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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Given an unsupervised novelty detection task on a new dataset, how can we automatically select a "best" detection model while simultaneously controlling the error rate of the best model? For novelty detection analysis, numerous detectors have been proposed to detect outliers on a new unseen dataset based on a score function trained on available clean data. However, due to the absence of labeled data for model evaluation and comparison, there is a lack of systematic approaches that are able to select a "best" model/detector (i.e., the algorithm as well as its hyperparameters) and achieve certain error rate control simultaneously. In this paper, we introduce a unified data-driven procedure to address this issue. The key idea is to maximize the number of detected outliers while controlling the false discovery rate (FDR) with the help of Jackknife prediction. We establish non-asymptotic bounds for the false discovery proportions and show that the proposed procedure yields valid FDR control under some mild conditions. Numerical experiments on both synthetic and real data validate the theoretical results and demonstrate the effectiveness of our proposed AutoMS method.

1 Introduction

With the development of big data, novelty detection has become critical for many machine learning applications, such as network intrusion detection [1], fraud detection in finance [2, 3], event detection in the earth sciences [4], and so on. For these applications, there are a very large number of "normal" instances (i.e., inlier) but a lack of "abnormal" data (i.e., novelty, or outlier), because collecting abnormal data is much more expensive or difficult than collecting normal data [5]. This leads us to a situation where the available historical dataset contains only normal observations, and the basic goal is to detect outliers on a new unseen test dataset.

Novelty detection has achieved remarkable success in many application domains with numerous methods/algorithms being developed in machine learning literature [6–8], however, two problems need to be addressed in the current novelty detection models, (1) lack of error rate control in detection algorithms, and (2) ad-hoc model selection of algorithms/hyperparameters.

The most common way to solve the novelty detection problem is to use a detection algorithm to learn a plausible score function that assigns a score to each point in the test dataset. Generally, data with large scores are more likely to be detected as outliers. A decision threshold is given to determine whether a data point is an outlier or not. There is an obvious trade-off between precision and recall: a small threshold would detect more outliers but make more false discoveries as well. Therefore, a more viable way is to reformulate the novelty detection problem as a multiple hypothesis testing problem, and adopt some error rate control procedure as suitable criterion to choose the threshold [9].

Model selection (including algorithm selection or hyperparameter tuning) is a well-known feature of these novelty detection algorithms and arises naturally since none of these methods can universally

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

outperform others on *all* tasks as noted in Wolpert and Macready [10], Zhao et al. [11]. Due to the special setting of novelty detection, model evaluation and model comparison are not feasible when only "clean" data are provided as training data. Those strategies are usually ad-hoc [11, 12], which may greatly hamper their applicability to novelty detection in general cases. Furthermore, these methods only target the recall through model selection without considering the false discoveries. Therefore, it is desirable to develop a unified framework to select the "best" detector with certain error rate control.

We see that designing a data-driven procedure to select the "best" model/detector while simultaneously achieving some error rate control remains an open challenge. In this paper, we propose a data-driven procedure, AutoMS, to implement automatic model selection for novelty detection with error rate control. We apply the Jackknife technique [13] to fully explore the clean data and avoid the randomness caused by data-splitting. With the help of Jackknife prediction, we select from a pool of diversified models/detectors the "best" detector that maximizes the number of detected outliers while controlling the false discovery rate. The theoretical guarantees for the FDR control of the selected "best" detector requires further investigation because the error distribution of the "best" detector is different from that of a given detector. We address this issue by carefully studying the bound of false discovery proportion (FDP) and FDR which is calculated by taking expectation of FDP.

To our best knowledge, this is the first work to propose a unified framework to systematically select a "best" model while achieving certain error rate control. The main contributions of this work are summarized as follows:

- 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.

2 Related Work

FDR Control. Reformulating the outlier detection problem as a multiple testing problem, a commonly-used error rate in this setting is the false discovery rate (FDR): the expected proportion of true inliers among the test points reported as outliers. The FDR control [14–16] has been fully studied in the statistical literature and is a particularly useful tool to maintain the ability to reliably detect true alternatives without excessive false discovery results when large-scale hypotheses are simultaneously tested.

Along this line, Bates et al. [17] pioneered a method to split the clean dataset into two independent parts, one for training the score function and one for validating the empirical cumulative distribution function of the score function which will be used for constructing p-values for the new observations. They showed that the FDR control can be achieved with theoretical finite sample guarantees. It involves, however, a one-time random split of the clean dataset, which leads to a "p-value lottery" problem. That is, those p-values based on sample-splitting are sensitive and difficult to be reproduced in practice. Moreover, the procedure in Bates et al. [17] can only control the FDR for a given novelty detection model, and no model efficiency is considered in their paper.

Model Selection in Novelty Detection. As described in Section 1, various detection algorithms have been developed for different problems or different types of data, such as LOF [18], one-class SVM [19], iForest[20] and so on. More discussions can be found in Pimentel et al. [6], Aggarwal [21]. However, no universal learning model performs well on all problem instances [10]. Therefore, it raises the problem of model selection, including algorithm selection and hyperparameter tuning. For example, AutoOD [12] proposed an automated outlier detection framework focusing on automatic neural architecture search. Recent work METAOD [11] built a measure for task similarity and use meta-learning to select the most similar task compared to the historical ones. Most existing work

is usually based on sophisticated black-box machine learning algorithms to learn the space of the normal data. While generally effective in practice, they rarely offer a clear guarantee about the output.

