Optimal human labelling for anomaly detection in industrial inspection

Tim Zander,[12] Ziyan Pan,[1] Pascal Birnstill[2], and Juergen Beyerer[12]

[1] Institut für Anthropomatik und Robotik, Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie (KIT)
[2] Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (IOSB)

Abstract Anomaly detection with machine learning in industrial inspection systems for manufactured products relies on labelled data. This rises the question how the labelling by humans should be conducted. We consider the case where we want to optimise the cost of the combined inspection process done by humans and an algorithm. This also influences the combined performance of the trained model as well as the knowledge of the performance of this model. We focus on so called one-class classification problem models which produce a continuous outlier score. We establish some cost model for human and machine combined inspection of samples. We then discuss in this cost model how to select two optimal boundaries of the outlier score where in between these two boundaries human inspection takes place. We also frame this established knowledge into an applicable algorithm.

Keywords Mathematical methods and models, artificial intelligence and machine learning, quality control

1 Introduction

The detection of non-common patterns in a batch of samples is a strong point of human visual cognition. Still there are many known limitations to human visual inspection as well as cost issues in real world production systems. The training of machine learning models for anomaly detection of industrial inspection problems is often done as a one-class classification problem where only good samples are presented to the algorithm. The background for this is that it is in general easy to acquire good samples but difficult and expensive to find anomalous samples. A dataset for benchmarking this type of algorithm is the MVTec-dataset [1] [2]. The best performing model[3] on this dataset is *"Patchcore"* [3]. For a given picture sample a *"Patchcore"*-model after training produces an outlier score together with a heat map on the likelihood of being an anomalous area. This is done by performing outlier-detection on the deep-features of a pretrained neural network of the images. The cutoff values for an anomaly in the outlier score of *"Patchcore"* are optimised in the paper by finding the cutoff-value with the highest F1-score. This already assumes that there are known outliers which are potentially very costly to acquire. Although we think of models designed for the MVTec dataset like *"Patchcore"* as the main application, our method of finding two boundaries for the outlier score, where in-between human inspection will take place, will work for any model of an one-class classification problem [4] with a continuous score.

More precisely, in this paper we formulate the problem of optimal usage of human inspection after acquiring of initial data for training. For this we assume that there are certain costs for inspection and costs for falsely classified samples. We are not are aware that such a human-in-the-loop machine learning consideration exists in the literature, although more generic considerations about iterative machine teaching and active learning can be found in [5]. A similar process by giving the human some sort of optimal presentation of data for labelling was done in [6]. However, this method is not applicable for the one-class outlier classification problems on images we consider here. In [7] it is shown, that for one-class classification models one can train an additional model on the bad samples and use a combined score on the good and bad sample models to find the most promising new samples for labelling. The authors show that using one of their active learning methods one can achieve faster convergence and better overall performance of the model. We refer to Munro's book [8] for a general overview of human-in-the-loop machine learning.

Another important concept which we will discuss and use is that of probabilistic classifiers. Probabilistic classifiers are classifiers which output a probability distribution on the target classes instead of just a

[3] `https://paperswithcode.com/sota/anomaly-detection-on-mvtec-ad`

score. Model calibration is a technique which achieves that a classifier will have a probabilistic output [9] [10]. A calibrated one-class classifier will give out a probability p which will represent the probability of being in the one class. In safety-critical applications it is important to have an idea of uncertainty of the model. Hence a probabilistic output is of great help with regard to such problems. Even in situations which are just cost-critical we will show that we can exploit having an uncertainty estimate of the classifier for a given sample to make better decisions.

2 Model

In this section we will describe the necessary pre-conditions and cost assumptions. Further we describe how, after initial training of our one-class classifier, we can establish our first optimal boundaries. We do describe multiple alternatives here. Then we pass on to acquiring more knowledge about the outliers we will encounter and their outlier scores. This will then be used to establish optimal decisions for the cutoff parameters of human inspection in the sense of our pre-made cost assumptions.

2.1 Pre-conditions

First we introduce a few more preliminary and formal assumptions and notations. We assume that there exists a set of images or more general data I which each have a hidden label $\{0, 1\}$ where images with label 0 are good samples and images with label 1 are anomalous samples. We will observe these samples in some process such as an industrial inspection task one after another. For our cost considerations we assume that the process of labelling a sample by a human has a cost c_l associated with it. Further we assume that human labelling perfectly assigns the correct label to the data. With N initially labelled data points we train and test a model M which will then produce an outlier-score $M(i) \in \mathbb{R}$ for every (new) image i we observe. We set a lower and upper decision boundary for manual inspection b_l and b_u such that any image i with outlier score $M(i)$, where $b_l < M(i) < b_u$ holds, will be inspected by a human.

