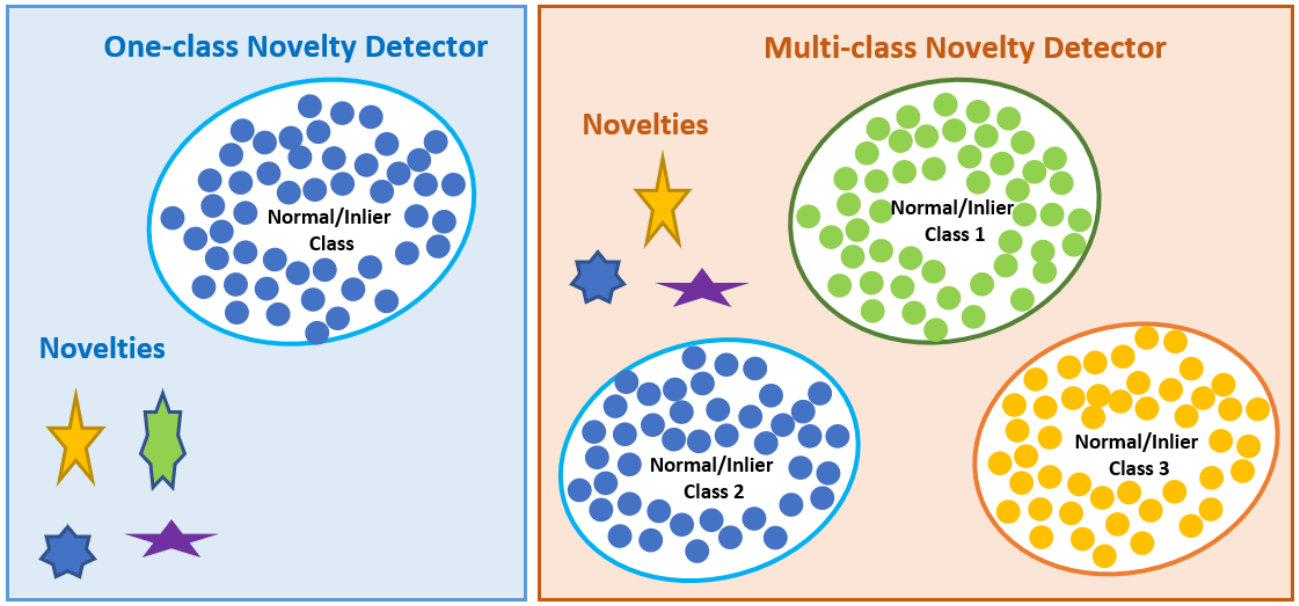
1 Introduction

Novelty Detection

A **clean** training dataset of inliers $\mathcal{D} = \{\widetilde{Z}_j\}_{j=1}^{m}$, i.e. $\widetilde{Z}_j \sim P_0$