We note that some statistical criteria can be used for model selection in novelty detection; e.g. using mass-volume and excess-mass [22] or some clustering validation metrics [23] to measure the quality of a score function, and using IREOS [24] to assess the separability. However, Ma et al. [25] indicated that those criteria are not useful in practice and cannot lead to a universal detector for general cases. Our proposed AutoMS is not limited to one specific algorithm or model, but can select the "best" one among any set of models/detectors.

3 Methodology: Automatic Model Selection

In an novelty detection problem, we are given a *clean* training dataset of inliers $\mathcal{D} = \{\tilde{Z}_j\}_{j=1}^m$ containing m independently and identically distributed (i.i.d.) observations $\tilde{Z}_j \sim \mathcal{P}_0$. The goal of novelty detection problem is to identify the outlier set $\mathcal{O} = \{Z_i \in \mathcal{U} : Z_i \text{ is } \textit{not} \text{ drawn from } \mathcal{P}_0\}$ among a new unseen test dataset $\mathcal{U} = \{Z_i\}_{i=1}^n$.

One of the most common approaches in addressing novelty detection is to employ a detection algorithm $\mathcal{M}$ to learn a plausible score function $S_{\mathcal{M}}(\cdot)$ with the clean dataset $\mathcal{D}$ and use the score $S_{\mathcal{M}}(Z_i)$ to assess whether $Z_i \in \mathcal{U}$ is an outlier. Generally, it is assumed that the larger the score $S_{\mathcal{M}}(Z_i)$, the more likely Z_i is an outlier. Hence, the decision set (detected outlier set) is selected as: $\widehat{\mathcal{O}}_{\mathcal{M}} = \{Z_i \in \mathcal{U} : S_{\mathcal{M}}(Z_i) \geq L_{\mathcal{M}}\}$, where $L_{\mathcal{M}}$ is a threshold associated with detector $\mathcal{M}$. Denote the subset of $\mathcal{U}$ containing all inliers as $\mathcal{O}^c = \{Z_i \in \mathcal{U} : Z_i \text{ is drawn from } \mathcal{P}_0\}$ with size n_0, and $\mathcal{O} \bigcup \mathcal{O}^c = \mathcal{U}$. Let $|\mathcal{A}|$ denote the number of the elements in set $\mathcal{A}$. Denote $a \vee b = \max\{a, b\}$.

Taking a multiple testing perspective, the null hypothesis is that a new observation is an inlier, and the alternative asserts it is an outlier. That is, the novelty detection problem is translated into a multiple testing problem as follows,

$$\mathbb{H}_{0i} : Z_i \sim \mathcal{P}_0 \text{ v.s. } \mathbb{H}_{1i} : Z_i \nsim \mathcal{P}_0 \quad i = 1, \cdots, n. \tag{1}$$

3.1 Problem Definition

Problem (Model Selection of Unsupervised Novelty Detection) The model selection problem for unsupervised novelty detection problem can be stated as follows.

Given a training dataset of inliers $\mathcal{D} = \{\tilde{Z}_j\}_{j=1}^m$, a new unlabeled dataset $\mathcal{U} = \{Z_i\}_{i=1}^n$ which may contain outliers, and all models $\mathcal{G} = \bigcup \mathcal{M}$;

Select a best model $\mathcal{M}^* \in \mathcal{G}$, *such that* the FDR of $\mathcal{M}^*$ is controlled.

The false discovery proportion (FDP) and true discovery proportion (TDP) are defined as follows

$$\text{FDP} = \frac{|\widehat{\mathcal{O}}_{\mathcal{M}} \bigcap \mathcal{O}^c|}{1 \vee |\widehat{\mathcal{O}}_{\mathcal{M}}|}, \quad \text{TDP} = \frac{|\widehat{\mathcal{O}}_{\mathcal{M}} \bigcap \mathcal{O}|}{|\mathcal{O}|}. \tag{2}$$

The false discovery rate (FDR) is the expectation of FDP, i.e. $\text{FDR} = \mathbb{E}(\text{FDP})$. The true detection rate (TDR) is defined as $\text{TDR} = \mathbb{E}(\text{TDP})$. A detector is superior to another if the former has a larger TDR for the same FDR level.

3.2 Construction of AutoMS

We begin with the FDR control for one given detector $\mathcal{M}$ and then give the AutoMS of how to select the "best" detector from $\mathcal{G}$. For one detector $\mathcal{M}$, we learn a score function $S_{\mathcal{M}}(\cdot)$ based on the historical clean dataset $\mathcal{D}$ and evaluate the score $S_{\mathcal{M}}(Z_i)$ at the new data point $Z_i \in \mathcal{U}$. The canonical approach to FDR control in multiple testing problem is to apply the Benjamini-Hochberg (BH) procedure [14–16] to the p-values of scores $S_{\mathcal{M}}(Z_i)$ for any $Z_i \in \mathcal{U}$. That is, the data-dependent threshold L is selected as follow,

$$L = \inf\left\{t > 0 : \frac{nG_{\mathcal{M}}(t)}{|\{Z_i \in \mathcal{U} : S_{\mathcal{M}}(Z_i) \geq t\}| \vee 1} \leq \alpha\right\},$$

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

However, it is quite a challenge to estimate reliable p-values $G_{\mathcal{M}}(t)$ since $\mathcal{P}_0$ is unknown and the score function $S_{\mathcal{M}}(\cdot)$ may be derived from a sophisticated black-box machine learning algorithm. Bates et al. [17] suggested randomly splitting $\mathcal{D}$ into two disjoint parts with equal sample size, one for training the score function $S_{\mathcal{M}}(\cdot)$ and the other for estimating $G_{\mathcal{M}}(t)$ with the help of conformal inference. Whereas, this single random-splitting based method suffers from, (1) the estimation of the p-values depends on the random split; (2) different split ratios also lead to different results.

