Choosing the operating threshold in evaluation of anomaly detection methods

March 19, 2020

1 Introduction

The problem of unsupervised anomaly detection in unknown conditions requires the determination of a suitable evaluation measure. It has been shown in previous experiments that two measures based on the ROC curve - partial AUC (AUC@p) and true positive rate (TPR@p) at a given false positive rate p - are more appropriate than other (e.g. AUC, F1 score, precision).

Now there is a question – when not given from the outside, what is the optimal operating false positive rate p at which we should measure TPR@p and AUC@p so that we select the best model? This depends on our definition of what is the best model. We may want to choose such a p at which the tested model performance is most distinguishable. We may want it to choose a model that is the most robust with respect to differences in testing and validation/application data.

2 Discriminability criteria

The experimental design is following. There are M different models, each with a set of $\theta_i = \{\theta_{ij}\}_{j=1}^{I_i}$ hyperparameter settings for a total of $N = \sum_{i=1}^{M} I_i$ different model/hyperparameter combinations. We split a dataset into a training and testing subsets. We train the model on the training subset and evaluate the measures on the testing subset. This training and valdiation loop is done in a k-fold validation scheme, that is the split is done k times for each basic dataset. This results in a total of $k \cdot K$ experiments. In our basic experiments, we have $M = 4, k = 10$. The models and hyperparameters settings are summarized in 1. Theres is a total of $k \sum_{i=1}^{M} I_i = 10 \cdot (27 + 3 + 3 + 9) = 420$ experiments done for each dataset.

The discriminability criteria are used to tell us how different the model performance is at different false positive rates p. We believe choosing such a value of p at which the models are most discriminable (there are the largest differences in their performance) leads to a more robust performance measure. Alternatively, we can choose the lowest value of p at which the models are already discriminable. There are different statistical tests that compare population means and variances and can be used for this.

2.1 Welch's t-test

Also called unequal variances t-test [**?**]. It is a test to decide whether two populations have equal means. It does not need the assumption of equality of variances. Also, the population sizes can be different. However, assumption of normality is still needed. The test statistic for comparing two populations (μ_i, σ_i) and (μ_j, σ_j) of sizes N_i, N_j is

$$t_{ij}^{W} = \frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2/N_i + \sigma_j^2/N_j}}. \tag{1}$$

The hypothesis H_0 is $\mu_i = \mu_j$. Under H_0, t_{ij}^{W} is from the t-distribution with ν degrees of freedom, where

$$\nu = \frac{\left(\sigma_i^2/N_i + \sigma_j^2/N_j\right)^2}{\frac{\sigma_i^4}{N_i^2(N_i-1)} + \frac{\sigma_j^4}{N_j^2(N_j-1)}}. \tag{2}$$

algorithm	hyperparameter	values
kNN	distance	$\{\kappa(x), \gamma(x), \delta(x)\}$
	k	$\{1, 3, 5, 7, 9, 13, 21, 31, 51\}$
LOF	k	$\{10, 20, 50\}$
IF	N_t	$\{50, 100, 200\}$
OCSVM	γ	$\{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100\}$

Table 1: Overview of used hyperparameter values.

max/loss	AUC	AUC_w	AUC@0.01	AUC@0.05	TPR@0.01	TPR@0.05	precision@0.01	precision@0.05	F1@0.01	F1@0.05	AUC@tukey-mean	TPR@tukey-mean	mean	mean@0.01	mean@0.05
AUC	–	0.4%	2.0%	1.2%	2.4%	1.0%	2.3%	2.0%	10.9%	7.1%	1.2%	1.1%	2.6%	4.4%	2.8%
AUC_w	0.1%	–	1.3%	0.5%	1.6%	0.6%	1.6%	1.7%	10.3%	6.7%	0.5%	0.5%	2.1%	3.7%	2.4%
AUC@0.01	0.7%	0.8%	–	1.0%	0.5%	2.1%	1.4%	2.5%	7.9%	6.3%	0.9%	1.0%	2.1%	2.5%	3.0%
AUC@0.05	0.3%	0.2%	0.7%	–	0.9%	0.4%	1.8%	1.5%	8.4%	5.3%	0.7%	0.6%	1.7%	2.9%	1.8%
TPR@0.01	0.6%	1.1%	0.5%	1.0%	–	2.2%	1.8%	2.3%	8.2%	6.2%	1.1%	0.7%	2.1%	2.6%	2.9%
TPR@0.05	0.3%	1.5%	5.7%	1.9%	4.5%	–	4.6%	2.2%	12.3%	6.2%	4.1%	1.8%	3.8%	6.8%	2.6%
precision@0.01	0.7%	1.0%	0.7%	1.5%	1.4%	2.8%	–	1.6%	12.5%	10.1%	1.4%	1.0%	2.9%	3.6%	4.0%
precision@0.05	0.5%	0.7%	2.3%	1.0%	2.2%	1.3%	2.4%	–	10.1%	6.6%	2.1%	1.3%	2.5%	4.3%	2.2%
F1@0.01	3.5%	17.7%	27.7%	16.9%	24.5%	10.1%	37.1%	13.8%	–	3.4%	26.6%	12.9%	16.2%	22.3%	11.0%
F1@0.05	3.6%	11.0%	13.1%	10.3%	12.4%	8.8%	21.2%	14.0%	6.0%	–	12.4%	9.3%	10.2%	13.2%	8.3%
AUC@tukey-mean	0.3%	0.5%	1.3%	1.1%	1.6%	1.5%	1.5%	2.7%	9.4%	6.7%	–	0.4%	2.2%	3.4%	3.0%
TPR@tukey-mean	2.4%	12.6%	26.1%	19.6%	24.4%	14.3%	23.1%	10.9%	28.0%	15.8%	21.8%	–	16.6%	25.4%	15.1%

Table 2: Means of relative loss in a column measure when optimal model and hyperparameters are selected using the row measure. 0% training contamination. Level of shading highlights three best results in a column.

We apply a two-tailed test on a confidence level α, where the critical value $t_c^W = f_q^{\mathrm{T}}(\nu, 1-\alpha/2)$, where f_q^T is the quantile function of the Student's t-distribution. If $t_{ij}^W \notin (-t_c^W, t_c^W)$, we reject H_0. The p-value of Welch statistic is

$$p^W = 1 - \mathrm{cdf}^{\mathrm{T}}(\nu, t_{ij}^W), \tag{3}$$

where $\mathrm{cdf}^{\mathrm{T}}$ is the cumulative distribution function of the Student's t-distribution. TODO: add the section for multiple comaprison as in [1].

2.2 Tukey's test

Assumes normality and homogenity of variance across groups. The test statistic is

$$q_{ij}^s = \frac{|\mu_i - \mu_j|}{\sqrt{\frac{MSW}{k}}}, \tag{4}$$

where MSW should be the mean squares within. We interpret it as the mean variance across the experiment

$$MSW = \frac{1}{N}\sum_i \sigma_i^2, \tag{5}$$

where N is the total number of populations (comapred model/hyperparameter combinations) and k is the number of samples (coming from k-fold crossvalidation). The test statistic is compared to the studentized range distribution with parameters $\nu = N(k-1)$ (degrees of freedom) and N.

3 Results

The tables 2–6 summarize an experiment that show us how robust a measure is with respect to the others. In other words, this is an answer to the question "How much worse in terms of measure A is model selected using measure B than that selected by using A?", where A is the column measure and B is the row measure. We select a model (one out of $\sum_{n=1}^N I_n$) that performs (on average over the k folds) the best on the testing dataset using the row measure. Then we look at the value of the column measure of the same model x_{base} and compare it to the best value of the column measure across all models x_{max}. Then we compute the relative measure loss $(x_{max} - x_{base})/x_{max}$. The table entries are means over all datasets.

For *welch mean, welch median* we compute all pairwise tests.

4 What are better alternatives to AUC?

It seems that the better alterantives for AUC are AUC@p and TPR@p. The first one is more preferable, since it is more robust due to being an integral. Also, it brings more discriminability at higher values of p. This is due to the fact that at a given p, it is more likely that two ROC curves will have the same value of TPR as opposed to the whole integral up to p, where even a difference in a single (FPR,TPR) pair leads to two different values of AUC@p.

What is the best weighing functions?

- universal

max/loss	AUC	AUC_w	AUC@0.01	AUC@0.05	TPR@0.01	TPR@0.05	precision@0.01	precision@0.05	F1@0.01	F1@0.05	AUC@tukey-median	TPR@tukey-median	mean	mean@0.01	mean@0.05
AUC	–	0.4%	2.0%	1.2%	2.4%	1.0%	2.3%	2.0%	10.9%	7.1%	1.0%	1.1%	2.6%	4.4%	2.8%
AUC_w	0.1%	–	1.3%	0.5%	1.6%	0.6%	1.6%	1.7%	10.3%	6.7%	0.4%	0.5%	2.1%	3.7%	2.4%
AUC@0.01	0.7%	0.8%	–	1.0%	0.5%	2.1%	1.4%	2.5%	7.9%	6.3%	0.6%	0.9%	2.1%	2.5%	3.0%
AUC@0.05	0.3%	0.2%	0.7%	–	0.9%	0.4%	1.8%	1.5%	8.4%	5.3%	0.5%	0.5%	1.7%	2.9%	1.8%
TPR@0.01	0.6%	1.1%	0.5%	1.0%	–	2.2%	1.8%	2.3%	8.2%	6.2%	0.9%	0.6%	2.1%	2.6%	2.9%
TPR@0.05	0.3%	1.5%	5.7%	1.9%	4.5%	–	4.6%	2.2%	12.3%	6.2%	2.9%	1.6%	3.7%	6.8%	2.6%
precision@0.01	0.7%	1.0%	0.7%	1.5%	1.4%	2.8%	–	1.6%	12.5%	10.1%	1.0%	0.8%	2.8%	3.6%	4.0%
precision@0.05	0.5%	0.7%	2.3%	1.0%	2.2%	1.3%	2.4%	–	10.1%	6.6%	1.7%	1.1%	2.5%	4.3%	2.2%
F1@0.01	3.5%	17.7%	27.7%	16.9%	24.5%	10.1%	37.1%	13.8%	–	3.4%	25.6%	12.6%	16.1%	22.3%	11.0%
F1@0.05	3.6%	11.0%	13.1%	10.3%	12.4%	8.8%	21.2%	14.0%	6.0%	–	10.5%	9.4%	10.0%	13.2%	8.3%
AUC@tukey-median	0.3%	0.5%	1.4%	1.1%	1.7%	1.5%	1.7%	2.7%	9.5%	6.8%	–	0.4%	2.3%	3.6%	3.0%
TPR@tukey-median	3.0%	13.2%	27.9%	21.4%	26.4%	15.9%	22.8%	12.3%	29.3%	17.2%	20.0%	–	17.5%	26.6%	16.7%

Table 3: Means of relative loss in a column measure when optimal model and hyperparameters are selected using the row measure. 0% training contamination. Level of shading highlights three best results in a column.