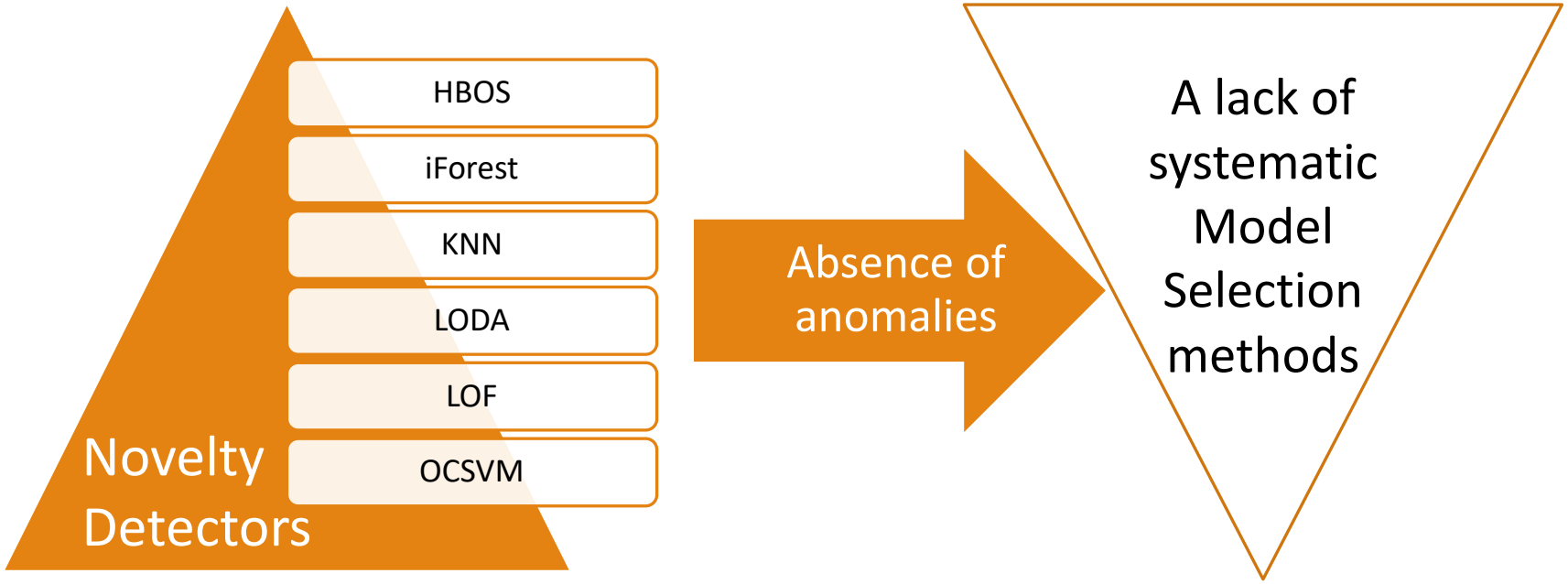
A new unseen test dataset $\mathcal{U} = \{Z_i\}_{i=1}^{n}$.

The **outlier** set $\mathcal{O} = \{Z_i \in \mathcal{U}: Z_i \text{ is not drawn from } P_0\}$



Why Model Selection

No single novelty detection methods can universally outperform others on all tasks.



Why error rate Control

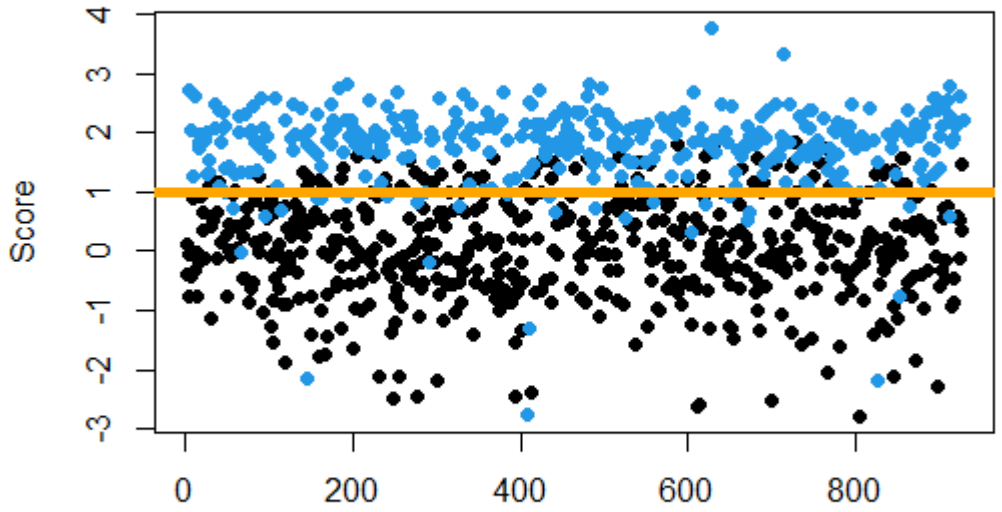
Score function $S(\cdot)$ based on Training dataset $\mathcal{D}$

Score $S_i = S(Z_i)$ to each Test data point $Z_i \in \mathcal{U}$

Decision set

outlier $\hat{\mathcal{O}}$: $S(Z_i) > L$

inlier $\hat{\mathcal{O}}^C$: $S(Z_i) \leq L$



Translate the novelty detection problem into a multiple testing problem:

$H_{0i}: Z_i \sim P_0$ v.s. $H_{1i}: Z_i \nsim P_0 \quad i = 1, \cdots, n.$

Goal of AutoMS

Select the best model $\mathcal{M}^*$ with largest **true discovery rate(TDR)** from a candidate detector set $\mathcal{G} = \bigcup \mathcal{M}$,

Control the **false discovery rate(FDR)** of $\mathcal{M}^* \in \mathcal{G}$ under a target level α.