2.2 A priori cost and anomalous data

For our cost considerations we further assume that there is a known (possibly non-linear[4]) cost-function C_f such that the absolute cost of missed outliers can be calculated as $C_f(\mathbf{FOR}) \cdot K$ where **FOR** is the false omission rate, i.e. the percentage of anomalies in the accepted samples, and K the absolute number of accepted samples. The cost of false positive samples are associated with a cost per sample of c_r. This could be for example lost revenue and disposal costs of an unnecessarily disposed sample of good state.

2.3 Initial cut-off boundaries

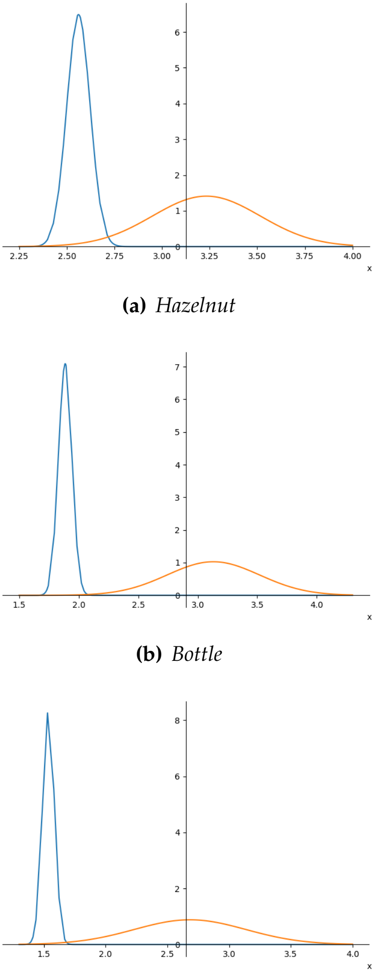
We assume now that the initial sampling and labelling of data D and the training of a model M is conducted. We update our initial belief p_o of the outlier percentage by taking the percentage of outliers in the sampled D into account. We are now interested in finding optimal cutoff parameters b_l, b_u in this stage. We discuss multiple alternatives now.

A priori anomaly distribution

In the first case we assume that the distribution of the outlier score of samples with label 0 and also of the samples with label 1 is both Gaussian[5]. For the good samples we can directly estimate this distribution. We get some distribution g_g with mean μ_g and variance σ_g. For the bad samples we also get some Gaussian distribution g_b. In the case where there are no bad samples available, we take some initial belief about the distribution, which we could take from former observations such as the MVTec dataset or a similar product line (see Figure 1), as our distribution. We can find the optimal parameters b_l, b_u in terms of cost. In order to find these parameters one would minimise Equation 1 of Section 2.4.

[4] One reason for non-linearity could be reputation costs, i.e., due to network effects reputation falls non-linearly with increasing fault-rate.

[5] A non-Gaussian distribution could also easily be considered here.



(a) *Hazelnut*

(b) *Bottle*

(c) *Leather*

Figure 1: These are the Gaussian distributions of anomaly scores for different items from the MVTec Dataset. Blue represents the good sample distribution and red represents the bad sample distribution. The model where the anomaly score stems from was Patchcore [3] and it was trained with training sample split of the MVTec dataset. Then the anomaly score output of the trained model on the good and bad samples of the test dataset split was used to find the shown Gaussian distributions. On these data-sets the established model has an AUC-score of 0.9996 for *Hazelnut*, 1.0 for *Bottle* and 1.0 for *Leather* on the test dataset samples.

Optimal cut-off sigma

Another approach would be to omit to define an a priori distribution of g_b and instead take a cutoff parameter x such that any sample with outlier score higher than $\mu_g + x \cdot \sigma_g$ is considered anomalous. The choice of the parameter x can be done as follows. We assume that we cannot inspect every piece which we observe but only some percentage p_i of it. Hence we have to find x in such a way that the expected amount of samples classified as anomalous is at most the amount that can be handled. Hence we have to pick x such that

$$p_i \geq (1 - p_o) \int_{\mu_g + x \cdot \sigma_g}^{\infty} g_g + p_o$$

holds. Note that we omitted the expected false negative classified samples in our considerations, but we assume that this amount is negligibly small. In case there is no sample to classify at the moment we might pick a random sample. In case we acquired enough bad samples we can infer the distribution g_b or update our initial belief about it. More details on the belief update of a Gaussian distribution can be found in [11].