To yield more stable estimates of p-values in practice, we propose to learn the score function $S_{\mathcal{M}}(\cdot)$ on the whole clean dataset $\mathcal{D}$, where our key idea stems from the **Jackknife estimate**. Denote $\mathcal{D}^{[-j]} = \mathcal{D} \backslash \{\tilde{Z}_j\}$ as the subset of $\mathcal{D}$ with the j-th observation removed. Define $S_{\mathcal{M}}^{[-j]}(\cdot)$ as the score function trained on $\mathcal{D}^{[-j]}$. Then we can evaluate m scores $\{S_{\mathcal{M}}^{[-j]}(\tilde{Z}_j)\}_{j=1}^m$ on the whole clean dataset $\mathcal{D}$. The p-value function can be estimated by

$$G_{m,\mathcal{M}}(t) = \frac{|\{\tilde{Z}_j \in \mathcal{D} : S_{\mathcal{M}}^{[-j]}(\tilde{Z}_j) \geq t\}|}{m}. \tag{3}$$

The Jackknife technique provides an approximation of the score, $S_{\mathcal{M}}^{[-j]}(\tilde{Z}_j)$, for any $\tilde{Z}_j \in \mathcal{D}$. Under the null $\mathbb{H}_0$ that $Z_i \in \mathcal{U}$ is from $\mathcal{P}_0$, $S_{\mathcal{M}}^{[-j]}(\tilde{Z}_j)$ can be used for estimating the empirical distribution of $S_{\mathcal{M}}(Z_i)$. And the corresponding threshold $L_{\mathcal{M}}$ is determined via

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

where $\bar{t}_{m,\mathcal{M}}$ is the upper bound for t and its choice will be discussed in Section 4 in detail, and $S_{\mathcal{M}}(\cdot)$ is the same score function trained on $\mathcal{D}$. Then, the detected set is identified as

$$\widehat{\mathcal{O}}_{\mathcal{M}} = \{Z_i \in \mathcal{U} : S_{\mathcal{M}}(Z_i) \geq L_{\mathcal{M}}\}. \tag{5}$$

Further, we select the "best" detector $\mathcal{M}^*$ via

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

We will refer to our procedure as AutoMS short for *Automatic Model Selection* and it is summarized in Algorithm 1. In fact, our AutoMS consists of two-phases: offline training in lines 2-6 and online prediction in lines 7-8 of pseudo-code in Algorithm 1. Note that the running time of offline training is not critical. Meanwhile, the additional running time is the detector selection in line 10, which should incur small run-time overhead due to only maximize a finite set.

Algorithm 1 AutoMS: Automatic Model Selection

1: **Input:** New dataset $\mathcal{U}$, clean dataset $\mathcal{D}$, target FDR level α and a pool of detectors $\mathcal{G}$.
2: **for** $\mathcal{M} \in \mathcal{G}$ **do**
3: **for** $j = 1, \cdots, m$ **do**
4: Compute the score at $\tilde{Z}_j \in \mathcal{D}$, i.e., $S_{\mathcal{M}}^{[-j]}(\tilde{Z}_j)$;
5: **end for**
6: Learn score function $S_{\mathcal{M}}(\cdot)$ based on $\mathcal{D}$;
7: 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);
8: The detected set is $\widehat{\mathcal{O}}_{\mathcal{M}} = \{Z_i \in \mathcal{U} : S_{\mathcal{M}}(Z_i) \geq L_{\mathcal{M}}\}$.
9: **end for**
10: Select the "best" detector $\mathcal{M}^*$ from $\mathcal{G}$ via Eq.(6).
11: **Input:** $\widehat{\mathcal{O}}_{\mathcal{M}^*} = \{Z_i \in \mathcal{U} : S_{\mathcal{M}^*}(Z_i) \geq L_{\mathcal{M}^*}\}$.

Note that the notation $\mathcal{G}$ can be a set of different algorithms or a set of hyperparameters for one typical algorithm, or both, which fits the definition of model selection including algorithm selection and hyperparameter tuning. Note that the size of $\mathcal{G}$, $|\mathcal{G}|$, can be infinite, and it is allowed to increase as the size of $\mathcal{D}$. More details will be discussed in Section 4.