	Detected outlier $\widehat{\mathcal{O}}_{\mathcal{M}}$	Detected inlier $\widehat{\mathcal{O}}^c_{\mathcal{M}}$	
True outlier $\mathcal{O}$	$\widehat{\mathcal{O}}_{\mathcal{M}} \cap \mathcal{O}$	$\widehat{\mathcal{O}}^c_{\mathcal{M}} \cap \mathcal{O}$	$\mathrm{TDP} = \frac{\lvert\widehat{\mathcal{O}}_{\mathcal{M}} \cap \mathcal{O}\rvert}{\lvert\mathcal{O}\rvert}$
True inlier $\mathcal{O}^c$	$\widehat{\mathcal{O}}_{\mathcal{M}} \cap \mathcal{O}^c$	$\widehat{\mathcal{O}}^c_{\mathcal{M}} \cap \mathcal{O}^c$	
	$\mathrm{FDP} = \frac{\lvert\widehat{\mathcal{O}}_{\mathcal{M}} \cap \mathcal{O}^c\rvert}{1 \vee \lvert\widehat{\mathcal{O}}_{\mathcal{M}}\rvert}$		

$$\mathrm{FDR} = \mathbb{E}(\mathrm{FDP}),\ \mathrm{TDR} = \mathbb{E}(\mathrm{TDP}).$$

2 Methodology

FDR control

- Benjamini-Hochberg (BH) procedure[14-16] to the p-values $G_{\mathcal{M}}(t)$

  where $G_{\mathcal{M}}(t) = \mathbb{P}_{\mathbb{H}_0}(S_{\mathcal{M}}(Z_i) \geq t|\mathcal{D})$ is the p-value of scores $S_{\mathcal{M}}(Z_i)$ for $Z_i \in \mathcal{U}$ when null hypothesis H_0 holds.

- The p-value $G_{\mathcal{M}}(t)$ is unknown as the null distribution of scores $S_{\mathcal{M}}(Z_i)$ is unknown.

- Bates et al.(2021): single-random-splitting(SRS) based method to estimate p-value $G_{\mathcal{M}}$

$$\widehat{G}_{\mathcal{M}}(t) = \frac{|\{\tilde{Z}_j \in \mathcal{D}_2 : S_{1,\mathcal{M}}(\tilde{Z}_j) \geq t\}|}{|\mathcal{D}_2|}$$

(1) the estimation of the p-values depends on the random split;
(2) different split ratios ($|\mathcal{D}_1|$: $|\mathcal{D}_2|$) lead to different results

Jackknife estimate

Jackknife technique to estimate p-value $G_{\mathcal{M}}(t)$

$$G_{m,\mathcal{M}}(t) = \frac{|\{\tilde{Z}_j \in \mathcal{D} : S_{\mathcal{M}}^{[-j]}(\tilde{Z}_j) \geq t\}|}{m}.$$

where $S_{\mathcal{M}}^{[-j]}(\cdot)$ is the score function trained on $\mathcal{D}^{[-j]} = \mathcal{D}\backslash\{\tilde{Z}_j\}$, $\mathcal{D}^{[-j]}$ is the subset of $\mathcal{D}$ with the j-th observation removed.

The corresponding threshold is determined via

$$L_{\mathcal{M}} = \inf\left\{0 < t < \bar{t}_{m,\mathcal{M}} : \frac{nG_{m,\mathcal{M}}(t)}{|\{Z_i \in \mathcal{U} : S_{\mathcal{M}}(Z_i) \geq t\}| \vee 1} \leq \alpha\right\}, \tag{4}$$

The decision set is $\widehat{\mathcal{O}}_{\mathcal{M}} = \{Z_i \in \mathcal{U} : S_{\mathcal{M}}(Z_i) \geq L_{\mathcal{M}}\}$.