Calibrated output

In some cases the model comes with a calibrated probabilistic output. This roughly means that the output value of the model $M(*)$ is a probability of being an outlier, e.g. we expect to find $q * 100$-many outliers of 100-samples i' with score $M(i') = q$. With such a calibrated model we can directly use the model output as our probability. We will not further assume that our model is calibrated although the following should be straightforward to adapt for directly using this output instead of learning some probability as in the previous paragraph.

Now we have found a priori parameters b_l, b_u or just $b_l (= \mu_g + x \cdot \sigma_g)$. With these we can set up our initial human in the loop process. After some time we will enrich our dataset of labelled pieces and therefore can update our believe about the Gaussian curves g_g, g_b as described in [11] or interfere the distributions g_g, g_b directly from all the gathered data. There is some caveat with the selection of the samples: Because of our parameters the selection of the samples is biased. This either needs to be corrected through enough random samples or giving the unlabelled data some pseudo label with continuous value greater 0 and smaller than 1. Additionally we could use the gathered data to further improve the model M or respectively re-train a new M with the new data and old data depending on the algorithm in use. In any case we now fix some model M, some p_0 and the Gaussian distributions g_g, g_b associated with it as well as the gathered data. In case we observed and classified a new sample we could continue to do a belief update of our estimated values p_o, g_g and g_b and retrain our model M in order to keep improving it. But we omit such considerations in the rest of the paper.

2.4 Cost-calculation

We calculate the cost associated for some fixed b_l and b_u for the next samples. We expect to see p_o-percent outliers which we have updated from the observations D. Additionally we can calculate the expected percentage that the next sample will be true positive: $\mathbf{TP}(b_l) = p_o \int_{b_l}^{\infty} g_b$, true negative: $\mathbf{TN}(b_u) = (1 - p_o) \int_{-\infty}^{b_u} g_g$, false negative: $\mathbf{FN}(b_l) = p_o \int_{-\infty}^{b_l} g_b$ and false positive: $\mathbf{FP}(b_u) = (1 - p_o) \int_{b_u}^{\infty} g_g$. From this we can calculate the false omission rate $\mathbf{FOR} = \frac{\mathbf{FN}}{\mathbf{FN}+\mathbf{TN}}$. Now for the next sample have the cost function $\mathcal{C}(b_l, b_u)$ defined as follows:

$$\begin{aligned} &C_f(\mathbf{FOR}(b_l)) \cdot [\mathbf{TN}(b_u) + \mathbf{FN}(b_l)] + c_r \cdot \mathbf{FP}(b_u) + \\ &c_l \cdot (1 - p_o) \int_{b_l}^{b_u} g_g + c_l \cdot p_o \int_{b_l}^{b_u} g_b. \end{aligned} \tag{1}$$

This function is our minimisation target for which we choose b_l and b_u accordingly:

$$\begin{aligned} \min_{b_l, b_u} \quad & \mathcal{C}(b_l, b_u, g_b, g_g, p_0) \\ \text{s.t.} \quad & b_l \leq b_u \\ & b_l, b_u \in \overline{\mathbb{R}} \end{aligned} \tag{2}$$

where $\overline{\mathbb{R}}$ is the set of the extended real numbers which additionally contains plus and minus infinity, i.e. $\mathbb{R} \cup \{-\infty, +\infty\}$.

In case of a very low outlier rate p_o we can simplify the cost by setting $b_u = \infty$ and the optimisation problem becomes a single variable problem. Often it will be the case that we have a fixed percentage of images, say p_f, which we can inspect due to such things as fixed amount of available human labour. In this case the lower part of the cost function 1 will be replaced by the constraint

$$p_f = (1 - p_o) \int_{b_l}^{b_u} g_g + p_o \int_{b_l}^{b_u} g_b.$$

If we additionally set $b_u = \infty$ we can already find the optimal b_l by just using this constraint. But these considerations are still useful as we can now estimate the cost of our system and further estimate whether it is useful to employ or dismiss a human at a certain cost or estimate the cost saving for a higher or lower rate of inspection of samples.

2.5 Non independence of outlier observations

In the case where we believe there is a non-independence of the series of observed data[6] we could increase the believed percentage of outliers p_o for the next few observed samples after observing an outlier. This ensures that the costs stay optimal for the next observed samples with higher anomaly probability. Note that in more complicated production environments we may observe pieces from multiple different machines. If possible one should keep track of the machines a piece went through to get more individual assessment of the anomalous probabilities.