Rationale. A good detector needs to achieve a large TDR while controlling the false discoveries under the target FDR level α. Here is the intuition why the selected detector obtained by AutoMS is the "best" one. For every detector $\mathcal{M} \in \mathcal{G}$, its FDP can be approximately controlled around the target FDR level α, as the proposed AutoMS resets the corresponding novelty threshold for $\mathcal{M}$. Therefore, there are roughly $|\widehat{\mathcal{O}}_{\mathcal{M}}|(1-\alpha)$ true outliers in the reported set $\widehat{\mathcal{O}}_{\mathcal{M}}$, with the TDR being approximated by $\frac{|\widehat{\mathcal{O}}_{\mathcal{M}}|(1-\alpha)}{|\mathcal{O}|}$. Since the number of true outliers $|\mathcal{O}|$ and the target α are fixed given the novelty detection task, selecting the largest $\frac{|\widehat{\mathcal{O}}_{\mathcal{M}^*}|(1-\alpha)}{|\mathcal{O}|}$, namely largest $|\widehat{\mathcal{O}}_{\mathcal{M}^*}|$, leads to the "best" detector $\mathcal{M}^*$.

Why Jackknife not Jackknife+. Barber et al. [26] introduced the jackknife+ as a new method for constructing predictive confidence intervals. They provide non-asymptotic coverage guarantees under no assumptions beyond the training and test data being exchangeable and also study the theoretical guarantees for the original jackknife with stability assumptions (see Section 5 of their paper). Jackknife+ needs to calculate m leave-one-out predictions for one test point while the original jackknife only calculates one predicted value [26]. The model evaluation cost (counts the number of times evaluating a fitted model on a new point) of jackknife+ is mn whereas the original jackknife is of cost n, indicating that jackknife+ has a much higher computation complexity compared to the original. Thus, since most novelty detection algorithms satisfy the stability assumption, we are more inclined to apply the original jackknife technique rather than jackknife+ to detectors that meet stability assumption, and we also give theoretical guarantees to achieve the non-asymptotic bounds of FDP and maintain FDR control.

4 Statistical Performance Guarantees

In this section, we will theoretically establish NON-asymptotic bounds for FDP of the selected detector $\mathcal{M}^*$ by AutoMS and further show that it can control the FDR under the following mild assumptions. Define the number of $\varpi_m = |\mathcal{G}|$ and let $A_n = o(n)$ be a pre-specified sequence.

Assumption 4.1 (Density)**.** For any detector $\mathcal{M} \in \mathcal{G}$ and sufficiently large m, the conditional density of $S_{\mathcal{M}}(Z)$ given training set $\mathcal{D}$ is uniformly bounded above by a universal constant $c_f > 0$, where $Z \in \mathcal{U}$ is a new point.

Assumption 4.2 (Learning stability)**.** For any fixed $\mathcal{M}$, the detection $S_{\mathcal{M}}$ satisfies: for large enough m,

$$\mathbb{P}(|S_{\mathcal{M}}(Z) - \tilde{\mu}_{\mathcal{M}}(Z)| \leq \kappa_m) \geq 1 - \zeta_m,$$

with some sequences satisfying $\kappa_m = o(1)$ and $\zeta_m = o(1)$ as $m \to \infty$, and some function $\tilde{\mu}_{\mathcal{M}}$.

Assumption 4.3 (Rates)**.** Denote $B_m = 2c_f\kappa_m + 3\zeta_m$. Suppose that $B_m = o(1)$, $n\varpi_m B_m = o(1)$ and $m\varpi_m/A_n^{2/3} = o(1)$.

Our main theoretical result on the validity of the ADNER method for both FDP and FDR control is given by the next theorem.

Theorem 4.1 (FDR control)**.** *Suppose Assumptions 4.1–4.3 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 $W_{mn} = \left(\frac{n}{mA_n^{2/3}} + \frac{n}{A_n^{4/3}}\right)(1 + 15B_m + 2m^2B_m)$.

It is worth mentioning that even if all detectors control the FDR, we cannot assert without theoretical proof that the optimal detector $\mathcal{M}^*$ can still guarantee the FDR control. And Bates et al. [17] only gives **the expectation version** that FDR can be controlled for a given detector. Briefly, we first prove

that the FDP of each detector can be bounded *uniformly* through the uniform convergence of the jackknife estimator. Hence, the selected "best" detector $\mathcal{M}^*$ with the largest number of discoveries will have the same non-asymptotic bounds for FDP. Then we give the asymptotic FDR control for the optimal detector. Detailed proofs for theoretical results are provided in the **Supplementary Material**.

In Eq.(4), we specify a data-driven upper bound $\bar{t}_{m,\mathcal{M}}$ when determining $L_{\mathcal{M}}$. We define $\bar{t}_{m,\mathcal{M}} = G^{-}_{m,\mathcal{M}}(A_n/n)$ where $G_{m,\mathcal{M}}(t)$ is defined in Eq.(3) and $G^{-}_{m,\mathcal{M}}(y) = \inf\{t \geq 0 : G_{m,\mathcal{M}}(t) \leq y\}$ with $0 \leq y \leq 1$. The upper bound $\bar{t}_{m,\mathcal{M}}$ for threshold $L_{\mathcal{M}}$ can facilitate the derivation of non-asymptotic bounds for FDP but is actually not required in practice. In other words, this makes an implicit assumption that a non-trivial (not ∞) threshold can be attained within the interval $(0, \bar{t}_{m,\mathcal{M}}]$. By Theorem 4.1, we can see that a large A_n would yield faster convergence of FDP. However, too large A_n may lead to obtain a trivial threshold. In fact, how large A_n could be depends on the number of identifiable outliers in $\mathcal{U}$ [27], which is certainly unknown to us. Alternatively, by examining the requirements in Assumption 4.3, we can set $A_n \sim \min\{(m\varpi_m)^2, n^{1/2}\}$ for simplicity.

Remark 4.1. Since we use the Jackknife to estimate FDP, certain stability conditions seem to be necessary. In Assumption 4.2, $\tilde{\mu}$ needs not to be close to an optimal population detector. We only need the $S_{\mathcal{M}}$ to concentrate around $\tilde{\mu}$. This is just a stability assumption rather than consistency assumption. When the sampling stability fails to hold, the $S^{[-j]}_{j,\mathcal{M}}$ may have a substantially different distribution than $G_{n,\mathcal{M}}(t)$. In that case, we may resort to the sample-splitting procedure instead.

Remark 4.2. Although the FDP bound might be slightly larger than the target FDR level α in finite samples, the FDP bound converges asymptotically to the target FDR level α. Actually, in novelty detection tasks, it is not necessary to obtain procedures of finite FDR control with additional assumptions, instead, a powerful and stable procedure which controls FDR asymptotically is good enough for most cases.

Remark 4.3. As discussed in Assumption 4.3, the number of detectors $\varpi_m = |\mathcal{G}|$ can go to infinity with some certain rate related to the training sample size m and test sample size n.

5 Experiments and Evaluation

5.1 Experiment Setting

In this section, we evaluate the performance of the proposed AutoMS in terms of FDR control and detection power (TDP) through extensive numerical studies on both synthetic and real data. We set the target FDR level α to $\alpha = 0.1$, and we report the results based on 100 replication experiments throughout this paper. We construct a detector set $\mathcal{G}$ with 72 candidate models by combining 6 novelty detection algorithms and their corresponding hyperparameters (See **Supplementary Material Table 1** for the complete list). These models/detectors are available in the Python "**PyOD**" [28] package. Note that the number of detector set $\mathcal{G}$ can be infinite and our AutoMS can still be used.

Performance Measures. Since we focus on both error rate control and detection power, we use $\text{FDR} = \mathbb{E}(\text{FDP})$ and $\text{TDR} = \mathbb{E}(\text{TDP})$ as performance measures for each model in this paper. We say a detection procedure is superior to its competitor if it has a larger TDR for the same FDR level. According to the definitions in Eq.(2) of Section 3, we have $\text{FDP} = 1 - \text{Precision}$, and $\text{TDP} = \text{Recall}$. Correspondingly, the empirical metrics of FDR and TDR are calculated by averaging of FDPs and TDPs over 100 replications, respectively.

Table 1: Model-selection and error-rate-control algorithms.

Method	Model Selection	Error Rate Control
SRS [17]	✗	✓
METAOD [11]	✓	✗
AutoMS	✓	✓

Model Selection and Error Rate Control. To evaluate the performance of the proposed **AutoMS** method, we first introduce two benchmark procedures, **SRS** [17] and **METAOD** [11], as shown in Table 1. Then we explore the performances of the three methods with respect to FDR and TDR. The SRS-based methods combine SRS with different base detectors. For example, SRS-LOF means that we use LOF as the base detector in combination with SRS, while SRS-LOF-best means SRS-LOF method with the best FDR performance. It is worth mentioning that SRS is guaranteed with FDR control for any given detector. METAOD is a model selection method using meta-learning without considering error rate control, while the **AutoMS** can do both model selection and error rate control.

5.2 Evaluation of Jackknife

Firstly, We would like to show the advantage of Jackknife prediction with a toy example using synthetic data. We generate a training dataset $\mathcal{D}$ and a test dataset $\mathcal{U}$ from a Gaussian mixture model, where their sample sizes are $m = n = 2000$ respectively and $\mathcal{U}$ contains 10% outliers. More details of the model setting can be found in Section 5.3. In Figure 1, we compare the Jackknife prediction to the single-random-splitting (SRS) method in Bates et al. [17] with different split ratios when the detector $\mathcal{M}$ is the OCSVM with its default hyperparameter in PyOD package[28]. Focusing on one training set $\mathcal{D}$, we randomly split $\mathcal{D}$ 100 times for each split ratio of SRS method while Jackknife technique only have one prediction result. We can see from Figure 1 that the Jackknife provides a quite tight FDP control and achieves a high TDP while finding more true outliers. As discussed above, the SRS-based methods depends on the random split and the split ratio and are usually unstable due to the randomness of the segmentation, making it hard to reproduce.

The Jackknife estimate helps improve the stability and the accuracy of the scores due to the usage of more samples to train the score function. Meanwhile, the Jackknife procedure will require a bit more computation compared to one-time data splitting, if m is not too small. As the Jackknife procedure for $S_{\mathcal{M}}^{[-j]}(\cdot)$ is trained offline, one can always replace Jackknife with cross-validation to reduce the computing burden to a certain extent. Instead of omitting one observation, we can randomly partition $\mathcal{D}$ into k equal-sized folds and remove one fold from $\mathcal{D}$ as the subsample to train the score function. In the following of this section, we use $k = 100$ folds in the Jackknife procedure.