Model Selection

AutoMS selects the "best" detector $\mathcal{M}^*$ via

$$\mathcal{M}^* = \arg\max_{\mathcal{M}\in\mathcal{G}} |\widehat{\mathcal{O}}_{\mathcal{M}}|. \tag{6}$$

Rationale:

$$\text{FDP} = \frac{|\widehat{\mathcal{O}}_{\mathcal{M}} \bigcap \mathcal{O}^c|}{1 \vee |\widehat{\mathcal{O}}_{\mathcal{M}}|} \approx \alpha \quad \Longrightarrow \quad \text{TDP} = \frac{|\widehat{\mathcal{O}}_{\mathcal{M}} \bigcap \mathcal{O}|}{|\mathcal{O}|} \approx \frac{(1-\alpha)|\widehat{\mathcal{O}}_{\mathcal{M}}|}{|\mathcal{O}|} \tag{R1}$$

Since the number of true outliers $|\mathcal{O}|$ and target FDR level α are fixed,

Select the detector $\mathcal{M}^*$ with **largest** $\left|\widehat{\mathcal{O}}_{\mathcal{M}^*}\right|$, leads to the "best" detector $\mathcal{M}^*$ with roughly the **largest TDP**.

Algorithm 1 AutoMS: Automatic Model Selection

Input: New dataset $\mathcal{U}$, clean dataset $\mathcal{D}$, target FDR level α and a pool of detectors $\mathcal{G}$.
for $\mathcal{M} \in \mathcal{G}$ **do**
  for $j = 1, \cdots, m$ **do**
    Train score function $S_{\mathcal{M}}^{[-j]}(\cdot)$ based on $\mathcal{D}^{[-j]}$ (i.e. $\mathcal{D}^{[-j]} = \mathcal{D} \backslash \{\tilde{Z}_j\}$);
    Compute the score at $\tilde{Z}_j \in \mathcal{D}$, i.e., $S_{\mathcal{M}}^{[-j]}(\tilde{Z}_j)$;
  end for
  Learn score function $S_{\mathcal{M}}(\cdot)$ based on $\mathcal{D}$;
  Compute the scores $S_{\mathcal{M}}(Z_i)$ for any $Z_i \in \mathcal{U}$, and find a threshold $L_{\mathcal{M}}$ by Eq. (4);
  The detected set is $\widehat{\mathcal{O}}_{\mathcal{M}} = \{Z_i \in \mathcal{U} : S_{\mathcal{M}}(Z_i) \geq L_{\mathcal{M}}\}$.
end for
Select the "best" detector $\mathcal{M}^*$ from $\mathcal{G}$ via Eq. (6).
Output: $\widehat{\mathcal{O}}_{\mathcal{M}^*} = \{Z_i \in \mathcal{U} : S_{\mathcal{M}^*}(Z_i) \geq L_{\mathcal{M}^*}\}$.

3 Statistical Guarantees

Theorem 1 (FDR control). *Suppose Assumptions 1–3 hold. Let* $0 < \delta < 1$ *and* $0 < \alpha < 1$. *There exist universal constants* $C_1 > 0$ *and* $C_2 > 0$ *so that the FDP of the proposed method satisfies*

$$\mathrm{FDP}(L_{\mathcal{M}^*}) \leq \alpha \left[1 + \frac{4nB_m}{A_n} + C_1 \sqrt{\frac{\varpi_m}{\delta A_n^{2/3}}} + C_2 \sqrt{\frac{\varpi_m W_{mn}}{\delta}} \right], \tag{7}$$

with probability at least $1 - 2\delta$, *and*

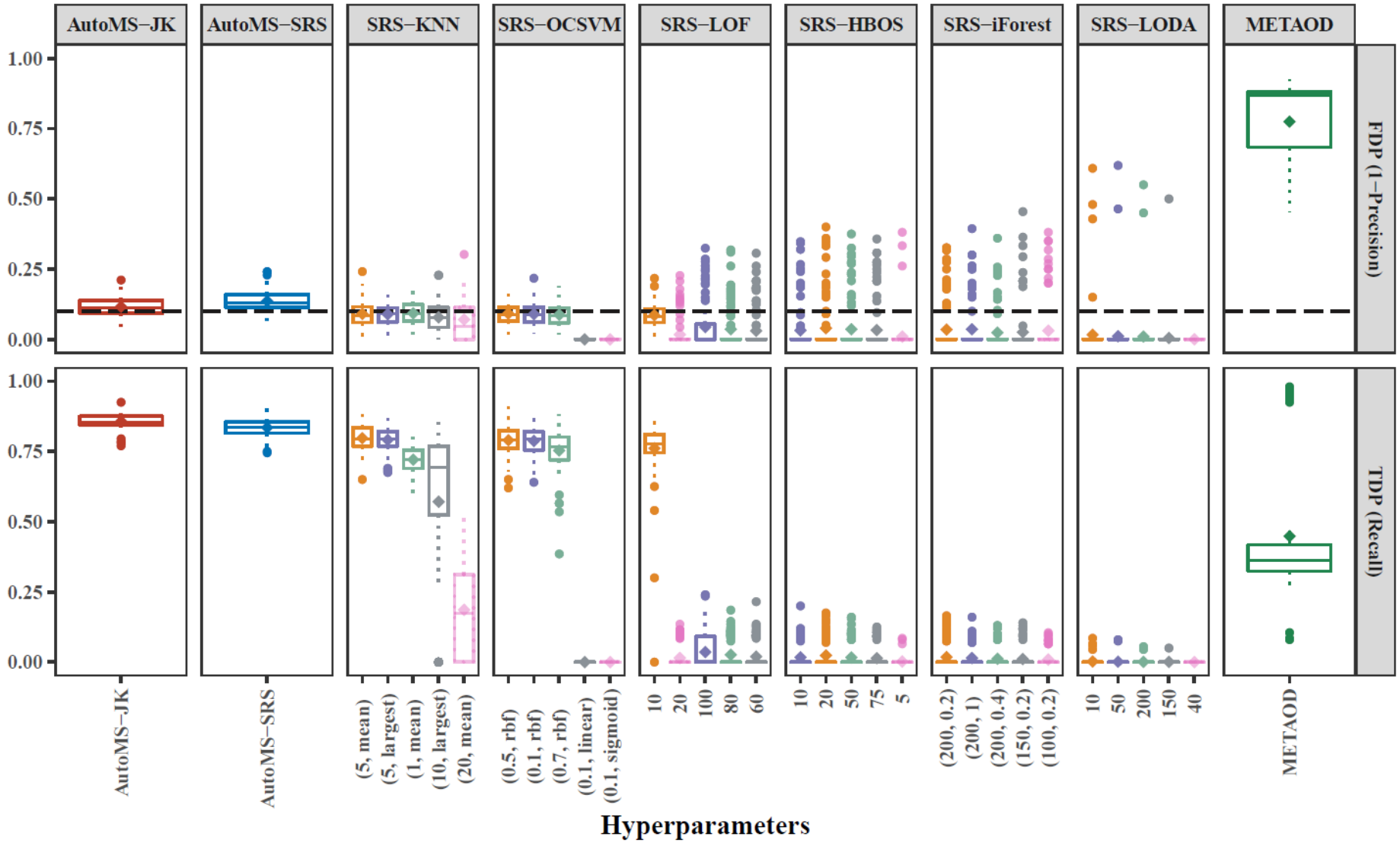
$$\limsup_{(m,n)\to\infty} \mathrm{FDR}(L_{\mathcal{M}^*}) \leq \alpha, \tag{8}$$

where $W_{mn} = \left(\frac{n}{mA_n^{2/3}} + \frac{n}{A_n^{4/3}} \right) (1 + 15B_m + 2m^2 B_m)$.

Theorem 1 shows that the selected "best" detector have **non-asymptotic bounds for FDP** and AutoMS procedure yields valid **FDR** control under some mild conditions.

4 Experiments

Figure 2: Comparisons of FDP and TDP on synthetic data. The dashed line is the target FDR level $\alpha = 0.1$.



5 Conclusion

Main Contributions of AutoMS

- The proposed AutoMS can select the best model and simultaneously control the error rate of the best model. Notably, AutoMS is a unified data-driven framework, and it does not rely on any labeled anomalous data for model selection.
- To our best knowledge, it is the **first effort to select a "best" model/detector with theoretical guarantee in the view of FDR control**. We establish non-asymptotic bounds for the FDP and show that the proposed AutoMS yields valid FDR control.
- The AutoMS can be easily coupled with commonly used novelty detection algorithms. Extensive numerical experiments indicate that AutoMS outperforms other methods significantly, with respect both error rate control and detection power.

Thank you for listening!