max/loss	AUC	AUC_w	AUC@0.01	AUC@0.05	TPR@0.01	TPR@0.05	precision@0.01	precision@0.05	F1@0.01	F1@0.05	AUC@tukey-q	TPR@tukey-q	mean	mean@0.01	mean@0.05
AUC	–	0.4%	2.0%	1.2%	2.4%	1.0%	2.3%	2.0%	10.9%	7.1%	2.3%	2.7%	2.9%	4.4%	2.8%
AUC_w	0.1%	–	1.3%	0.5%	1.6%	0.6%	1.6%	1.7%	10.3%	6.7%	1.6%	1.8%	2.3%	3.7%	2.4%
AUC@0.01	0.7%	0.8%	–	1.0%	0.5%	2.1%	1.4%	2.5%	7.9%	6.3%	0.3%	0.8%	2.0%	2.5%	3.0%
AUC@0.05	0.3%	0.2%	0.7%	–	0.9%	0.4%	1.8%	1.5%	8.4%	5.3%	1.0%	1.2%	1.8%	2.9%	1.8%
TPR@0.01	0.6%	1.1%	0.5%	1.0%	–	2.2%	1.8%	2.3%	8.2%	6.2%	0.8%	0.2%	2.1%	2.6%	2.9%
TPR@0.05	0.3%	1.5%	5.7%	1.9%	4.5%	–	4.6%	2.2%	12.3%	6.2%	5.9%	4.8%	4.2%	6.8%	2.6%
precision@0.01	0.7%	1.0%	0.7%	1.5%	1.4%	2.8%	–	1.6%	12.5%	10.1%	0.6%	1.4%	2.9%	3.6%	4.0%
precision@0.05	0.5%	0.7%	2.3%	1.0%	2.2%	1.3%	2.4%	–	10.1%	6.6%	2.2%	2.2%	2.6%	4.3%	2.2%
F1@0.01	3.5%	17.7%	27.7%	16.9%	24.5%	10.1%	37.1%	13.8%	–	3.4%	28.0%	24.8%	17.3%	22.3%	11.0%
F1@0.05	3.6%	11.0%	13.1%	10.3%	12.4%	8.8%	21.2%	14.0%	6.0%	–	13.4%	12.7%	10.5%	13.2%	8.3%
AUC@tukey-q	0.6%	0.7%	0.1%	0.9%	0.5%	1.9%	1.4%	2.3%	7.9%	6.2%	–	0.5%	1.9%	2.5%	2.8%
TPR@tukey-q	0.6%	1.1%	0.6%	1.0%	0.1%	2.3%	1.8%	2.2%	8.3%	6.3%	0.5%	–	2.1%	2.7%	3.0%

Table 4: Means of relative loss in a column measure when optimal model and hyperparameters are selected using the row measure. 0% training contamination. Level of shading highlights three best results in a column.

max/loss	AUC	AUC_w	AUC@0.01	AUC@0.05	TPR@0.01	TPR@0.05	precision@0.01	precision@0.05	F1@0.01	F1@0.05	AUC@welch-mean	TPR@welch-mean	mean	mean@0.01	mean@0.05
AUC	–	0.4%	2.0%	1.2%	2.4%	1.0%	2.3%	2.0%	10.9%	7.1%	2.3%	2.1%	2.8%	4.4%	2.8%
AUC_w	0.1%	–	1.3%	0.5%	1.6%	0.6%	1.6%	1.7%	10.3%	6.7%	1.6%	1.3%	2.3%	3.7%	2.4%
AUC@0.01	0.7%	0.8%	–	1.0%	0.5%	2.1%	1.4%	2.5%	7.9%	6.3%	0.3%	0.6%	2.0%	2.5%	3.0%
AUC@0.05	0.3%	0.2%	0.7%	–	0.9%	0.4%	1.8%	1.5%	8.4%	5.3%	1.0%	0.8%	1.8%	2.9%	1.8%
TPR@0.01	0.6%	1.1%	0.5%	1.0%	–	2.2%	1.8%	2.3%	8.2%	6.2%	0.8%	0.2%	2.1%	2.6%	2.9%
TPR@0.05	0.3%	1.5%	5.7%	1.9%	4.5%	–	4.6%	2.2%	12.3%	6.2%	6.0%	2.4%	4.0%	6.8%	2.6%
precision@0.01	0.7%	1.0%	0.7%	1.5%	1.4%	2.8%	–	1.6%	12.5%	10.1%	0.7%	0.9%	2.8%	3.6%	4.0%
precision@0.05	0.5%	0.7%	2.3%	1.0%	2.2%	1.3%	2.4%	–	10.1%	6.6%	2.3%	1.7%	2.6%	4.3%	2.2%
F1@0.01	3.5%	17.7%	27.7%	16.9%	24.5%	10.1%	37.1%	13.8%	–	3.4%	28.0%	11.8%	16.2%	22.3%	11.0%
F1@0.05	3.6%	11.0%	13.1%	10.3%	12.4%	8.8%	21.2%	14.0%	6.0%	–	13.4%	8.4%	10.2%	13.2%	8.3%
AUC@welch-mean	0.6%	0.7%	0.1%	1.0%	0.5%	2.0%	1.3%	2.3%	7.9%	6.3%	–	0.3%	1.9%	2.5%	2.9%
TPR@welch-mean	2.2%	12.5%	24.4%	18.3%	22.7%	13.2%	23.3%	9.7%	26.2%	14.1%	25.2%	–	16.0%	24.2%	13.8%

Table 5: Means of relative loss in a column measure when optimal model and hyperparameters are selected using the row measure. 0% training contamination. Level of shading highlights three best results in a column.

max/loss	AUC	AUC_w	AUC@0.01	AUC@0.05	TPR@0.01	TPR@0.05	precision@0.01	precision@0.05	F1@0.01	F1@0.05	AUC@welch-median	TPR@welch-median	mean	mean@0.01	mean@0.05
AUC	–	0.4%	2.0%	1.2%	2.4%	1.0%	2.3%	2.0%	10.9%	7.1%	2.3%	2.1%	2.8%	4.4%	2.8%
AUC_w	0.1%	–	1.3%	0.5%	1.6%	0.6%	1.6%	1.7%	10.3%	6.7%	1.6%	1.3%	2.3%	3.7%	2.4%
AUC@0.01	0.7%	0.8%	–	1.0%	0.5%	2.1%	1.4%	2.5%	7.9%	6.3%	0.3%	0.6%	2.0%	2.5%	3.0%
AUC@0.05	0.3%	0.2%	0.7%	–	0.9%	0.4%	1.8%	1.5%	8.4%	5.3%	1.0%	0.8%	1.8%	2.9%	1.8%
TPR@0.01	0.6%	1.1%	0.5%	1.0%	–	2.2%	1.8%	2.3%	8.2%	6.2%	0.8%	0.2%	2.1%	2.6%	2.9%
TPR@0.05	0.3%	1.5%	5.7%	1.9%	4.5%	–	4.6%	2.2%	12.3%	6.2%	6.0%	2.4%	4.0%	6.8%	2.6%
precision@0.01	0.7%	1.0%	0.7%	1.5%	1.4%	2.8%	–	1.6%	12.5%	10.1%	0.7%	0.9%	2.8%	3.6%	4.0%
precision@0.05	0.5%	0.7%	2.3%	1.0%	2.2%	1.3%	2.4%	–	10.1%	6.6%	2.3%	1.7%	2.6%	4.3%	2.2%
F1@0.01	3.5%	17.7%	27.7%	16.9%	24.5%	10.1%	37.1%	13.8%	–	3.4%	28.0%	11.8%	16.2%	22.3%	11.0%
F1@0.05	3.6%	11.0%	13.1%	10.3%	12.4%	8.8%	21.2%	14.0%	6.0%	–	13.4%	8.4%	10.2%	13.2%	8.3%
AUC@welch-median	0.6%	0.7%	0.1%	1.0%	0.5%	2.0%	1.3%	2.3%	7.9%	6.3%	–	0.3%	1.9%	2.5%	2.9%
TPR@welch-median	2.2%	12.5%	24.4%	18.3%	22.7%	13.2%	23.3%	9.7%	26.2%	14.1%	25.2%	–	16.0%	24.2%	13.8%

Table 6: Means of relative loss in a column measure when optimal model and hyperparameters are selected using the row measure. 0% training contamination. Level of shading highlights three best results in a column.

- data-specific

We should not use a cross–model information because that is a very difficult and different task. So instead of doing optimal fpr level selection based on discriminability, we should try to propose a universal measure independent on precomputed model performance.

Or we go forwards with the discriminability:

- use adjusted pvals from Garcia paper
- use some other way to measure discriminability, e.g. the number of discriminable pairs of models

We should also check if we ahve not already answered some of the issues that the reviewers had with the paper, such as explaining why AUC@5 is so good.

References

[1] Bernard Lewis Welch. On the comparison of several mean values: an alternative approach. *Biometrika*, 38(3/4):330–336, 1951.