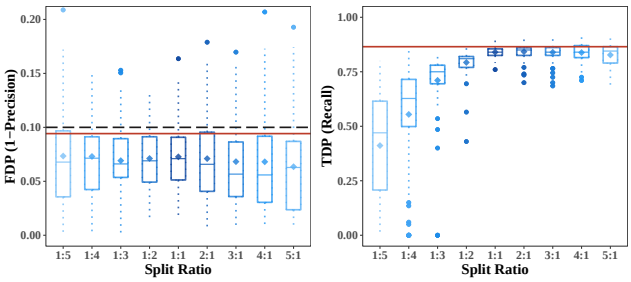

Figure 1: Performance of SRS-OCSVM [17] under different split ratios. Each box-plot shows the distribution of FDP and TDP under 100 simulations given one split ratio. The red line shows the corresponding results from Jackknife prediction. The black dashed line is the target FDR level $\alpha = 0.1$.

5.3 Synthetic Data

Data Description. We consider the same settings as in Bates et al. [17]. The data are generated by sampling each data point $Z_i \in \mathbb{R}^{50}$ from a multivariate Gaussian mixture model, $Z_i = W_i + \sqrt{a}V_i$, where W_i is one of 50 cluster center sampled independently from the uniform distribution on $[-3, 3]^{50}$ and $V_i \in \mathbb{R}^{50}$ are i.i.d from $N(0, 1)$. The difference between inliers and outliers is the constant a with $\mathbb{H}_0 : a = 1$ v.s. $\mathbb{H}_1 : a = 2$. We generate a clean dataset $\mathcal{D}$ with $m = 2000$ observations and a test dataset $\mathcal{U}$ with $n = 2000$ observations in which 90% are sampled as inliers. Our aim is to identify the remaining 10% of outliers in $\mathcal{U}$.

Results. We show the performance of the proposed AutoMS, the SRS-based methods with different hyperparameters (e.g., SRS-LOF) and METAOD in Figure 2. All the SRS-based methods can perfectly control the FDR with FDPs scattering around the target FDR level $\alpha = 0.1$. But in terms of detection power, the TDR performance of SRS-based methods vary greatly with different hyperparameters and different algorithms. Some SRS-based methods, such as SRS-LODA, SRS-HOBOS, SRS-iForest, even lose detection power with almost zero TDP/recall. This necessitates model selection among numerous novelty detection methods. On the other hand, METAOD gives a large FDR, which means that most of the reported points are not true outliers. This phenomenon can be explained as METAOD does not employ any FDR control procedure. And the TDR of METAOD does not preform well because the model selection based on historical data may not be suitable for the data we use. AutoMS is superior to all SRS-based methods and METAOD with respect to FDR and TDR, in the sense that AutoMS gives the highest TDR with the FDR controlled. It is clear that AutoMS is able to control the FDR and performs more stably compared to SRS-based methods. Additionally, AutoMS always achieves better detection power with a larger TDP compared to SRS-based methods.

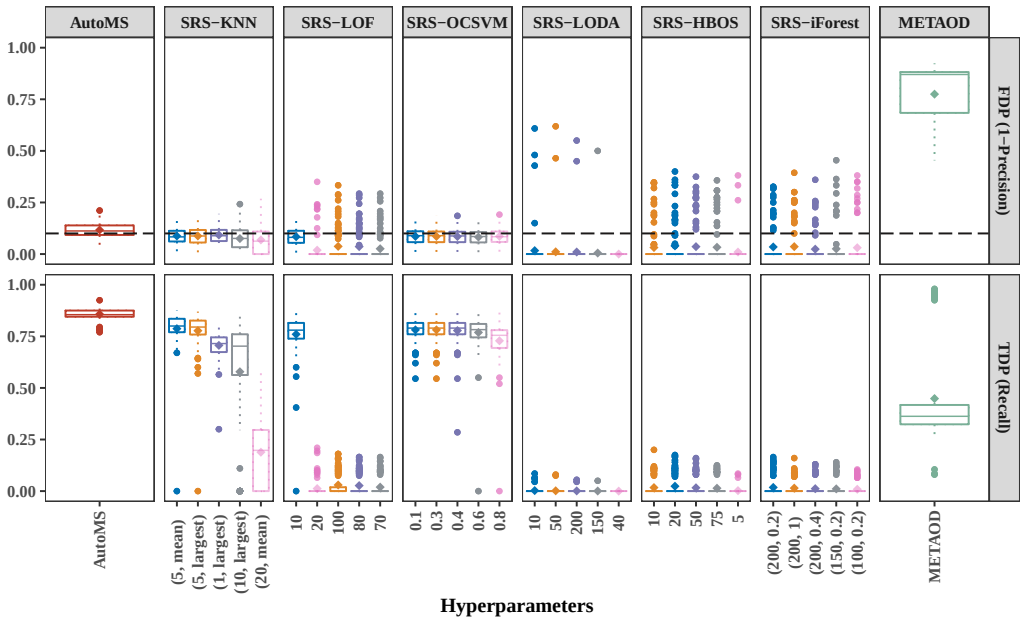

Figure 2: Comparisons of FDP (more stable around the dashed line) and TDP (the larger the better) among AutoMS, SRS-based methods (with different hyperparameters) and METAOD. The dashed line is the target FDR level $\alpha = 0.1$.

Figure 3 shows the performance of AutoMS and SRS-based-best methods under different target FDR levels. The SRS-based-best methods includes SRS-KNN-best, SRS-LOF-best, SRS-OCSVM-best. In particular, SRS-LOF-best is the best SRS-LOF detector with the largest TDR among all its candidate hyperparameters. All the SRS-based-best methods can control the FDR under the target level α. The empirical FDR of AutoMS is slightly larger than the given target FDR level α, but their difference does not increase as α becomes larger. The TDR of AutoMS is higher than all the SRS-based-best methods at every target FDR level. Actually, in novelty detection tasks, it is not necessary to obtain procedures with finite FDR control, instead, a powerful and stable procedure that controls FDR asymptotically around the target α is good enough for most cases. That is what AutoMS does, in the sense that AutoMS can uniformly offer much higher TDRs than SRS-based-best methods under different target FDR levels.

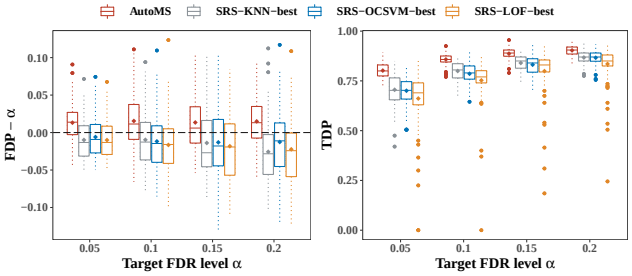

Figure 3: Performance of AutoMS and SRS-based-best methods with different target FDR level α. The difference between the FDP and the target FDR level α shows the deviation of FDP from α.

5.4 Real Data

We compare the proposed AutoMS with two different types of baselines: (i) model selection algorithms and (ii) error-rate-controlled algorithms, as in Table 1. For model selection algorithms, we use METAOD [11] as a representative, since it deals with automatic unsupervised outlier model selection which uses the performances of historical outlier detection benchmark datasets as prior experience to automatically do model selection via meta learning. For error-rate-controlled algorithms, we choose SRS-based methods [17] as typical algorithms to give statistical guarantees for the error rate control, but there is no model selection considered in SRS-based methods.

Table 2: Summary of real-world datasets for novelty detection.

	Credit Card[29]	Covertype[30]	Shuttle[31]	Satellite[32]
# Features d	30	10	36	9
# Inliers N_0	284,807	283,301	5,732	45,586
# Outliers N_1	492	2747	703	12,414

Datasets. Four real-world datasets for outlier detection are considered, and the details of the datasets are summarized in Table 2. We use TDR (Recall) and FDR (1-Precision) as performance measures. For real data implementation, we focus on a test dataset generating mechanism under which 90% of the $n = 2000$ or 1000 observations in each $\mathcal{U}$ are sampled from N_0 inliers, and we seek to identify the remaining 10% of outliers. The training set $\mathcal{D}$ is constructed by randomly sampling $m = \min\{20000, N_0/3\}$ inlier samples from the total N_0 inliners.

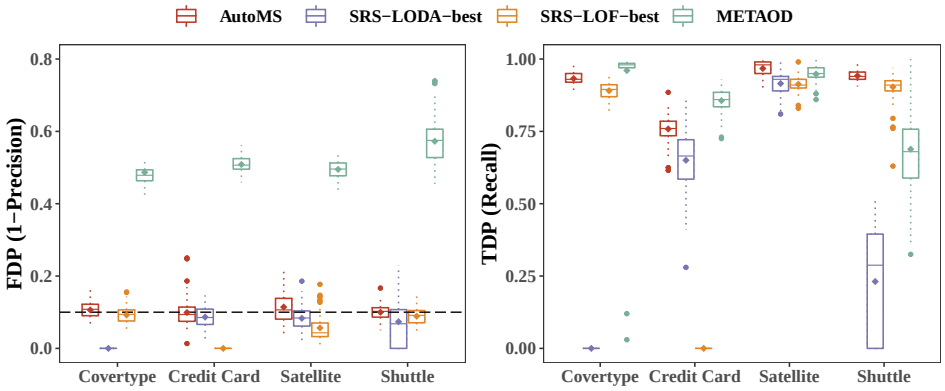

Figure 4: Performance on four different real-world datasets. Each box-plot shows the distribution of FDP and TDP. The dashed line is the target FDR level $\alpha = 0.1$.

Results. From Figure 4, we can observe that AutoMS can maintain the FDR level well around the target level $\alpha = 0.1$, while METAOD has inflated the FDR to values around or above 0.5 on all real-world datasets, indicating that in the detected set $\widehat{\mathcal{O}}$ obtained by METAOD, around 50% of the detected outliers are not true outliers. It is worth mentioning that, all SRS-based methods can maintain the FDR level around 0.1 as well as expected from their theoretical guarantees. As for the true detection rate (TDR), AutoMS has reasonable TDR on all datasets, while SRS-based methods sometimes perform too conservatively and have no detected outliers, and its performance can vary greatly from dataset to dataset. In addition, the performances of SRS-LODA-best and SRS-LOF-best are rather different, particularly on the Covertype dataset and Credit Card dataset, which means SRS-based-best methods can not give a consistent good result on different tasks and its performance can be severely affected by the base detector the SRS used. On the other hand, AutoMS keeps accurate and stable FDR levels on all datasets and has reasonable TDR for different tasks. Overall, AutoMS can uniformly offer much higher TDRs than SRS-based-best methods across all settings and that AutoMS can maintain the FDR level with a comparable TDR compared with METAOD. **Section B in Supplementary Material** summarizes the empirical FDR and TDR of different detection methods across all data sets.

6 Conclusion

The detector selection problem in novelty detection analysis is a challenging problem and it has not been thoroughly investigated in the literature, especially in terms of theoretical guarantees. In this paper, we propose an effective data-driven procedure, named AutoMS, that can select the "best" detector which offers high TDR and controls FDR simultaneously. The AutoMS is fully automatic without any supervision and is easily implemented. We give theoretical results showing that the proposed AutoMS can achieve the non-asymptotic bounds of FDP and maintain FDR control under mild conditions. Extensive experiments on synthetic and real data show AutoMS significantly improves the detection power and yields accurate FDR control.

References

[1] Hung-Jen Liao, Chun-Hung Richard Lin, Ying-Chih Lin, and Kuang-Yuan Tung. Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, 36 (1):16–24, 2013.

[2] Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, and J Christopher Westland. Data mining for credit card fraud: A comparative study. *Decision support systems*, 50(3):602–613, 2011.

[3] Tzu-Hsuan Lin and Jehn-Ruey Jiang. Credit card fraud detection with autoencoder and probabilistic random forest. *Mathematics*, 9(21):2683, 2021.

[4] Jan Verbesselt, Achim Zeileis, and Martin Herold. Near real-time disturbance detection using satellite image time series. *Remote Sensing of Environment*, 123:98–108, 2012.

[5] Rémi Domingues. Probabilistic modeling for novelty detection with applications to fraud identification. *arXiv preprint arXiv:1903.01730*, 2019.

[6] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[7] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021.

[8] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021.

[9] Quentin Rebjock, Baris Kurt, Tim Januschowski, and Laurent Callot. Online false discovery rate control for anomaly detection in time series. *Advances in Neural Information Processing Systems*, 34, 2021.

[10] David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[11] Yue Zhao, Ryan Rossi, and Leman Akoglu. Automatic unsupervised outlier model selection. *Advances in Neural Information Processing Systems*, 34, 2021.

[12] Yuening Li, Zhengzhang Chen, Daochen Zha, Kaixiong Zhou, Haifeng Jin, Haifeng Chen, and Xia Hu. Autood: Neural architecture search for outlier detection. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 2117–2122. IEEE, 2021.

[13] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[14] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, 57(1):289–300, 1995.

[15] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[16] John D Storey, Jonathan E Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

[17] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *arXiv preprint arXiv:2104.08279*, 2021.

[18] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[19] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[20] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

[21] Charu C Aggarwal. *Outlier Analysis*. Springer, 2017.

[22] Nicolas Goix. How to evaluate the quality of unsupervised anomaly detection algorithms? *arXiv preprint arXiv:1607.01152*, 2016.

[23] Thanh Trung Nguyen, Uy Quang Nguyen, et al. An evaluation method for unsupervised anomaly detection algorithms. *Journal of Computer Science and Cybernetics*, 32(3):259–272, 2016.

[24] Henrique O Marques, Ricardo JGB Campello, Jörg Sander, and Arthur Zimek. Internal evaluation of unsupervised outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(4):1–42, 2020.

[25] Martin Q Ma, Yue Zhao, Xiaorong Zhang, and Leman Akoglu. A large-scale study on unsupervised outlier model selection: Do internal strategies suffice? *arXiv preprint arXiv:2104.01422*, 2021.

[26] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

[27] Lilun Du, Xu Guo, Wenguang Sun, and Changliang Zou. False discovery rate control under general dependence by symmetrized data aggregation. *Journal of the American Statistical Association*, just-accepted:1–34, 2021.

[28] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019. URL http://jmlr.org/papers/v20/19-011.html.

[29] Credit Card. Credit card fraud detection, 2016. URL https://www.kaggle.com/mlg-ulb/creditcardfraud.

[30] Covertype. Forest covertype, 1998. URL http://odds.cs.stonybrook.edu/forestcovercovertype-dataset.

[31] Shuttle. Statlog (shuttle), 1993. URL https://archive.ics.uci.edu/ml/datasets/Statlog+(Shuttle).

[32] Satellite. Satellite image, 1993. URL http://odds.cs.stonybrook.edu/satimage-2-dataset.

[33] Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*, pages 59–63, 2012.

[34] Tomáš Pevný. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 102(2):275–304, 2016.

[35] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 427–438, 2000.

[36] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 11:2973–3009, 2010.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See at the end of Section 1, we summarize our main contributions. And the experimental results in Section 5 validate the theoretical results in Section 4.
   (b) Did you describe the limitations of your work? [Yes] See Remark 4.1 in Section 4, we discuss the stability limitation (Assumption 4.2) of detector algorithm.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is foundational research and not tied to particular applications.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We read the ethics review guidelines and ensured that our paper conforms to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumption 4.1-4.3 in Section 4.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section C in Supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Algorithm 1 in Section 3, we give the pseudo-code of our procedure AutoMS.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We describe the generation of our synthetic data in Section 5.3 and give the candidate detectors we use in Section A of Supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Our experimental results are all shown with boxplots.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use the Python "**PyOD**" package and have cited the creator. The real data we use is open source and the URLs are cited.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]