3 Algorithm

In this section we combine the observations established in the last section into an combined algorithm (see Algorithm 1). As an input to our algorithm there is a one-class classification model M that needs N-many samples for initial training and testing, and there is also a belief about the percentage of outliers p_o in the samples to be observed. Additionally we have the cost function C_f and a real value c_r representing the cost of a false positive sample. Moreover we fix an amount of outliers we want to observe L. The algorithm starts by letting a human

[6] A broken machine could for example produce a sudden stream of defect parts.

label samples till we receive a set D containing N-many labelled samples with label 0. We use this dataset D to train some model M_D which then is used to produce some outlier scores for the test data split of D. This is then used to find more anamalous samples in order to form a probabilistic model by inferring a Guassian curve of the good and a Gaussian curve of the bad samples. With this we are finally able to find the cost optimal parameters b_l and b_u which mark the outlier score interval where human inspection takes place.

Algorithm 1 Find optimal interval for human inspection

```
initialization: $p_o, C_f, c_r, c_l, N, L$
$n \leftarrow 0$
for $n < N$ do
    wait for next sample $s$
    get label $l(s)$ (by human)
    $n \leftarrow n + 1 - l(s)$
    $p_o \leftarrow$ belief update through observed $l(s)$
end for
return  training dataset $D$ , $p_0$
$M_D \leftarrow$train model with $D$
$b'_l \leftarrow$ (see Section 2.3 for possible computations)
$k \leftarrow 0$
for $k < L$ do
    get next sample $s$
    if $b_l < M_D(s)$ then
        get label $l(s)$ (by human)
    end if
    $k \leftarrow k + l(s)$
    $p_o \leftarrow$ belief update through observed $l(s)$
end for
return  updated dataset $D$, $p_0$
$g_g, g_b \leftarrow$ interfere Gaussian from data $D$
solve $\min_{b_l,b_u} \mathcal{C}(b_l, b_u, g_b, g_g, p_0)$
return  Model $M_D$ and inspection interval values $b_l, b_u$
```

4 Discussion and future work

We establish theory for the cost-optimal selection of samples of one-class classifications models. For this we established a cost-model and showed how to infer probabilistic knowledge of the samples online and offline in order to establish a cost-optimal decision for a human inspection boundary in the outlier score. Moreover, we have merged this into an algorithm which can be applied in production. For now we have not considered the case of retraining the model and we can assume that this will be done occasionally till the economic evaluation stabilises or the performance is satisfactory. Also the problem of a timely dependence of the occurrence of outliers which could stem from faulty machines was discussed. At worst there could be no outlier samples or only a very biased selection of them. A detailed analysis of the practical relevance of this problem could be an interesting topic for future investigation. There could also be potential for future work especially in the case where the one-class problem is a moving target, i.e. the golden sample changes over time. The case for selecting valuable examples for improving the model performance also seems an interesting area not yet considered and will probably require an extra model which is also trained with the outliers. Another not yet used feature is utilising the presentation of anomalous areas on the image for better outlier visualisation for the user decision. There, another optimisation problem arises which is the optimisation of the cutoff parameter for the selection of the anomalous area. A more general question is the question of a good visualisation to improve human performance.

References

1. P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, “Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9592–9600.
2. P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, “The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection,” *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1038–1059, 2021.

3. K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 318–14 328.
4. P. Perera, P. Oza, and V. M. Patel, "One-class classification: A survey," *arXiv preprint arXiv:2101.03064*, 2021.
5. E. Mosqueira-Rey, D. Alonso-Ríos, and A. Baamonde-Lozano, "Integrating iterative machine teaching and active learning into the machine learning loop," *Procedia Computer Science*, vol. 192, pp. 553–562, 2021.
6. C. Chai, L. Cao, G. Li, J. Li, Y. Luo, and S. Madden, "Human-in-the-loop outlier detection," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020, pp. 19–33.
7. P. Schlachter and B. Yang, "Active learning for one-class classification using two one-class classifiers," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1197–1201.
8. R. Munro, *Human-in-the-loop machine learning*. New York, NY: Manning Publications, Oct. 2021.
9. J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, and T. Schön, "Evaluating model calibration in classification," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 3459–3467. [Online]. Available: https://proceedings.mlr.press/v89/vaicenavicius19a.html
10. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.
11. K. Murphy, "Conjugate bayesian analysis of the gaussian distribution," 11 2007. [Online]. Available: https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf