Data Mining and Knowledge Discovery (2024) 38:3719–3757
https://doi.org/10.1007/s10618-024-01056-5

Evaluating outlier probabilities: assessing sharpness, refinement, and calibration using stratified and weighted measures

Philipp Röchner[1] · Henrique O. Marques[2] · Ricardo J. G. B. Campello[2] · Arthur Zimek[2]

Received: 4 December 2023 / Accepted: 29 June 2024 / Published online: 19 July 2024

Abstract
An outlier probability is the probability that an observation is an outlier. Typically, outlier detection algorithms calculate real-valued outlier scores to identify outliers. Converting outlier scores into outlier probabilities increases the interpretability of outlier scores for domain experts and makes outlier scores from different outlier detection algorithms comparable. Although several transformations to convert outlier scores to outlier probabilities have been proposed in the literature, there is no common understanding of good outlier probabilities and no standard approach to evaluate outlier probabilities. We require that good outlier probabilities be sharp, refined, and calibrated. To evaluate these properties, we adapt and propose novel measures that use ground-truth labels indicating which observation is an outlier or an inlier. The refinement and calibration measures partition the outlier probabilities into bins or use kernel smoothing. Compared to the evaluation of probability in supervised learning, several aspects are relevant when evaluating outlier probabilities, mainly due to the imbalanced and often unsupervised nature of outlier detection. First, stratified and weighted measures are necessary to evaluate the probabilities of outliers well. Second, the joint use of the sharpness, refinement, and calibration errors makes it possible to independently measure the corresponding characteristics of outlier probabilities. Third, equiareal bins, where the product of observations per bin times bin length is constant, balance the number of observations per bin and bin length, allowing accurate evaluation of different outlier probability ranges. Finally, we show that good outlier probabilities, according to the proposed measures, improve the performance of the follow-up task of converting outlier probabilities into labels for outliers and inliers.

Keywords Outlier detection · Anomaly detection · Unsupervised learning · Outlier probabilities · Calibration · Refinement · Sharpness · Outlier ensembles

Responsible editor: Rita P. Ribeiro.

Extended author information available on the last page of the article

1 Introduction

An *outlier probability* is the probability that an observation is an outlier, where "an outlier [is] an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"(Hawkins 1980). For example, outliers can be implausible electronic health records in cancer registries (Röchner and Rothlauf 2023). Typically, outlier detection algorithms calculate real-valued outlier scores to identify outliers. Converting outlier scores to outlier probabilities increases the interpretability of outlier scores for domain experts and makes outlier scores from different outlier detection algorithms comparable. For example, outlier probabilities can normalize outlier scores from multiple outlier detection algorithms to build outlier ensembles (Kriegel et al. 2011; Rayana and Akoglu 2016; Campos et al. 2018) or to select outlier detection models (Marques et al. 2020, 2022).

Although several transformations have been proposed to convert outlier scores into outlier probabilities (Gao and Tan 2006; Kriegel et al. 2011; Bouguessa 2012; Bauder and Khoshgoftaar 2017; Perini et al. 2021; Muhr et al. 2023), a common approach to evaluate these transformations is lacking. Only Kriegel et al. (2011) propose a measure for outlier probabilities. To our knowledge, there is no common understanding of good outlier probabilities and no standard approach to evaluating outlier probabilities. We believe that research on outlier probabilities will benefit from a standard approach to evaluating different characteristics of outlier probabilities, as researchers can more easily demonstrate the strengths and weaknesses of outlier detection algorithms and outlier score transformations.

The evaluation of probability estimates in supervised learning has been studied extensively (Platt et al. 1999; Niculescu-Mizil and Caruana 2005; Kull et al. 2017). In this context, *proper scoring rules* (Murphy and Winkler 1970) yield the best (lowest) score when the predicted probabilities match the true probabilities. Notable examples of proper scoring rules include the Brier score (Brier 1950) and cross-entropy (also called log-loss or log-likelihood) (Shuford Jr et al. 1966). Both measures are special cases of an integral over a Beta density of costs (Buja et al. 2005).

Compared to supervised learning, the evaluation of outlier probabilities poses several challenges. First, outlier detection is a highly imbalanced problem with typically only a very small percentage of outliers, which can make it harder to evaluate the probabilities of the rare outliers. Second, probabilities generated by unsupervised outlier detection algorithms or unsupervised outlier score transformations are typically inferior to probabilities generated by supervised classification and supervised calibration algorithms because unsupervised outlier detection does not use datasets with ground-truth labels to identify outliers and adjust probabilities. Therefore, outlier probability measures must be able to evaluate probabilities well over a wide quality range. For example, log-loss is less suitable for evaluating outlier probabilities because it can be undefined or unbounded for low-quality probabilities.

This article investigates what good outlier probabilities are and how to evaluate them. In the following discussion, both questions are independent of how the

outlier detection algorithm or the outlier score transformation computed the outlier scores or outlier probabilities. In other words, the outlier probabilities to be evaluated can be computed in an unsupervised, semi-supervised, or supervised way.

Following DeGroot and Fienberg (1982) for evaluating probability estimates in supervised classification, we require good outlier probabilities to be *refined* and *calibrated* (Dawid 1982), also called reliable (Murphy 1973). Refined means that the probabilities of outliers and inliers are well distinguished (DeGroot and Fienberg 1982). Calibrated means that the outlier probabilities match the empirical frequency of observations with a similar outlier probability to be an outlier (DeGroot and Fienberg 1982). In addition, we require that good outlier probabilities are *sharp* according to Gneiting et al. (2007), that is, concentrated around zero or one. As pointed out by Gneiting et al. (2007), one typically tries to increase the refinement and sharpness of probability estimates while keeping the probabilities calibrated.

To assess the sharpness, refinement, and calibration of outlier probabilities, we adapt measures for evaluating probability estimates in supervised learning (the Brier score and the calibration error) and propose novel sharpness and refinement measures using ground-truth labels that indicate which observation is an outlier or an inlier. The calibration and refinement measures partition the outlier probabilities into bins or use kernel smoothing. We find that several aspects are relevant when evaluating outlier probabilities: First, to properly evaluate the probabilities of the outliers, it is necessary to use stratified and weighted versions of the measures. Second, the joint use of the sharpness, refinement, and calibration errors makes it possible to independently measure the corresponding characteristics of outlier probabilities. Third, equiareal bins, where the product of observations per bin and bin length are constant, balance the number of observations and bin length, allowing us to accurately evaluate different ranges of outlier probabilities. Finally, we demonstrate that good outlier probabilities, according to the proposed measures, perform well in the follow-up tasks of converting outlier probabilities into binary labels (outliers and inliers) based on a predefined threshold.

We organize the remainder of the paper as follows. In Sect. 2, we review articles that propose outlier probabilities and how the outlier probabilities have been evaluated. We specify what we consider to be outliers, outlier scores, and outlier probabilities in Sect. 3. In Sect. 4, we define the characteristics of good outlier probabilities — sharpness, refinement, and calibration — and adapt the Brier score and calibration error, typically used for evaluating probability estimates in supervised classification, for evaluating outlier probabilities. In Sect. 5, we propose measures to evaluate sharpness and refinement. To address the imbalanced nature of outlier detection, we adjust the sharpness, refinement, and calibration errors using stratified and weighted variants in Sect. 6. Since some calibration and refinement errors use bins, we discuss appropriate bin types for outlier probabilities in Sect. 7. Our experiments, which show that the adapted and proposed measures are appropriate for outlier probabilities, are described in Sect. 8. Finally, we discuss the results of our experiments in Sect. 9, elaborate on practical considerations when using outlier probabilities in follow-up tasks in Sect. 10, and conclude our work in Sect. 11.

2 Related work: evaluating outlier probabilities

As mentioned above, there is no standard approach to evaluating outlier probabilities (Gao and Tan 2006; Kriegel et al. 2011; Bouguessa 2012; Bauder and Khoshgoftaar 2017; Perini et al. 2021; Muhr et al. 2023).

Some studies evaluate outlier probabilities similarly to outlier scores: Kriegel et al. (2009, 2012) and Clifton et al. (2014) use outlier probabilities to rank observations and evaluate the ranked observations against a reference ground-truth ranking using the area under the receiver operating characteristic curve (AUC ROC), which is typical for outlier scores.

Other studies visualize observations and their outlier probabilities. Kriegel et al. (2009, 2012), Achtert et al. (2010), Sotiris et al. (2010), and Clifton et al. (2014) show one-dimensional and two-dimensional datasets jointly with their outlier probabilities and discuss the quality of the probabilities.

When evaluating outlier probabilities as outlier scores, one does not explicitly assess the properties of probabilities. For example, when only evaluating the ranking of observations according to outlier probabilities, one does not assess if the outlier probabilities match the empirical frequency of outliers. We discuss the characteristics relevant for outlier probabilities compared to outlier scores in Sect. 4.1.

Many studies proposing outlier detection algorithms computing outlier probabilities (Sotiris et al. 2010; Clifton et al. 2014) or outlier score transformations (Bouguessa 2012; Bauder and Khoshgoftaar 2017; Perini et al. 2021) do not directly evaluate the corresponding outlier probabilities. Instead, outlier probabilities are used for follow-up tasks. The performance of the follow-up task is evaluated using the ground-truth labels that indicate which observation is an outlier or an inlier.

Gao and Tan (2006), Sotiris et al. (2010), and Clifton et al. (2014) use outlier probabilities to select an outlier score threshold that separates outliers from inliers. After converting outlier scores to binary outlier labels, the binary outlier labels are evaluated against the ground-truth labels using classification measures such as accuracy, precision, recall, F_1 score, false alarm rate, and AUC ROC (Gao and Tan 2006; Sotiris et al. 2010; Clifton et al. 2014).

Outlier ensembles are another follow-up task using outlier probabilities (Gao and Tan 2006; Kriegel et al. 2011; Bauder and Khoshgoftaar 2017). Outlier ensembles combine the outlier probabilities of different outlier detection algorithms, for example, by averaging. Measures, such as the AUC ROC, evaluate the ensemble's performance (Gao and Tan 2006; Kriegel et al. 2011).

Perini et al. (2021) use outlier probabilities to determine the confidence that an outlier detection algorithm will detect an observation as an outlier. The confidence that an observation will be an outlier is then evaluated by comparing the confidence to the empirical frequency of that observation being an outlier for multiple perturbed datasets. Finally, the average difference between the confidence and the empirical frequency of outliers and inliers is weighted equally.

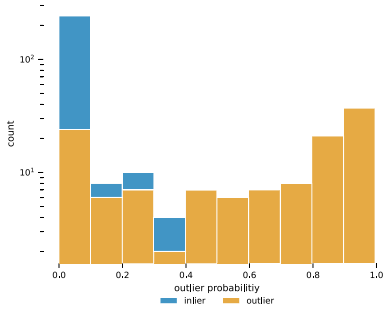
Indirect evaluation of outlier probabilities makes it challenging to distinguish the outlier probabilities' quality from the follow-up task's performance due to

potential interaction effects. In particular, it is difficult to generalize the quality of outlier probabilities from one follow-up task to another. In other words, just because outlier probabilities work well for one follow-up task does not necessarily mean they will work well for another. For example, Gao and Tan (2006) found that outlier score transformations using sigmoid functions outperformed calibration using mixture modeling for converting outlier probabilities into labels for outliers and inliers. However, the reverse is true for outlier ensembles, where transformations using mixture modeling outperform transformations using sigmoid functions (Gao and Tan 2006).

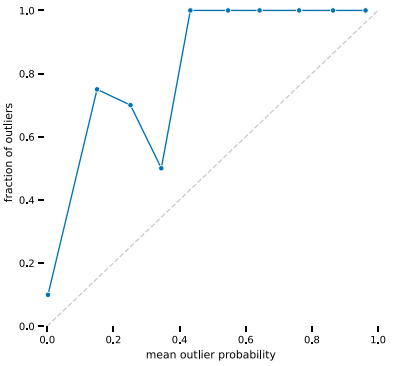
Few studies directly evaluate outlier probabilities as probabilities. Gao and Tan (2006) visually evaluate outlier probabilities using *calibration plots*, also called reliability diagrams (DeGroot and Fienberg 1983); see, for example, Fig. 1b. Calibration plots bin outlier probabilities, comparing the average outlier probability per bin on the x-axis with the proportion of outliers per bin on the y-axis. Calibrated outlier probabilities lie close to the diagonal from the lower left to the upper right corner.

Kriegel et al. (2011) assess outlier probabilities using an error measure that calculates the average probability that an outlier is an inlier and the probability that an inlier is an outlier. The error measure averages these probabilities equally weighted to account for the imbalanced nature of outlier detection.

A systematic evaluation approach, however, that measures different characteristics of outlier probabilities is still missing in the literature.



(a) Stacked histogram of outlier probabilities. Sharp outlier probabilities are concentrated around zero or one. For refined outlier probabilities, each bin contains only outliers or only inliers. The y-axis with the number of outlier probabilities has a logarithmic scale.



(b) Calibration plot of outlier probabilities. Calibrated outlier probabilities are on the diagonal.

Fig. 1 Example visualizing the sharpness, refinement, and calibration of outlier probabilities. We computed outlier probabilities using Gaussian scaling (Kriegel et al. 2011) and outlier scores using the *k*-Nearest Neighbors Detector (Ramaswamy et al. 2000) on the Ionosphere dataset (Campos et al. 2016). Each plot has ten equidistant bins (see Sect. 4.3)

3 Notation: outliers, outlier scores, and outlier probabilities

We characterize outliers and define outlier scores and outlier probabilities. There are several definitions of outliers (Hawkins 1980; Barnett et al. 1994; Ruff et al. 2021). Typically, outliers are considered to be significantly different from the remaining observations (Barnett et al. 1994; Hawkins 1980). Formally, this can be described in terms of probability, which means that outliers come from low-probability regions of the data (Ruff et al. 2021). However, the origin of outliers can vary: The same or different processes may generate outliers and inliers (Hawkins 1980). When different processes generate outliers and inliers, outliers are often referred to as anomalies. If different processes generate outliers and inliers, and the origins of the observations are known, then outliers and inliers can be labeled as outliers and inliers directly. If the same process generates outliers and inliers, or the origin of the observations is unknown, then outliers can be defined based on low-probability regions according to the data's probability distribution, assuming the data's probability distribution is known. Alternatively, domain experts can label outliers (Röchner and Rothlauf 2023), which can be time-consuming and, to some extent, subjective. Overall, the definition of outliers and the origin of ground-truth labels for outliers and inliers are problem and dataset specific. To evaluate outlier probabilities in the next sections, we assume that observations have ground-truth labels indicating whether they are outliers or inliers, as is common practice in outlier detection research (Campos et al. 2016).

In the following, we study a dataset $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ with ground-truth labels $\boldsymbol{y} = \{y_i\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^n$ and $y_i \in \{0, 1\}$ for all $1 \leq i \leq N$. Inliers are labeled 0, and outliers are labeled 1. We assume that the dataset $\boldsymbol{X}$ and the ground-truth labels $\boldsymbol{y}$ are a sample jointly drawn from the Cartesian product $\mathbb{R}^n \times \{0, 1\}$ under the conditional probability distribution $P : \mathbb{R}^n \times \{0, 1\} \to [0, 1]$.

For an outlier detection algorithm $D : \mathbb{R}^n \to \mathbb{R}$, we obtain outlier scores $\boldsymbol{s} = \{s_i\}_{i=1}^N$ where $s_i := D(\boldsymbol{x}_i) \in \mathbb{R}$ for all $1 \leq i \leq N$. Either the outlier scores $\boldsymbol{s}$ are directly interpretable as probabilities or an outlier score transformation $T : \mathbb{R} \to [0, 1]$ converts the outlier scores $\boldsymbol{s}$ to outlier probabilities $\boldsymbol{p} = \{p_i\}_{i=1}^N$, where $p_i := T(s_i) = T(D(\boldsymbol{x}_i)) \in [0, 1]$ for all $1 \leq i \leq N$. For example, the outlier scores $\boldsymbol{s}$ can be linearly scaled to the interval [0, 1].

The outlier probability evaluation discussed in this article is independent of how the outlier scores and outlier probabilities are computed; that is, the outlier detection algorithms D and the outlier score transformation T can compute the outlier scores and outlier probabilities to be evaluated in an unsupervised, semi-supervised, or supervised way.

4 Background: evaluating probability estimates

The evaluation of probability estimates for binary classification is related to the evaluation of outlier probabilities if we consider outliers and inliers as classes. Therefore, this section adapts results for the evaluation of probability estimates to the evaluation of outlier probabilities and discusses necessary adjustments.

4.1 Sharpness, refinement, and calibration for outlier probabilities

We require that *ideal* outlier probabilities are sharp, refined, and calibrated and define these characteristics in the following. Our definition of sharpness below follows Gneiting et al. (2007).

Definition 1 (Sharpness) Outlier probabilities are sharp if they are concentrated at zero or one, that is, if $p_i = 0$ or $p_i = 1$ for all $1 \leq i \leq N$.

As noted by Gneiting et al. (2007), sharpness only depends on the outlier probabilities $\boldsymbol{p}$ and is independent of the ground-truth labels $\boldsymbol{y}$.

Histograms of outlier probabilities visualize sharpness; see, for example, Fig. 1a. The more outlier probabilities are sharp, the more they have a U-shaped histogram: many observations have an outlier probability near one or zero, and few observations have an outlier probability in between.

Perfectly separated probabilities of outliers and inliers, that is, the probabilities of outliers are always larger than the probabilities of inliers, are not necessarily sharp: a transformation could map the scores of outliers to 0.51 and the scores of inliers to 0.49. Thus, sharpness is a relevant property even when considering separated outlier probabilities.

Outlier probabilities that are randomly zero or one, without distinguishing between outliers and inliers, are sharp. Thus, sharpness does not indicate whether observations with the same outlier probability are all outliers or all inliers. To describe the purity of observations with similar outlier probabilities, we define *refinement*.

Definition 2 (Refinement) Outlier probabilities are refined if observations with similar outlier probabilities for some equivalence relation $\sim$ are all outliers or all inliers; that is, if $p_i \sim p_j$, then $y_i = y_j$ for all $1 \leq i, j \leq N$.

As an equivalence relation for refinement (Definition 2), one can take, for example, the equality of outlier probabilities. Refinement then requires that if two outlier probabilities are equal, both must be outliers or inliers.

Another choice of equivalence relation for refinement (Definition 2) is that two outlier probabilities are equivalent if they are in the same bin for some binning of the interval [0, 1]. Section 4.3 and Sect. 7 discusses the binning of outlier probabilities in more detail.

Refinement according to DeGroot and Fienberg (1982) compares two calibrated sets of probabilities using stochastic transformations or the availability of probability functions for the probabilities being compared. Refinement in Definition 2 is independent of calibration and is the characteristic of a single set of outlier probabilities $\boldsymbol{p}$ with ground-truth labels $\boldsymbol{y}$; this allows to discuss refinement separately from calibration and to measure refinement of a single set of outlier probabilities in Sect. 5.2.

Visually, refined outlier probabilities have histograms where each bin contains only outliers or only inliers; see, for example, Fig. 1a. Here, the bins with outlier probabilities greater than 0.4 contain only outliers, and these outlier probabilities are refined. The bins with outlier probabilities less than 0.4 contain outliers and inliers; the outlier probabilities of these bins are not refined.

Outlier probabilities that alternate between only outliers or only inliers in each bin, such as the interval [0, 0.1) contains only outliers, the interval [0.1, 0.2) contains only inliers, and so on, are refined in each bin. These refined outlier probabilities, however, would not be helpful.

Often, outlier score transformations can only partially sharpen or refine outlier probabilities. For example, there are inliers and outliers with similar outlier probabilities, as shown in Fig. 1a for outlier probabilities less than 0.4. For outlier probabilities that cannot be sharp or refined, calibration requires that the outlier probabilities have the 'correct' value between zero and one. For example, out of 100 observations with a similar outlier score, where 70% of the observations are outliers and 30% are inliers, the 'correct' outlier probability would be approximately 0.7. Following DeGroot and Fienberg (1982) and Dawid (1982), we define calibration for outlier probabilities.

Definition 3 (Calibration) Outlier probabilities are calibrated if all outlier probabilities p_i equal the probability P that given the outlier score s_i, an observation is an outlier; that is, outlier probabilities are calibrated if $p_i = P(y = 1|s_i)$ for all $1 \leq i \leq N$, where $P(y = 1|s_i)$ is the conditional probability that, given an outlier score s_i, an observation is an outlier.

Calibration plots, such as Fig. 1b, visualize the calibration of outlier probabilities. In histograms of outlier probabilities, calibration requires that bins have an average outlier probability equal to their proportion of outliers. In other words, bins with a high proportion of outliers are closer to one, and bins with a high proportion of inliers are closer to zero.

Changing the calibration of outlier probabilities can affect their sharpness depending on their refinement. For example, to be calibrated, outlier probabilities with similar values that belong primarily to outliers must correspond to the high proportion of outliers. Therefore, increasing the calibration will push the outlier probabilities toward one. Likewise, increasing the calibration will push outlier probabilities with similar values that belong mostly to inliers toward zero. In contrast, to be calibrated, similar outlier probabilities, half of which belong to outliers and half to inliers, must have outlier probabilities equal to the outlier

proportion of 0.5: increasing the calibration of such outlier probabilities will push them toward 0.5. As a result, increasing the calibration of outlier probabilities increases their sharpness when their refinement is high; it decreases their sharpness when their refinement is low.

Sharpness, refinement, or calibration alone are insufficient for good outlier probabilities. For example, outlier probabilities equal to the proportion of outliers in the dataset are calibrated, but are not refined or sharp and are rarely helpful. Another example is outlier probabilities, which are zero for outliers and one for inliers. Because these outlier probabilities are only zero or one, they are sharp. In addition, these outlier probabilities are refined because all observations with the same outlier probability have the same ground-truth label. The outlier probabilities are, however, uncalibrated.

If outlier probabilities are calibrated and sharp, they are also refined (DeGroot and Fienberg 1983). If outlier probabilities are sharp, they are either one or zero. If the outlier probabilities are additionally calibrated, then the observations with an outlier probability of zero have a zero probability of being an outlier. Thus, all observations with an outlier probability of zero are inliers. Similarly, observations with an outlier probability of one have a probability of one of being an outlier. Consequently, all observations with an outlier probability of one are outliers.

As mentioned above, DeGroot and Fienberg (1982) discuss refinement only for calibrated outlier probabilities. Since DeGroot and Fienberg (1983) assume that probabilities are calibrated when discussing refinement, sharp probabilities are also refined. We introduce refinement independently of calibration in Definition 2. Consequently, according to our definition of refinement, refinement and sharpness are different characteristics of outlier probabilities. In particular, this makes it necessary to define sharpness in addition to refinement.

In summary, we require that *ideal* outlier probabilities are sharp, refined, and calibrated.

4.2 Brier score

Typically, outlier probabilities are not perfectly sharp, refined, or calibrated. To evaluate the quality of probabilities, one can use *scoring rules* (Murphy and Winkler 1970), which can be thought of as loss functions for probability estimates: A scoring rule provides a quality measure of the estimated class probabilities based on their deviation from the true underlying probabilities. A *proper scoring rule* returns the minimal expected loss when the true probabilities are estimated. Therefore, algorithms should aim to minimize proper scoring rules to provide sharp, refined, and calibrated probabilities.

In this section, we review and discuss the Brier score (Brier 1950), a well-known proper scoring rule to evaluate probability estimates in supervised classification.

Definition 4 (Brier Score) The Brier score BS is the mean squared difference between the outlier probabilities p_i and the ground-truth labels y_i, that is,

$$\mathrm{BS}(\boldsymbol{p}, \boldsymbol{y}) := \frac{1}{N} \sum_{i=1}^{N} \left(p_i - y_i\right)^2.$$

The Brier score uses ground-truth labels that indicate which observation is an outlier or an inlier and takes values between zero and one. A lower Brier score corresponds to better outlier probabilities. Perfect outlier probabilities with one for outliers and zero for inliers have a Brier score of zero. Conversely, flawed outlier probabilities with zero for outliers and one for inliers have a Brier score of one.

The Brier score can be decomposed into a *calibration term* and a *refinement term* in several ways. For example, Hernández-Orallo et al. (2012) derived an exact decomposition of the Brier score in continuous terms. The most common decomposition of the Brier score, however, remains the one introduced by Murphy (1972) for empirical distributions. In this decomposition, instead of summing over the observations in Definition 4 of the Brier score, one takes the sum over the different outlier probabilities, which gives

$$\mathrm{BS}(\boldsymbol{p}, \boldsymbol{y}) = \underbrace{\frac{1}{N} \sum_{i=1}^{K} n_i \overline{y_i}\left(1 - \overline{y_i}\right)}_{\text{refinement term}} + \underbrace{\frac{1}{N} \sum_{i=1}^{K} n_i \left(p_i - \overline{y_i}\right)^2}_{\text{calibration term}} \tag{1}$$

where K is the number of different outlier probabilities, n_i is the number of observations with outlier probability p_i, and $\overline{y_i}$ is the proportion of outliers in all observations with outlier probability p_i. Since the calibration and refinement terms are non-negative, the Brier score has its minimum value of zero when both decomposition terms have their minimum value of zero.

The refinement term measures how pure each group of observations with the same outlier probability is. To do so, the refinement term averages the product of the proportion of outliers $\overline{y_i}$ times the proportion of inliers $(1 - \overline{y_i})$ weighted by the number of observations n_i with outlier probability p_i. The refinement term has its minimum value of zero if, for each outlier probability, all observations with that outlier probability are either outliers or inliers; the refinement term is maximal if half of the observations are outliers and the other half are inliers. The refinement term of the Brier score in Equation (1) was also the motivation for Definition 2 (Refinement).

The calibration term measures how well the outlier probabilities $\boldsymbol{p}$ approximate the conditional probabilities $P(y = 1|s_i)$ in Definition 3 by averaging the squared difference between the outlier probability p_i and the proportion of outliers $\overline{y_i}$ in all observations with outlier probability p_i, weighted by the number of observations n_i with outlier probability p_i. The calibration term has its minimum value of zero if the outlier probabilities are equal to the proportion of outliers for all different outlier probabilities.

Many relationships between the Brier score and AUC ROC have been explored in the literature. For example, Flach and Matsubara (2007) derive a decomposition of the Brier score into calibration and refinement using ROC curves: They compute the calibration term using the empirical probabilities obtained from the slopes of the ROC curve segments. In contrast to the calibration and refinement decomposition,

another possible decomposition of the Brier score is the Brier curve (Hernández-Orallo et al. 2011). The Brier curve can be thought of as an example-wise decomposition of the Brier score, such that the area under the Brier curve is the Brier score. While the ROC curve evaluates the performance of rankings, ignoring the magnitude of the scores, the Brier curve is useful for representing probability estimates. Additionally, Hernández-Orallo et al. (2012) find that when using evenly spaced scores, the AUC ROC and the Brier score are equivalent (linearly related) performance metrics, i.e., the AUC ROC is equivalent to a Brier score that takes into account all evenly spaced scores.

It is unclear to what degree the Brier score measures sharpness, refinement, or calibration of outlier probabilities. For outlier detection, there are typically few outliers and, consequently, few observations with high outlier probability values. Thus, we usually have some observations that share their outlier probability with only a few other observations. For outlier probabilities with only a few observations, the interpretation of the refinement and calibration terms in Equation (1) is limited. For example, in the extreme case that the outlier probabilities of all observations are different, the refinement term in Equation (1) vanishes, and the calibration term is equal to the Brier score in Definition 4. In this situation, the Brier score measures that the outlier probabilities of the outliers are close to one and the outlier probabilities of the inliers are close to zero. Concentrating outliers around one and inliers around zero is a stricter property than sharpness because it also measures the correct concentration of probabilities for outliers at one and inliers at zero. Also, in our experiments in Sect. 9.2, we observe that the Brier score applied to outlier probabilities measures a mixture of sharpness, refinement, and calibration, where the contribution of each part is unclear.

Wallace and Dahabreh (2014) studied the quality of probability estimates for imbalanced binary classification tasks. Because the Brier score of Definition 4 calculates the average across all observations, with each observation weighted equally, the majority class can dominate the Brier score. For example, suppose we have an imbalanced dataset where 90% of the observations are inliers and 10% are outliers. If all outlier probabilities are zero, the Brier score is equal to 0.1, the Brier score computed only for the inliers is zero, and the Brier score computed only for the outliers is one. Thus, the outlier probability of the frequent inliers dominates the Brier score in this example. To measure the outlier probability of outliers and inliers separately, we adopt the stratified Brier score from Wallace and Dahabreh (2014) for outlier probabilities, that is, the Brier score calculated for inliers and outliers separately.

Definition 5 (Stratified Brier Score) The stratified Brier score for inliers BS^{inlier} is the Brier score BS restricted to inliers, that is,

$$\begin{aligned}\mathrm{BS}^{inlier}(\boldsymbol{p},\boldsymbol{y}) := \mathrm{BS}|_{y=0}(\boldsymbol{p},\boldsymbol{y}) &= \frac{1}{N^{inlier}} \sum_{\{i|y_i=0\}} (p_i - y_i)^2 \\ &= \frac{1}{N^{inlier}} \sum_{\{i|y_i=0\}} p_i^2\end{aligned}$$

where N^{inlier} is the number of inliers.

The stratified Brier score for outliers $\text{BS}^{outlier}$ is the Brier score BS restricted to outliers, that is,

$$\begin{aligned}\text{BS}^{outlier}(\boldsymbol{p},\boldsymbol{y}) := \text{BS}|_{y=1}(\boldsymbol{p},\boldsymbol{y}) &= \frac{1}{N^{outlier}} \sum_{\{i|y_i=1\}} (p_i - y_i)^2 \\ &= \frac{1}{N^{outlier}} \sum_{\{i|y_i=1\}} \big(p_i - 1\big)^2\end{aligned}$$

where $N^{outlier}$ is the number of outliers.

Like the Brier score, the stratified Brier score has values between zero and one, with lower values corresponding to better outlier probabilities. When comparing different outlier score transformations, the stratified Brier score makes it possible to compare the quality of the outlier probabilities for inliers and outliers separately.

Another widely used proper scoring rule for evaluating probability estimates in supervised learning is cross-entropy (Shuford Jr et al. 1966), also called log-loss or log-likelihood.

Definition 6 (Binary Cross-Entropy) The binary cross-entropy cH of the outlier probabilities $\boldsymbol{p}$ and the corresponding ground-truth labels $\boldsymbol{y}$ is

$$\text{cH}(\boldsymbol{p},\boldsymbol{y}) := \sum_{i=1}^{N} \big(-y_i \log(p_i) - (1-y_i)\log(1-p_i)\big).$$

According to the definition of proper scoring rules, the binary cross-entropy has its minimum value if the outlier probabilities $\boldsymbol{p}$ are equal to the ground-truth labels $\boldsymbol{y}$ and probability estimates closer to the true underlying probability are rewarded. Like the Brier score, the binary cross-entropy can be decomposed into a refinement term and a calibration term (Ramos et al. 2018). However, the binary cross-entropy is harsher than the Brier score in penalizing overconfident wrong predictions since it is unbounded for probabilities of outliers close to zero or probabilities of inliers close to one. In particular, binary cross-entropy is undefined for outliers with a probability of zero or inliers with a probability of one. This low quality of outlier probabilities can occur because of the imbalanced and often unsupervised nature of outlier detection. For example, suppose we convert the outlier scores of the k-Nearest Neighbors Detector (Ramaswamy et al. 2000) on the Ionosphere dataset (Campos et al. 2016) to outlier probabilities using Gaussian scaling (Kriegel et al. 2011). In this case, there are outliers with an outlier probability of zero, as shown in Fig. 1a. In this example, the binary cross-entropy is less suitable than the Brier score for evaluating the quality of outlier probabilities.

4.3 Calibration error

When comparing the strengths and weaknesses of outlier probabilities, it is desirable to measure sharpness, refinement, and calibration separately. Therefore, we first

adapt measures for evaluating the calibration of probability estimates in supervised learning and discuss their suitability for evaluating outlier probabilities.

For probability estimates in supervised learning, several measures have been proposed (Naeini et al. 2015; Nixon et al. 2019) to evaluate how much the probabilities $\boldsymbol{p}$ deviate from the conditional probabilities $P(y = 1|s_i)$ in Definition 3 (Calibration).

A major challenge in measuring calibration is that the conditional probabilities $P(y = 1|s_i)$ in Definition 3 are unknown because the conditional probability distribution P is unknown. We only evaluate outlier probabilities using a finite set of binary ground-truth labels. Following the idea of calibration plots, one approach is to divide the interval [0, 1] into bins and approximate the conditional probabilities $P(y = 1|s_i)$ by the proportion of outliers among observations with outlier probabilities in the same bin (DeGroot and Fienberg 1983). One can then define calibration measures based on the deviation between the average outlier probability and the proportion of outliers per bin (Naeini et al. 2015; Nixon et al. 2019).

In the following, we review ways to partition the interval [0, 1] into bins that are common to measure calibration (Naeini et al. 2015; Nixon et al. 2019). Let $B := \{b_i\}_{i=1}^{M}$ be a set of bins where $b_i \subseteq [0, 1]$ for all $1 \leq i \leq M$. Moreover, the union of all bins contains the outlier probabilities $\boldsymbol{p}$, that is, $\boldsymbol{p} \subseteq \bigcup_{i=1}^{M} b_i$. This implies that for every outlier score s_i with outlier probability $p_i = T(s_i)$ there exists a bin b_j such that $p_i \in b_j$. The number of outlier probabilities in the bin b_j is denoted by n_{b_j}.

There are several ways to divide the interval [0, 1] into bins. The most granular bins contain only equal outlier probabilities. In this case, one gets as many bins as there are unique values of outlier probabilities. Often, bins only containing equal outlier probabilities contain very few observations.

Equidistant bins, also called uniform bins, are commonly used and divide the interval [0, 1] into bins of equal length. For example, the intervals $[0.0, 0.1), [0.1, 0.2), \ldots, [0.9, 1.0]$ are ten equidistant bins.

Quantile bins each contain the same proportion of observations. For example, one can divide the interval [0, 1] into ten bins with the same number of observations per bin. Then, the first bin contains the observations with the smallest 10% of outlier probabilities, and the second bin contains the observations with the smallest 10% of the remaining outlier probabilities without the first bin.

In addition to categorizing bins based on how long they are and how many observations they contain, such as equidistant or quantile bins, we can also categorize them into overlapping and non-overlapping bins. Unlike non-overlapping bins described above, where different bins do not share observations, *overlapping bins* (Caruana and Niculescu-Mizil 2004) can share observations and require defining the degree of overlap between subsequent bins. For example, consider dividing the interval [0, 1] into quantile bins of 10 observations each and an overlap of 9 observations between subsequent bins. When sorting the outlier probabilities from low to high, the first bin contains the first 10 observations, the second bin contains the 2^{nd} to 11^{th} observation, the third bin contains the 3^{rd} to 12^{th} observation, until the last bin contains the last 10 observations. In the following, we assume

non-overlapping bins, which are commonly discussed in the literature (Naeini et al. 2015; Nixon et al. 2019).

As in the calibration plots (DeGroot and Fienberg 1983), we are interested in the average outlier probability and the proportion of outliers in each bin to measure the quality of outlier probabilities. The average outlier probability $\overline{p_{b_j}}$ of the bin b_j is

$$\overline{p_{b_j}} := \frac{1}{n_{b_j}} \sum_{\{i|p_i \in b_j\}} p_i$$

and the set of average outlier probabilities for the bins B is $\overline{\boldsymbol{p}_B} := \left\{\overline{p_{b_j}}\right\}_{j=1}^{M}$.

Similarly, the proportion of outliers $\overline{y_{b_j}}$ of the bin b_j is

$$\overline{y_{b_j}} := \frac{1}{n_{b_j}} \sum_{\{i|p_i \in b_j\}} y_i \tag{2}$$

and the set of proportions of outliers for the bins B is $\overline{\boldsymbol{y}_B} := \left\{\overline{y_{b_j}}\right\}_{j=1}^{M}$.

Following Vaicenavicius et al. (2019), we define the *binned calibration error* for a general distance function that measures the deviation between probability vectors.

Definition 7 (Binned Calibration Error) For a set of bins B and a distance function $d : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$, the binned calibration error CE_B^d of the probabilities $\boldsymbol{p}$ and ground-truth labels $\boldsymbol{y}$ is the distance between the average outlier probabilities $\overline{\boldsymbol{p}_B}$ and the proportions of outliers $\overline{\boldsymbol{y}_B}$ for the bins B; that is,

$$\mathrm{CE}_B^d(\boldsymbol{p}, \boldsymbol{y}) := d(\overline{\boldsymbol{p}_B}, \overline{\boldsymbol{y}_B}).$$

Because of the positivity property of the distance d, the binned calibration error has its minimum value of zero if and only if the proportions of outliers $\overline{y_{b_j}}$ are equal to the average outlier probabilities $\overline{p_{b_j}}$ for each bin. The smaller the binned calibration error, the better the outlier probabilities are calibrated.

We obtain different calibration errors for different choices of distance functions d and bins B. For the maximum norm and non-overlapping equidistant bins, Definition 7 is the *maximum calibration error* MCE defined by Naeini et al. (2015):

$$\mathrm{MCE} := \mathrm{CE}_{B_{equidis}}^{\max} = \max_{j=1}^{M} r_{b_j}$$

where $r_{b_j} := \left|\overline{p_{b_j}} - \overline{y_{b_j}}\right|$ are the residuals between the average outlier probabilities $\overline{p_{b_j}}$ and the proportion of outliers $\overline{y_{b_j}}$ per bin b_j. Visually, the maximum calibration error is the maximum vertical deviation between the diagonal and the calibration line in calibration plots, such as in Fig. 1b.

The residuals r_{b_j} can be volatile, because of the low proportion of outliers. In Fig. 1b, for example, the proportion of outliers in the 4^{th} bin is much closer to the diagonal than the proportion of outliers in the 5^{th} to 10^{th} bin. Since only the largest deviation between the average probabilities $\overline{\boldsymbol{p}_B}$ and proportions of outliers $\overline{\boldsymbol{y}_B}$

per bin determines the maximum calibration error, the maximum calibration error may be too pessimistic for measuring the calibration of outlier probabilities.

A common choice to measure the deviation between the average probabilities $\overline{\boldsymbol{p}_B}$ and the proportions of outliers $\overline{\boldsymbol{y}_B}$ is the p-norm (without the p-th root) weighted by the number of outlier probabilities per bin (Naeini et al. 2015; Nixon et al. 2019; Vaicenavicius et al. 2019).

Definition 8 (L^p Calibration Error) The L^p calibration error $\mathrm{CE}_B^{L^p}$ is the p-norm to the power of p between the average outlier probabilities $\overline{p_{b_j}}$ and the proportions of outliers $\overline{y_{b_j}}$ weighted by the number of outlier probabilities n_{b_j} per bin; that is,

$$\mathrm{CE}_B^{L^p}(\boldsymbol{p},\boldsymbol{y}) := \frac{1}{N}\sum_{j=1}^{M} n_{b_j} r_{b_j}^p = \frac{1}{N}\sum_{j=1}^{M} n_{b_j}\left|\overline{p_{b_j}} - \overline{y_{b_j}}\right|^p.$$

The L^p calibration error is between zero and one. Increasing p makes the L^p calibration error less sensitive to changes in r_{b_j} for r_{b_j} close to zero and more sensitive to changes in r_{b_j} for r_{b_j} close to one.

By choosing the bins B such that there are only equal outlier probabilities in the same bin, the L^2 calibration error (Definition 8) is equal to the calibration term of the Brier score decomposition in Equation (1).

For the 1-norm and non-overlapping equidistant bins, the L^p calibration error is called the *expected calibration error* (Naeini et al. 2015); for the 1-norm and non-overlapping quantile bins, we have the *adaptive calibration error* (Nixon et al. 2019); and for the 1-norm and overlapping quantile bins, we have the calibration measure *CAL* (Caruana and Niculescu-Mizil 2004). The expected, adaptive, and CAL calibration errors still depend on the number of bins M. As mentioned above, CAL also depends on the degree of overlap between subsequent bins.

Instead of binning the outlier probabilities when determining the calibration error, Blasiok and Nakkiran (2023) have proposed a calibration error using kernel smoothing. Because the probability mass of kernels, such as the Gaussian kernel, is unbounded and not uniformly distributed, smoothing around outlier probabilities near the boundary of the interval [0, 1] can be affected differently than smoothing around outlier probabilities in the center of the interval [0, 1]. To avoid distortions near the boundary of the interval [0, 1] when smoothing outlier probabilities, Blasiok and Nakkiran (2023) use a *reflected Gaussian kernel*. Intuitively, the reflected Gaussian kernel folds the probability mass of the unbounded Gaussian kernel into the interval [0, 1].

Definition 9 (Reflected Gaussian Kernel) The reflected Gaussian kernel $\widetilde{\phi}_\sigma : [0,1]\times[0,1] \to \mathbb{R}$ with scale σ on the interval [0, 1] is

$$\widetilde{\phi}_\sigma(x,y) := \sum_{\tilde{x}\in\pi^{-1}(x)} \phi_\sigma(\tilde{x}-y) = \sum_{\tilde{y}\in\pi^{-1}(y)} \phi_\sigma(x-\tilde{y})$$

where $\pi : \mathbb{R} \to [0,1]$ is the projection

$$\pi(x) := \begin{cases} x \mod 2 & \text{if } x \mod 2 \leq 1 \\ 2 - (x \mod 2) & \text{otherwise.} \end{cases}$$

and $\phi_\sigma : \mathbb{R} \to \mathbb{R}$ is the probability density function of the normal distribution centered at zero with scale σ:

$$\phi_\sigma(x) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right).$$

The set $\pi^{-1}(x)$ contains all numbers that can be mapped to x by a composition of reflections around integers.

Following Blasiok and Nakkiran (2023), we define the smooth calibration error for a finite set of outlier probabilities.

Definition 10 (Smooth Calibration Error) The smooth calibration error smCE_σ of the outlier probabilities $\boldsymbol{p}$ and the corresponding ground-truth labels $\boldsymbol{y}$ is

$$\text{smCE}_\sigma(\boldsymbol{p}, \boldsymbol{y}) := \int_0^1 \left| \frac{\sum_{i=1}^N \widetilde{\phi}_\sigma(t, p_i)(y_i - p_i)}{\sum_{i=1}^N \widetilde{\phi}_\sigma(t, p_i)} \right| dt$$

where $\widetilde{\phi}_\sigma$ is the reflected Gaussian kernel on the interval [0, 1] and the scale σ is chosen such that

$$\text{smCE}_\sigma(\boldsymbol{p}, \boldsymbol{y}) = \sigma. \tag{3}$$

Blasiok and Nakkiran (2023) show that there exists a unique scale σ that satisfies Equation (3).

In the following, we will focus on the L^p calibration error that is often used in the literature (Naeini et al. 2015; Nixon et al. 2019).

5 Measures for outlier probabilities

We propose error measures to independently assess how much outlier probabilities deviate from ideal sharp and refined outlier probabilities.

5.1 Sharpness error

To measure sharpness (Definition 1), we determine the degree to which outlier probabilities $\boldsymbol{p}$ deviate from ideal outlier probabilities concentrated around zero or one. If the probability p_i of being an outlier is close to zero, then the probability $1 - p_i$ of being an inlier is close to one, and vice versa. Therefore, for sharp outlier probabilities, the probability p_i deviates more from the probability $1 - p_i$ than for less sharp outlier probabilities. Consequently, to measure sharpness, we can determine the *purity* of the probabilities p_i and $1 - p_i$, that is, how different the two probabilities are. The *sharpness error* of the outlier probabilities $\boldsymbol{p}$ is then

the average purity of the probabilities p_i and $1 - p_i$ for a measure of purity for probabilities with binary outcomes. Without loss of generality, we assume that the purity measures below have values between zero and one, where lower values corresponds to higher purity and higher values to lower purity.

Definition 11 (Sharpness Error) The sharpness error SE is the average purity of the probability p_i of being an outlier and the probability $(1 - p_i)$ of being an inlier, that is,

$$\mathrm{SE}(\boldsymbol{p}) := \frac{1}{N} \sum_{i=1}^{N} \mathrm{PM}(p_i)$$

for a measure of purity PM for probabilities with binary outcomes.

The sharpness error has values between zero and one, with lower values corresponding to better outlier probabilities.

There are several purity measures for probabilities with binary outcomes. In general, for any distance function $d : [0, 1] \times [0, 1] \to [0, 1]$ we can define a purity measure $1 - d(p, 1 - p)$ for the probability p. Many purity measures can be derived from proper scoring rules (Buja et al. 2005). Common measures of purity for probabilities with two outcomes are the *binary entropy* (MacKay 2003) derived from binary cross-entropy (Definition 6), the *Gini index* (Hastie et al. 2009) derived from Brier score (Definition 4), and the *misclassification error* (Hastie et al. 2009).

Definition 12 (Purity Measures) Let p be a probability of an event with two outcomes.

1. The binary entropy of the probability p is

$$\mathrm{H}(p) := \begin{cases} -p \log_2(p) - (1-p) \log_2(1-p) & \text{if } p \in (0, 1) \\ 0 & \text{if } p = 0 \text{ or } p = 1. \end{cases}$$

2. The Gini index of the probability p is

$$\mathrm{G}(p) := 4p(1 - p).$$

3. The misclassification error of the probability p is

$$\mathrm{M}(p) := 2(1 - \max(p, 1 - p)).$$

We have scaled the Gini index and the misclassification error to cover the entire interval [0, 1]. For the binary entropy, the Gini index, or the misclassification error, the sharpness error is zero if all outlier probabilities are zero or one; the sharpness error has its maximum value of one if all outlier probabilities are 0.5. As discussed by Hastie et al. (2009), the binary entropy and the Gini index are more sensitive to changes for probabilities p close to zero or one than the misclassification error.

The sharpness error depends on the outlier probabilities $\boldsymbol{p}$ and is independent of the ground-truth labels $\boldsymbol{y}$. Thus, one could evaluate outlier probabilities using the sharpness error in the absence of ground-truth labels. Whether the sharpness error is a useful unsupervised or internal measure of outlier probabilities requires further investigation.

5.2 Refinement error

For refinement (Definition 2), we need a suitable definition of similarity for outlier probabilities. As for the binned calibration error (Definition 7), we consider outlier probabilities to be similar for refinement if they are in the same bin.

The sharpness error (Definition 11) measures the average purity of the probabilities $\boldsymbol{p}$ and $1-\boldsymbol{p}$. Similarly, we argue below to measure refinement by the average purity of the ground-truth labels $\boldsymbol{y}$ for outlier probabilities in the same bin. To quantify refinement, we determine the degree to which outlier probabilities in the same bin deviate from the ideal outlier probabilities, which belong either all to outliers or inliers. For a bin b_j, more refined outlier probabilities have a proportion of outliers $\overline{y_{b_j}}$ closer to zero or one than less refined outlier probabilities. If the proportion of outliers $\overline{y_{b_j}}$ is close to zero, then the proportion of inliers $1-\overline{y_{b_j}}$ is close to one, and the other way around. Consequently, refined outlier probabilities have more pure proportions of outliers $\overline{y_{b_j}}$ and inliers $1-\overline{y_{b_j}}$ per bin b_j than less refined outlier probabilities.

Definition 13 (Binned Refinement Error) The binned refinement error RE_B is the average purity of the proportion of outliers $\overline{y_{b_j}}$ per bin weighted by the number of outlier probabilities n_{b_j} per bin; that is,

$$\mathrm{RE}_B(\boldsymbol{p},\boldsymbol{y}) := \frac{1}{N}\sum_{j=1}^{M} n_{b_j}\,\mathrm{PM}\left(\overline{y_{b_j}}\right)$$

for some purity measure PM.

The purity measures discussed for the sharpness error (Definition 12) are also candidates for the binned refinement error. For the binary entropy, the Gini index, or the misclassification error, the binned refinement error has a maximum value of one if all bins contain half outliers and half inliers; the refinement error has a minimum value of zero if each bin contains only outliers or only inliers.

The binned refinement error implicitly depends on the outlier probabilities $\boldsymbol{p}$ because we select the ground-truth labels $\boldsymbol{y}$ to compute $\overline{\boldsymbol{y}_B}$ (Equation 2) based on their outlier probabilities $\boldsymbol{p}$. In addition, the bins B may depend on the outlier probabilities $\boldsymbol{p}$, such as for quantile bins.

For the Gini index and by choosing the bins B of the binned refinement error (Definition 13) such that there are only equal outlier probabilities in the same bin, the refinement error is a multiple of the refinement term of the Brier score decomposition in Equation (1).

Sharpness and binned refinement errors could be defined differently. First, instead of averaging the purity values $\mathrm{PM}(p_i)$ for the sharpness error (Definition 11) and the binned refinement error (Definition 13), we could aggregate them in some other way, such as taking the maximum or the median. We chose to use the average because it is more robust to outlier probabilities with extreme values than the maximum, and more sensitive to changes in outlier probabilities than the median. Second, we could also define sharpness and refinement errors using measures of impurity, since any measure of impurity is also a measure of purity.

Analogous to the smooth calibration error (Definition 10), we define a smooth refinement error that avoids the binning of outlier probabilities.

Definition 14 (Smooth Refinement Error) The smooth refinement error smRE_σ of the outlier probabilities $\boldsymbol{p}$ and the corresponding ground-truth labels $\boldsymbol{y}$ for a purity measure PM is

$$\mathrm{smRE}_\sigma(\boldsymbol{p}, \boldsymbol{y}) := \int_0^1 \mathrm{PM}\left(\frac{\sum_{i=1}^N \widetilde{\phi}_\sigma(t, p_i) y_i}{\sum_{i=1}^N \widetilde{\phi}_\sigma(t, p_i)} \right) dt$$

where $\widetilde{\phi}_\sigma$ is the reflected Gaussian kernel on the interval [0, 1].

Whether there is a similar canonical choice for the scale σ of the smooth refinement error smRE_σ as for the smooth calibration error smCE_σ requires further investigation.

6 Stratified and weighted measures

In Sect. 4.2, we argued, following Wallace and Dahabreh (2014), that the probabilities of the inliers can dominate the Brier score. Therefore, Wallace and Dahabreh (2014) proposed the stratified Brier score (Definition 5). Analogously to the Brier score, the sharpness, refinement, and L^p calibration errors measure the average quality of outlier probabilities, weighting each observation equally. Since there are typically significantly more inliers than outliers, the inliers can dominate the sharpness, refinement, and L^p calibration errors. In the following, we compute the measures for outliers and inliers separately to evaluate the quality of the probabilities for inliers or outliers only.

The binned refinement error (Definition 13) and the L^p calibration error (Definition 8) are the average of functions depending on the average outlier probabilities $\overline{p_{b_j}}$ and the proportion of outliers $\overline{y_{b_j}}$ weighted by the number of observations n_{b_j} per bin b_j. To measure the error for outliers and inliers separately, we can change the weight per bin according to the number of outliers or inliers in each bin. The following definition of *stratified measures* generalizes this idea.

Definition 15 (Stratified Binned Measure) For a measure EM_B evaluating outlier probabilities $\boldsymbol{p}$ using binary ground-truth labels $\boldsymbol{y}$, a binning B and that can be written as

$$\mathrm{EM}_B(\boldsymbol{p},\boldsymbol{y}) = \frac{1}{N}\sum_{j=1}^{M} n_{b_j} f(\overline{p_{b_j}}, \overline{y_{b_j}}) \tag{4}$$

for some real-valued function f, we define its stratified measure for inliers EM_B^{inlier} by weighting each bin by the number of inliers instead of the number of observations n_{b_j} in the bin b_j; that is,

$$\mathrm{EM}_B^{inlier}(\boldsymbol{p},\boldsymbol{y}) := \frac{1}{N^{inlier}}\sum_{j=1}^{M} n_{b_j}^{inlier} f(\overline{p_{b_j}}, \overline{y_{b_j}})$$

where N^{inlier} is the number of inliers and $n_{b_j}^{inlier}$ is the number of inliers in bin b_j.

The stratified measure for outliers $\mathrm{EM}_B^{outlier}$ is

$$\mathrm{EM}_B^{outlier}(\boldsymbol{p},\boldsymbol{y}) := \frac{1}{N^{outlier}}\sum_{j=1}^{M} n_{b_j}^{outlier} f(\overline{p_{b_j}}, \overline{y_{b_j}})$$

where $N^{outlier}$ is the number of outliers and $n_{b_j}^{outlier}$ is the number of outliers in bin b_j.

The stratified measures (Definition 15) include the stratified variant of the sharpness error (Definition 11): when we choose bins such that there are only equal outlier probabilities in the same bin, then Equation (4) for $f(\overline{p_{b_j}}, \overline{y_{b_j}}) = \mathrm{PM}(\overline{p_{b_j}}) = \mathrm{PM}(p_j)$ is equal to the sharpness error (Definition 11).

The stratified measures (Definition 15) also generalize the stratified Brier score (Definition 5): when we choose bins such that there are only equal outlier probabilities in the same bin, then the Brier score decomposition in Equation (1) has the form as in Equation (4) for $f(\overline{p_{b_j}}, \overline{y_{b_j}}) = f(\overline{p_j}, \overline{y_j}) = \overline{y_j}\left(1-\overline{y_j}\right) + \left(p_j - \overline{y_j}\right)^2$.

Similar to the approach taken to stratify measures that bin the outlier probabilities (Definition 15), we define stratified measures for the smooth calibration error (Definition 10) and the smooth refinement error (Definition 14).

Definition 16 (Stratified Smooth Measures) For a measure EM_K that evaluates outlier probabilities $\boldsymbol{p}$ using binary ground-truth labels $\boldsymbol{y}$ of the form

$$\mathrm{EM}_K(\boldsymbol{p},\boldsymbol{y}) = \int_0^1 f\left(\frac{\sum_{i=1}^N K(t,p_i)g(p_i,y_i)}{\sum_{i=1}^N K(t,p_i)}\right) dt \tag{5}$$

for real-valued functions f and g and a kernel K, we define its stratified measure for inliers EM_K^{inlier} by restricting the measure EM_K to inliers; that is,

$$\begin{aligned}\mathrm{EM}_K^{inlier}(\boldsymbol{p},\boldsymbol{y}) &:= \mathrm{EM}_K|_{\boldsymbol{y}=0}(\boldsymbol{p},\boldsymbol{y}) \\ &= \int_0^1 f\left(\frac{\sum_{\{i|y_i=0\}} K(t,p_i)g(p_i,y_i)}{\sum_{\{i|y_i=0\}} K(t,p_i)}\right)dt.\end{aligned}$$

The stratified measure for outliers $\mathrm{EM}_K^{outlier}$ is the measure EM_K restricted to outliers; that is,

$$\begin{aligned}\mathrm{EM}_K^{outlier}(\boldsymbol{p},\boldsymbol{y}) &:= \mathrm{EM}_K|_{\boldsymbol{y}=1}(\boldsymbol{p},\boldsymbol{y}) \\ &= \int_0^1 f\left(\frac{\sum_{\{i|y_i=1\}} K(t,p_i)g(p_i,y_i)}{\sum_{\{i|y_i=1\}} K(t,p_i)}\right)dt.\end{aligned}$$

The stratified measures (Definition 15 and Definition 16) evaluate the probabilities of outliers and inliers separately. However, examining only the stratified measure for outliers or inliers can be insufficient. For example, for outlier probabilities that are always one, the stratified Brier score for outliers is zero, but the stratified Brier score for inliers is one. Similarly, for outlier probabilities that are always zero, the stratified Brier score for inliers is zero, but the stratified Brier score for outliers is one.

Therefore, we summarize stratified measures for outliers and inliers into a single number using *weighted measures*, so that we can weight each stratified measure as required by the problem at hand.

Definition 17 (Weighted Measure) For a measure EM as in Equation (4) or Equation (5) that evaluates outlier probabilities $\boldsymbol{p}$ using binary ground-truth labels $\boldsymbol{y}$, its weighted measure EM_λ is the convex combination of the stratified measures for inliers EM^{inlier} and outliers $\mathrm{EM}^{outlier}$; that is,

$$\mathrm{EM}_\lambda := (1-\lambda)\,\mathrm{EM}^{inlier} + \lambda\,\mathrm{EM}^{outlier}$$

for $\lambda \in [0,1]$.

The weight λ depends on the problem. For example, in supervised learning, λ is commonly set according to the class proportion (Hernández-Orallo et al. 2012). Hernández-Orallo et al. (2012) also propose a prior-independent Brier score that weights both classes equally and corresponds to λ equal to 0.5 in Definition 17. However, if good probabilities for outliers are important, one can choose a higher value for λ. For λ equal to zero, the weighted measure EM_λ is equal to the stratified measure for inliers EM^{inlier}; for λ equal to one, the weighted measure is equal to the stratified measure for outliers $\mathrm{EM}^{outlier}$. If the stratified measures for outliers $\mathrm{EM}^{outlier}$ and inliers EM^{inlier} are equal, the weighted measure EM_λ is equal to the original measure EM.

The outlier probability measure proposed by Kriegel et al. (2011), as discussed in Sect. 2,

$$\frac{1}{2}\frac{1}{N^{inlier}}\sum_{\{i|y_i=0\}} p_i + \frac{1}{2}\frac{1}{N^{outlier}}\sum_{\{i|y_i=1\}}(1-p_i), \tag{6}$$

is related to the weighted Brier score $\mathrm{BS}_{0.5}$ (Definition 17). The measure in Equation (6) is a class-weighted average of the absolute differences $|p_i - y_i|$, and the weighted Brier score $\mathrm{BS}_{0.5}$ is a class-weighted average of the squared differences $(p_i - y_i)^2$.

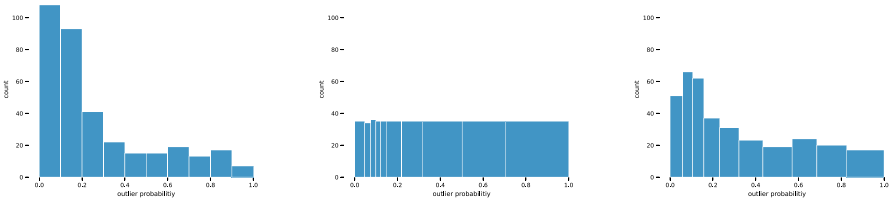
7 Binning outlier probabilities

The L^p calibration error (Definition 8) and the binned refinement error (Definition 13) require outlier probabilities to be in bins. We will discuss limitations of equidistant and quantile binning (see Sect. 4.2) of outlier probabilities and recommend a different binning approach.

As pointed out by Arrieta-Ibarra et al. (2022), "calibration errors based on binning intrinsically trade-off resolution for statistical confidence and vise versa." On the one hand, each bin should cover only a small part of the interval [0, 1], such that one accurately evaluates the outlier probabilities in that bin. On the other hand, each bin should contain enough outlier probabilities to compute stable estimates for that bin.

Figure 2 shows histograms of outlier probabilities computed using the *k*-Nearest Neighbors Detector (Ramaswamy et al. 2000) on the Ionosphere dataset (Campos et al. 2016). The outlier scores have been linearly scaled to the interval [0, 1], and each histogram has ten bins.

Figure 2a shows a histogram of outlier probabilities with ten non-overlapping equidistant bins. As expected, the number of observations in each equidistant bin varies: There are few outliers and, thus, few outlier probabilities near one. Consequently, equidistant bins near one contain few observations, in this example, less than ten observations, making it challenging to compute reliable properties for these bins.



(a) equidistant bins (b) quantile bins (c) equiareal bins

Fig. 2 Histograms for different bin types. We computed outlier scores for the Ionosphere dataset (Campos et al. 2016) using the *k*-Nearest Neighbors Detector (Ramaswamy et al. 2000). The outlier scores were linearly scaled to the probability interval [0, 1], and each histogram has ten non-overlapping bins. Equidistant bins may contain only a few observations, whereas quantile bins may cover wide intervals of different outlier probabilities. Equiareal bins provide a good balance between bin length and the number of observations per bin

Figure 2b shows a histogram of outlier probabilities with ten non-overlapping quantile bins. Although quantile bins all contain the same number of observations, the length of the bins varies. The higher-end bin covers approximately 30% of the interval [0, 1] because there are few outlier probabilities with large values. As a result, one only gets general information for the bins that cover a large part of the interval [0, 1].

To mitigate the limitations of equidistant and quantile bins for outlier probabilities, we recommend using *equiareal bins*, which divide the outlier probabilities such that the product of the number of observations per bin times the length of each bin is constant (wrkyle and farenorth 2016; Fung 2023a, b). The bars of histograms with equiareal bins all have approximately the same area. Compared to equidistant and quantile bins, equiareal bins have the advantage that the bin length and the number of observations per bin can vary. In other words, equiareal bins balance the trade-off between stable and accurate estimates for outlier probabilities more flexibly than equidistant or quantile bins.

Figure 2c shows a histogram of outlier probabilities with ten non-overlapping equiareal bins. As expected, the length and number of observations per bin vary for equiareal bins. We observe that each equidistant bin contains no fewer than approximately 20 observations, and each bin covers less than 10% of the interval [0, 1]. Thus, as this example illustrates and as we will show empirically in Sect. 9.1, equiareal bins are adequate to evaluate outlier probabilities in a detailed and reliable manner.

The L^p calibration error and the binned refinement error depend not only on the bin type, but also on the number of bins. How to set the number of bins for the L^p calibration error and the binned refinement error is generally unknown. If the evaluation approach is too sensitive to the number of bins, then the evaluation could be manipulated by choosing an appropriate number of bins. To reduce the dependence of the binned refinement and L^p calibration errors on the number of bins, we recommend computing the measures for different numbers of bins, such as 5, 6, 7 to 20, and reporting the center and scale of the measured error values, for example, using mean, median, standard deviation, or confidence intervals.

8 Experiments

We show empirically that the above measures are appropriate and helpful for evaluating outlier probabilities.

In the experiments below, we evaluated refinement using the binned refinement error (Definition 13) with the Gini index (Definition 12), as in the refinement term of the Brier score decomposition (Equation 1). For the sharpness error (Definition 11), we chose the binary entropy (Definition 12) because it is more sensitive to changes in outlier probabilities near zero and one than the Gini index (Hastie et al. 2009). We computed the refinement and L^1 calibration errors for 5, 6, 7 to 20 non-overlapping equiareal bins and reported their average.

8.1 Synthetic outlier probabilities

Experiments with synthetic outlier probabilities allow us to vary sharpness, refinement, and calibration and study the impact on the outlier probability measures.

Each synthetic dataset contains 1 000 outlier probabilities, where 900 belong to inliers and 100 belong to outliers. Following the observation of Gao and Tan (2006), we sample the outlier probabilities of inliers from an exponential distribution and the outlier probabilities of outliers from a Gaussian distribution centered at one. The outlier probabilities are calibrated with Platt scaling using a logistic sigmoid function (Platt et al. 1999).

We vary the sharpness and refinement of the synthetic outlier probabilities by changing the scale parameters of the exponential and Gaussian distributions. In the following, we refer to the rate parameter of the exponential distribution used to sample the inliers as *inlier sampling parameter*, and the standard deviation of the Gaussian distribution used to sample the outliers as *outlier sampling parameter*. In our experiments, both sampling parameters have values from 0.05, 0.15, 0.25 to 0.45.

Figure 3 shows synthetic outlier probabilities for equal sampling parameters of 0.05, 0.25, and 0.45 without calibration. The y-axis with the number of outlier probabilities has a logarithmic scale. As intended, the lower the outlier and inlier sampling parameters, the more sharp and refined the probabilities of the inliers and outliers are.

We change the degree of calibration of the synthetic outlier probabilities by multiplying the coefficients of the logistic sigmoid function used for Platt scaling by a *calibration factor* with values from 0.2, 0.4, 0.6 to 1.0. The higher the calibration factor, the better the outlier probabilities are calibrated. For a calibration factor of one, we obtain calibrated outlier probabilities. For all combinations of inlier and outlier sampling parameters and calibration factors, we have 125 synthetic sets of outlier probabilities.

Figure 4 shows synthetic outlier probabilities for different calibration factors. The outlier and inlier sampling parameters are both 0.15. The histograms in

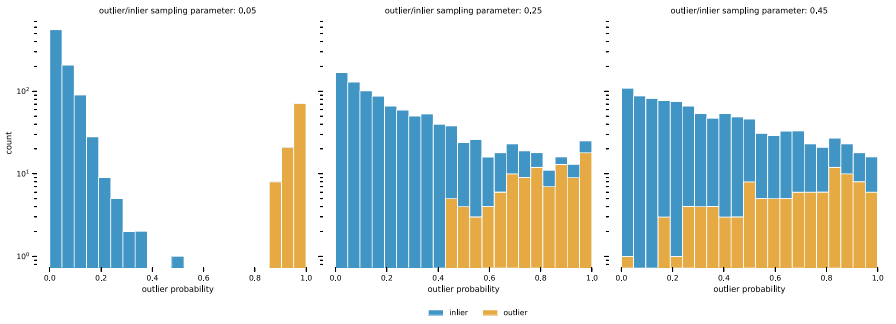
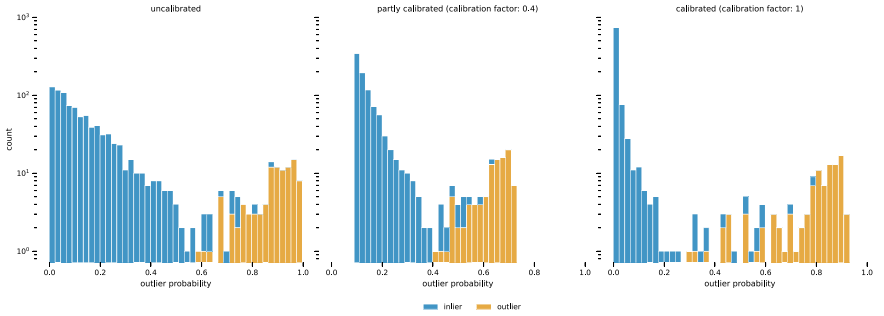
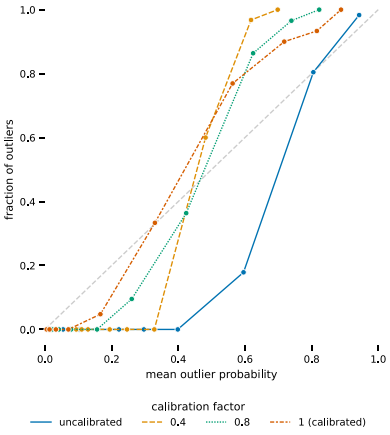


Fig. 3 Stacked histogram of synthetic outlier probabilities without calibration for different inlier and outlier sampling parameters. As intended, the lower the inlier and outlier sampling parameters, the more the outlier probabilities of inliers and outliers are sharp and refined (subfigures from right to left). The y-axes with the number of outlier probabilities have logarithmic scales

Fig. 4a illustrate that the higher the calibration factor, the more the outlier probabilities are pushed toward zero and one. Thus, the more we calibrate the synthetic outlier probabilities, the sharper they are, as discussed in Sect. 4.1. The y-axis with the number of outlier probabilities has a logarithmic scale. As expected, the calibration plot in Fig. 4b displays that the higher the calibration factor, the better



(a) Stacked histogram of uncalibrated (left), partially calibrated (middle) and calibrated (right) outlier probabilities. As expected, the more calibrated (i.e., the higher the calibration factor), the more the outlier probabilities are pushed toward zero and one. The y-axes with the number of outlier probabilities have logarithmic scales.



(b) Calibration plot with ten non-overlapping equiareal bins for outlier probabilities. Calibrated outlier probabilities lie close to the diagonal. As intended, the higher the calibration factor, the better the mean outlier probabilities match the proportion of outliers per bin.

Fig. 4 Synthetic outlier probabilities for different calibration factors. The outlier and inlier sampling parameters are both 0.15. The higher the calibration factor, the better the outlier probabilities are calibrated. For a calibration factor of one, we obtain calibrated probabilities

the mean outlier probabilities match the proportion of outliers per bin. The calibration plot has ten equiareal bins.

8.2 Real-world outlier probabilities

For the real-world experiments, we took 21 datasets from Campos et al. (2016), where we excluded the KDD and ALOI datasets for computational reasons. We computed outlier scores for all of the above datasets using 11 outlier detection algorithms implemented by Zhao et al. (2019) resulting in 231 sets of outlier scores. The outlier detection algorithms are: Principal Component Analysis (Shyu et al. 2003), Kernel Principal Component Analysis (Hoffmann 2007), Gaussian Mixture Model (Dempster et al. 1977), k-Nearest Neighbors Detector (Ramaswamy et al. 2000), Local Outlier Factor (Breunig et al. 2000), Isolation Forest (Liu et al. 2012), Histogram-based outlier score (Goldstein and Dengel 2012), Lightweight on-line detector of anomalies (Pevný 2016), Connectivity-Based Outlier Factor (Tang et al. 2002), Outlier detection based on Sampling (Sugiyama and Borgwardt 2013), and Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions (Li et al. 2023). For the hyperparameters of the outlier detection algorithms, we choose the default values according to Zhao et al. (2019).

8.2.1 Varying sharpness, refinement, and calibration

As with the synthetic outlier probabilities, we vary the sharpness, refinement, and calibration of the real-world outlier probabilities to examine the effect on the outlier probability measures.

First, we transform the 231 sets of outlier scores into outlier probabilities using linear scaling (Kriegel et al. 2011).

We then change the refinement and sharpness of the probabilities of outliers using an *outlier scaling parameter* and of inliers using an *inlier scaling parameter*: For an outlier scaling parameter $\alpha_{outlier}$ and an inlier scaling parameter α_{inlier}, we change an outlier probability probability p by

$$p \mapsto \begin{cases} 1 - \alpha_{outlier}(1-p) & \text{if } p \text{ is an outlier} \\ \alpha_{inlier} p & \text{if } p \text{ is an inlier.} \end{cases}$$

Thus, for an outlier and inlier scaling parameter of one, the probabilities of outliers and inliers are not changed; for an outlier scaling parameter of zero, the probabilities of outliers are one; and for an inlier scaling parameter of zero, the probabilities of inliers are zero. In general, as with the outlier and inlier sampling parameters for synthetic outlier probabilities, the lower the outlier or inlier scaling parameter, the sharper and more refined the real-world probabilities of outliers and inliers, respectively.

Finally, we vary the degree of calibration of the real-world outlier probabilities using Platt scaling (Platt et al. 1999) with a *calibration factor*, as for the synthetic outlier probabilities (see Sect. 8.1).

The outlier and inlier scaling parameters and the calibration factor for the real-world outlier probabilities ranged from 0.2, 0.4, 0.6 to 1.0.

8.2.2 Follow-up task

We study the impact of outlier probabilities generated on real-world datasets on follow-up tasks. As a follow-up task, we transform outlier probabilities into binary labels for outliers and inliers.

We converted the 231 real-world outlier scores to outlier probabilities using four outlier score transformations: ExCeeD (Perini et al. 2021), Gaussian scaling (Kriegel et al. 2011), gamma scaling (Kriegel et al. 2011), and linear scaling (Kriegel et al. 2011). In total, this results in 924 sets of outlier probabilities.

For each set of outlier scores computed by an outlier detection algorithm on a dataset, we ranked the outlier probabilities generated by the outlier score transformation using the weighted Brier score, calibration, refinement, and sharpness errors for weights λ between zero and one in Definition 17 (Weighted Measures).

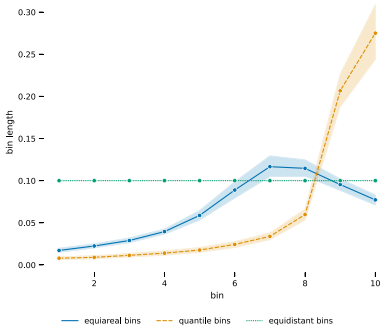
Following Gao and Tan (2006), we label observations with outlier probabilities greater than 0.5 as outliers and observations with outlier probabilities equal to or less than 0.5 as inliers. As argued by Gao and Tan (2006) using Bayesian risk models with a zero–one loss, this means that misclassifying outliers as inliers has the same cost as misclassifying inliers as outliers. For each set of outlier scores computed by an outlier detection algorithm on a dataset, we ranked the outlier labels generated from the outlier score transformations by the classification measures precision, recall, specificity, and F_1 score.

Finally, we examine the relationship between the outlier score transformations ranked by the quality of the outlier probabilities and the quality of the outlier and inlier labels. For each set of outlier scores and weights λ between zero and one, we computed the Spearman's rank correlation between the outlier score transformations ranked by the weighted outlier probability measures and the outlier score transformations ranked by the classification measures that evaluate the labels. We then averaged the correlation coefficients for each combination of outlier probability measure and outlier label measure across the sets of outlier scores.

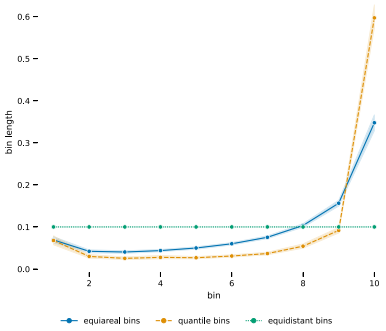
9 Results

9.1 Which bin types are appropriate for the refinement and calibration errors of outlier probabilities?

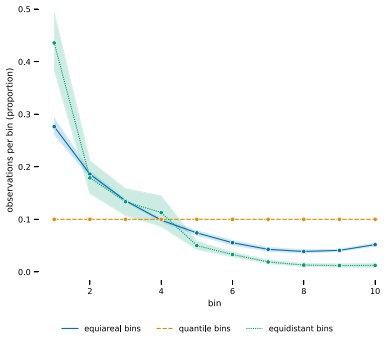
First, we examine the influence of bin types on bin length and number of observations per bin for synthetic and real-world outlier probabilities. The top row of Fig. 5 shows the bin length for the 125 synthetic (Fig. 5a) and 231 real-world (Fig. 5b) sets of outlier probabilities, which we scaled linearly to the interval [0, 1], for the 1^{st} to 10^{th} equidistant, quantile, and equiareal bins; the bottom row shows the proportion of synthetic (Fig. 5c) and real-world (Fig. 5d) sets of outlier probabilities per



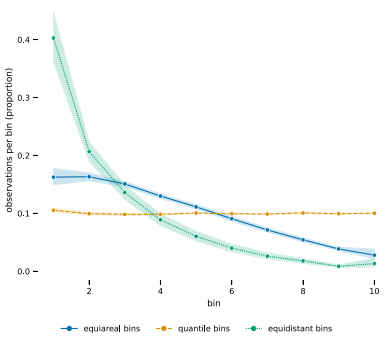
(a) bin length for synthetic outlier probabilities



(b) bin length for real-world outlier probabilities



(c) proportion of observations per bin for synthetic outlier probabilities



(d) proportion of observations per bin for real-world outlier probabilities

Fig. 5 Length (top row) and proportion of observations (bottom row) per bin of the 1^{st} to 10^{th} bin for different bin types. We divide each of the 125 synthetic (left column) and 231 real-world sets of outlier probabilities, which we scaled linearly to the interval [0, 1] (right column), into ten bins and report the average (points) and 95% confidence interval (area) of the length and proportion of observations per bin. The y-axes have different scales. **Top row:** The 10^{th} quantile bin contains, on average, more than 25% of the interval [0, 1] for synthetic and nearly 60% for real-world outlier probabilities. Equiareal bins contain, on average, less than 15% of the interval [0, 1] for synthetic and less than 35% for real-world outlier probabilities. **Bottom row:** The 1^{st} equidistant bin contains, on average, more than 40% of the synthetic and approximately 40% of the real-world outlier probabilities. Using equiareal bins reduces this to less than 30% for the synthetic and less than 20% for the real-world outlier probabilities

bin for the different bin types. We divide each synthetic and real-world set of outlier probabilities into ten bins and report the average and 95% confidence interval of the length and proportion of observations per bin.

On average, the 10^{th} quantile bin covers approximately 25% of the interval [0, 1] for the synthetic (Fig. 5a) and nearly 60% for the real-world outlier probabilities (Fig. 5b); equiareal bins contain, on average, only less than 15% of the interval [0, 1] for synthetic and less than 35% for real-world outlier probabilities. By design, each equidistant bin contains 10% of the interval [0, 1].

The 1st equidistant bin contains, on average, more than 40% of the observations for the synthetic (Fig. 5c) and approximately 40% of the observations for the real-world outlier probabilities (Fig. 5d). Equiareal bins reduce this to an average of less than 30% of the observations for synthetic and less than 20% of the observations for real-world outlier probabilities. As intended, each quantile bin contains 10% of the observations.

Overall, equiareal bins are short enough to accurately evaluate the entire interval [0, 1]; at the same time, each equiareal bin contains enough observations to compute stable measures per bin. Therefore, when evaluating outlier probabilities, we recommend using equiareal bins for binned refinement and L^p calibration errors.

9.2 Do the measures evaluate the desired characteristics of outlier probabilities?

We then examine if the Brier score, sharpness, refinement, and L^1 calibration errors evaluate the desired characteristics of outlier probabilities.

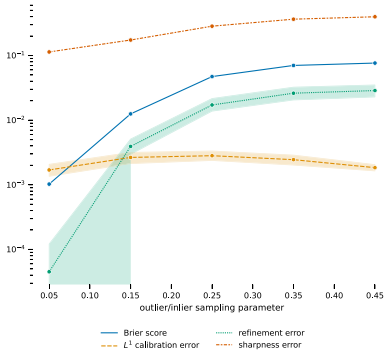
The top row of Fig. 6 shows the Brier score, sharpness, refinement, and L^1 calibration errors depending on the outlier and inlier sampling parameters for synthetic (Fig. 6a) and depending on the outlier and inlier scaling parameters for real-world (Fig. 6b) outlier probabilities. We changed the parameters for inliers and outliers simultaneously and calibrated all outlier probabilities with a calibration factor of one. The points are the means of the measures, and the area is the 95% confidence intervals over the bin numbers and sets of real-world outlier probabilities. The y-axes are logarithmic and have different scales.

The Brier score, sharpness, and refinement errors increase as the outlier and inlier sampling (Fig. 6a) and scaling (Fig. 6b) parameters increase. In contrast, the L^1 calibration error does not change much as the outlier and inlier sampling and scaling parameters increase.

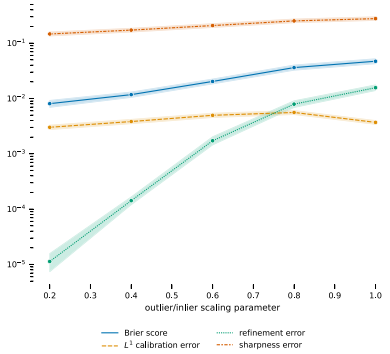
The bottom row of Fig. 6 shows the Brier score, sharpness, refinement, and L^1 calibration errors depending on the calibration factor of the synthetic (Fig. 6c) and real-world (Fig. 6d) outlier probabilities. The outlier and inlier sampling parameters for the synthetic outlier probabilities are both 0.25; the outlier and inlier scaling parameters for the real-world outlier probabilities are one.

The Brier score, L^1 calibration, and sharpness errors decrease as the calibration factor increases, with a higher calibration factor corresponding to better calibrated probabilities (Fig. 6c and Fig. 6d). The refinement error is almost independent of the investigated calibration factors.

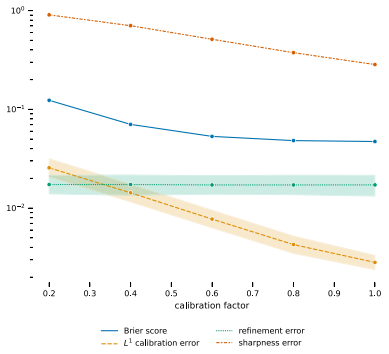
These results suggest that the L^1 calibration error measures only the calibration, not the refinement or sharpness of the outlier probabilities. As intended, the refinement error evaluates the mixture of outliers and inliers with similar probabilities and does not depend on the calibration of the outlier probabilities. The sharpness error is a good measure of the concentration of probabilities at zero or one. In our experiments, smaller sampling or scaling parameters or a higher calibration factor can cause a higher concentration of probabilities at zero or one. That calibration pushes



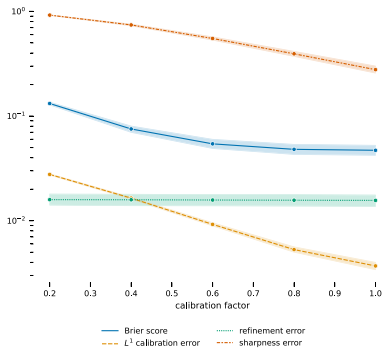
(a) measures depending on outlier and inlier sampling parameters for synthetic outlier probabilities



(b) measures depending on outlier and inlier scaling parameters for real-world outlier probabilities



(c) Measures depending on calibration factor for synthetic outlier probabilities. The outlier and inlier sampling parameters are both 0.25.



(d) Measures depending on calibration factor for real-world outlier probabilities. The outlier and inlier scaling parameters are both one.

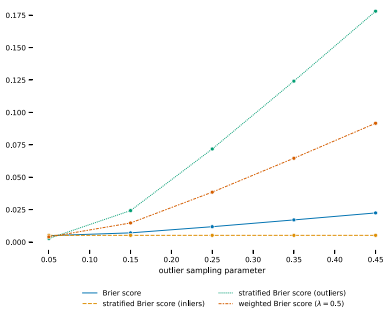
Fig. 6 Measures depending on the outlier and inlier sampling and scaling parameters (top row) and calibration factor (bottom row) of the synthetic (left column) and real-world (right column) outlier probabilities. The points are the means of the measures, and the area is the 95% confidence intervals over the bin numbers and sets of real-world outlier probabilities. The y-axes are logarithmic and have different scales. **Top row:** The Brier score, the refinement, and sharpness errors are appropriate measures of the sharpness and refinement of the outlier probabilities. The L^1 calibration error is almost independent of the sampling and scaling parameters. We have calibrated all outlier probabilities with a calibration factor of one. **Bottom row:** The Brier score, the L^1 calibration, and sharpness errors are appropriate measures of the calibration and sharpness of the outlier probabilities. The refinement error is not affected by the calibration factor

the outlier probabilities toward zero or one can also be observed in Fig. 4 and is discussed in Sect. 4.1. As expected, the Brier score measures a mixture of sharpness, refinement, and calibration.

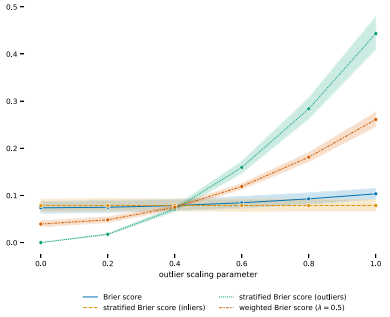
In summary, the sharpness, refinement, and L^1 calibration errors are suitable measures of the corresponding characteristics of the outlier probabilities. Therefore, we recommend evaluating outlier probabilities using the Brier score, sharpness, refinement, and L^1 calibration errors.

9.3 Do the stratified and weighted measures evaluate the probabilities of outliers well?

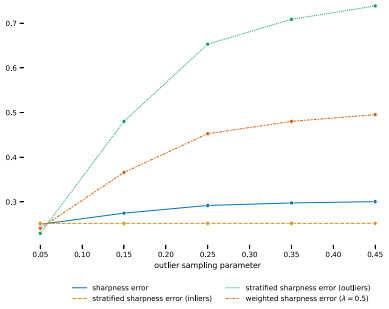
We now turn to the sensitivity of the measures to changes in the probabilities of outliers. Figure 7 shows the Brier score, sharpness, refinement, and L^1 calibration errors, as well as their stratified and weighted variants, depending on the outlier sampling



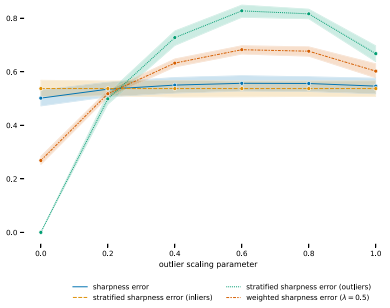
(a) Brier score for synthetic outlier probabilities



(b) Brier score for real-word outlier probabilities

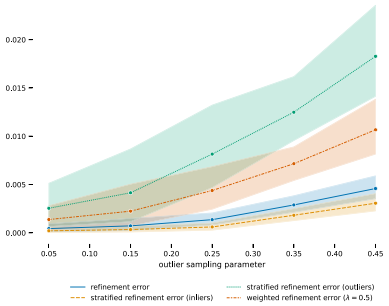


(c) sharpness error for synthetic outlier probabilities

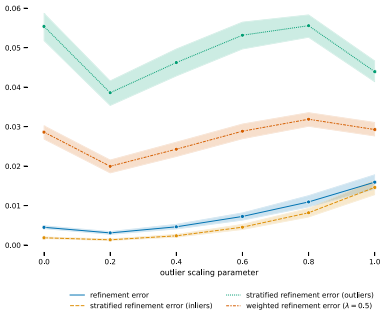


(d) sharpness error for real-word outlier probabilities

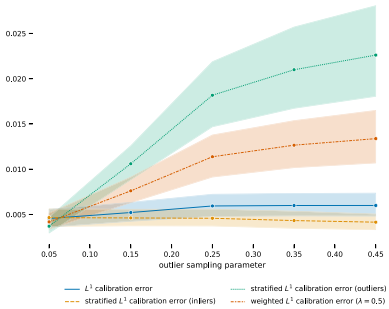
Fig. 7 Measures and their stratified and weighted variants for λ equal to 0.5 depending on the outlier sampling parameter for the synthetic outlier probabilities (left column) and depending on the outlier scaling parameter for the real-world outlier probabilities (right column). The inlier sampling parameter for the synthetic outlier probabilities is 0.05, the inlier scaling parameter for the real-world outlier probabilities is one, and the outlier probabilities are uncalibrated. The points are the means of the measures, and the area is the 95% confidence intervals over the bin numbers and sets of real-world outlier probabilities. The outlier probabilities of the inliers dominate the unstratified or unweighted measures. The weighted and stratified measures for outliers capture well changes in the probabilities of outliers. The y-axes do not start at the origin and have different scales.



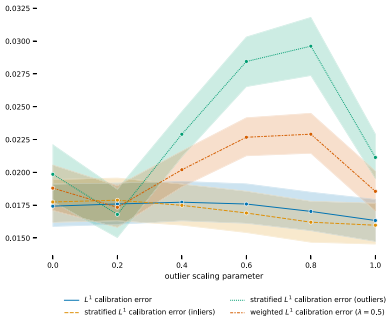
(e) refinement error for synthetic outlier probabilities



(f) refinement error for real-word outlier probabilities



(g) L^1 calibration error for synthetic outlier probabilities



(h) L^1 calibration error for real-word outlier probabilities

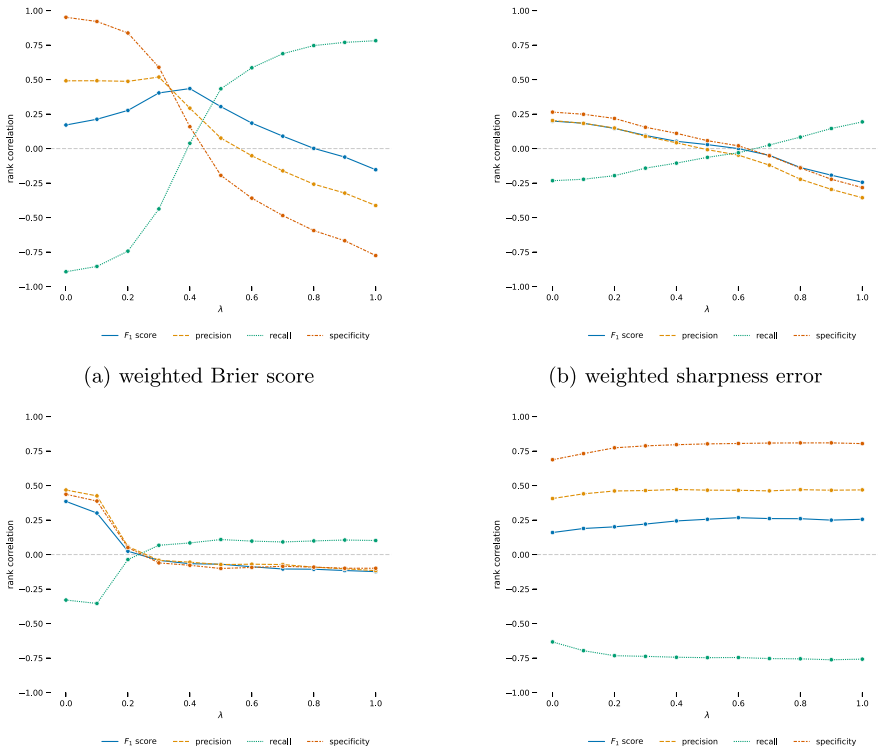
Fig. 7 (continued)

parameter for synthetic outlier probabilities (left column) and depending on the outlier scaling parameter for real-world outlier probabilities (right column). For the weighted variants of the outlier probabilities, we chose λ equal to 0.5: the stratified measures (Definition 17) for outliers and inliers are weighted equally. The inlier sampling parameter for synthetic outlier probabilities is 0.05; the inlier scaling parameter for real-world outlier probabilities is one; and the outlier probabilities are uncalibrated. The points are the means of the measures, and the area is the 95% confidence intervals over the bin numbers and sets of real-world outlier probabilities.

As expected, the unstratified or unweighted and the stratified measures for inliers are nearly independent of the outlier sampling and scaling parameters. In contrast, the stratified measures for outliers and the weighted measures are sensitive to changes in the probabilities of outliers. Therefore, we recommend using weighted and stratified measures to evaluate the probabilities of outliers.

9.4 Do good outlier probabilities, according to the proposed measures, improve the performance of follow-up tasks?

Finally, we examine the influence of the quality of outlier probabilities on the performance of follow-up tasks. As a follow-up task, we convert outlier probabilities to inlier and outlier labels. Figure 8 shows the average Spearman's rank correlation between the quality of the outlier probabilities, as assessed by the proposed weighted measures, and the quality of the outlier labels on the y-axes depending on the weight λ for the weighted outlier probability measures on the x-axes. For



(a) weighted Brier score

(b) weighted sharpness error

(c) weighted refinement error

(d) weighted L^1 calibration error

Fig. 8 Average Spearman's rank correlation between the quality of the outlier probabilities, as assessed by the proposed weighted measures, and the quality of the outlier labels on the y-axes depending on the weight λ for the weighted measures on the x-axes. We label observations with outlier probabilities greater than 0.5 as outliers and observations with outlier probabilities equal to or less than 0.5 as inliers. We computed outlier scores using 11 outlier detection algorithms on 21 different datasets. Outlier probabilities were transformed using ExCeeD, Gaussian scaling, gamma scaling, and linear scaling resulting in 924 sets of outlier probabilities. For each set of outlier scores, we computed the Spearman's rank correlation between the four sets of outlier probabilities ranked by the weighted probability measures and ranked by the classification metrics. For example, selecting the outlier score transformation by the best weighted Brier score (a), sharpness (b), refinement (c), or L^1 calibration (d) errors increases the F_1 score of the corresponding labels for λ less than 0.2

example, the blue solid line in Fig. 8a shows the average rank correlation between the F_1 score and the weighted Brier score on the y-axes for different weights λ on the x-axes.

For a smaller λ, selecting outlier probabilities by the weighted Brier score (Fig. 8a), the weighted sharpness error (Fig. 8b), or the weighted refinement error (Fig. 8c) results in a better F_1 score, precision, and specificity but decreases recall. For a larger λ, the weighted Brier score, sharpness error, and refinement error correlate positively with recall and negatively with F_1 score, precision, and specificity.

Selecting outlier probabilities by the weighted L^1 calibration error (Fig. 8d) results in a better F_1 score, precision, and specificity but decreases recall for all values of λ.

The high positive correlation between the weighted L^1 calibration error ranking and the specificity ranking, combined with the high negative correlation between the weighted L^1 calibration error ranking and the recall ranking, suggests that the outlier probabilities with low weighted L^1 calibration error label many observations as inliers. Consequently, many inliers are correctly labeled as inliers, and many outliers are incorrectly marked as inliers.

The results of our experiments indicate that selecting outlier probabilities using follow-up task-specific outlier probability measures, as we will discuss in more detail in Sect. 10, can improve the quality of the follow-up task.

10 Practical considerations

As mentioned in Sect. 4.1, there are pathological examples where sharpness, refinement, and calibration are independent or redundant. In general, however, the dependency between sharpness, refinement, and calibration when evaluating outlier probabilities is unclear and depends on the dataset and the outlier algorithm being studied. Often, improving only calibration, sharpness, or refinement may worsen the other properties. One typically tries to increase the refinement and sharpness of the outlier probabilities while keeping the probabilities calibrated, as mentioned in Sect. 1. Therefore, we suggest to use the Brier score together with the sharpness, refinement, and calibration errors. When outlier probabilities are used for a specific follow-up task, the choice of outlier probability measure can be weighted differently depending on the follow-up task.

In the following, we discuss practical considerations for the follow-up task of converting outlier probabilities into labels for outliers and inliers. When converting outlier probabilities into outlier labels using a threshold t, it is only important that the probabilities of outliers are greater than t and the probabilities of inliers are less than t; the values of the probabilities in the intervals $[0, t]$ or $(t, 1]$ are not relevant.

As mentioned in Sect. 4.2, the sharpness error only measures the concentration of probabilities around zero and one; it does not measure the concentration of outliers around one and inliers around zero. Consequently, two different outlier probabilities can have the same sharpness error, but the quality of the labels can be different. The Brier score, in contrast, also measures the desired concentration of outliers around

one and inliers around zero. For example, for a threshold of 0.5, two outliers with outlier probabilities of 0.3 and 0.7 have the same sharpness error using the binary entropy or the Gini index, but we incorrectly label the outlier with a probability of 0.3 as an inlier; the outlier with a probability of 0.7 is correctly labeled as an outlier. The outlier with a probability of 0.3 has a Brier score of 0.49; the outlier with a probability of 0.7 has a Brier score of 0.09. Thus, unlike the sharpness error, the Brier score evaluates the outlier probabilities in order to convert outlier probabilities into labels as desired. This may explain why the Brier score has a higher correlation with the quality of the outlier and inlier labels than the sharpness error with the labels in Sect. 9.4.

The purity of ground-truth labels for observations with an outlier probability in the same bin is less critical for converting probabilities into labels, since the refinement of outlier probabilities is independent of whether the outlier probabilities are greater than the threshold t or less than the threshold t. For example, a bin containing only outliers improves the quality of outlier and inlier labels if the bin is in the interval $(t, 1]$. If the bin containing only outliers is in the interval $[0, t]$, the quality of the outlier and inlier labels is reduced. In both cases, the refinement error of the bins containing only outliers is zero, but with an opposite effect on the label quality. In general, a higher proportion of outliers in a bin is positively related to the quality of the labels if the bin is in the interval $[t, 1]$. Similarly, a higher proportion of inliers in a bin is positively related to the quality of the labels if the bin is in the interval $[0, t)$.

Calibration requires that bins with a proportion of outliers larger than the threshold t have an average outlier probability larger than the threshold t; bins with a proportion of outliers less than the threshold t have an average outlier probability less than the threshold t. Thus, we expect better calibrated outlier probabilities to have better label quality.

For follow-up tasks other than converting outlier probabilities into labels, the outlier probability measures may have a different relevance. For example, as mentioned in Sect. 1, outlier probabilities can be used to normalize outlier scores from different outlier detection algorithms to build outlier ensembles. Outlier ensembles then aggregate the outlier probabilities, for example, by averaging them. The extent to which the outlier probabilities of different outlier detection algorithms are comparable can be measured by the calibration error.

11 Conclusions

We define *good* outlier probabilities as sharp, refined, and calibrated. To evaluate these characteristics, we adapt and propose novel measures that use the ground-truth labels that indicate which observation is an outlier or an inlier. Refinement and calibration measures require partitioning the outlier probabilities into bins or using kernel smoothing.

Compared to probability estimates in supervised learning, we argue and empirically show that several aspects are essential when evaluating outlier probabilities, mainly due to the imbalanced and often unsupervised nature of outlier detection. First, we advocate using stratified and weighted versions of the measures to evaluate

the probabilities of outliers well. Otherwise, the frequent inliers tend to dominate the measures. Second, we recommend using the sharpness, refinement, and calibration errors to independently evaluate the different characteristics of good outlier probabilities. Third, measures using equiareal bins make it possible to accurately evaluate different ranges of outlier probabilities.

Finally, we show empirically that good outlier probabilities, according to the proposed measures, improve the performance of converting outlier probabilities into labels for outliers and inliers.

In future work, we plan to investigate binning-free outlier probability measures using the empirical cumulative distribution function as an alternative to kernel smoothing (Gupta et al. 2020; Arrieta-Ibarra et al. 2022).

Funding Open Access funding enabled and organized by Projekt DEAL. This study was partly funded by the Independent Research Fund Denmark in the project Reliable Outlier Detection.

Declarations

Conflict of interest of potential Conflict of interest Arthur Zimek is an Action Editor for the journal Data Mining and Knowledge Discovery.

References

Achtert E, Kriegel H, Reichert L, et al. (2010) Visual evaluation of outlier detection models. In: DASFAA (2), Lecture Notes in Computer Science, vol 5982. Springer, pp 396–399

Arrieta-Ibarra I, Gujral P, Tannen J et al (2022) Metrics of calibration for probabilistic predictions. J Mach Learn Res 23(1):15886–15940

Barnett V, Lewis T et al (1994) Outliers in statistical data, vol 3. Wiley, New York

Bauder RA, Khoshgoftaar TM (2017) Estimating outlier score probabilities. In: 2017 IEEE International Conference on Information Reuse and Integration (IRI), IEEE, pp 559–568

Blasiok J, Nakkiran P (2023) Smooth ECE: Principled reliability diagrams via kernel smoothing. In: The Twelfth International Conference on Learning Representations

Bouguessa M (2012) Modeling outlier score distributions. In: ADMA, Springer, pp 713–725

Breunig MM, Kriegel H, Ng RT, et al. (2000) LOF: identifying density-based local outliers. In: SIGMOD Conference. ACM, pp 93–104

Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78(1):1–3

Buja A, Stuetzle W, Shen Y (2005) Loss functions for binary class probability estimation and classification: Structure and applications. Working draft, November 3

Campos GO, Zimek A, Sander J et al (2016) On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Min Knowl Disc 30:891–927

Campos GO, Zimek A, Jr. WM (2018) An unsupervised boosting strategy for outlier detection ensembles. In: PAKDD (1), Lecture Notes in Computer Science, vol 10937. Springer, pp 564–576

Caruana R, Niculescu-Mizil A (2004) Data mining in metric space: an empirical analysis of supervised learning performance criteria. In: KDD. ACM, pp 69–78
Clifton LA, Clifton DA, Zhang Y et al (2014) Probabilistic novelty detection with support vector machines. IEEE Trans Reliab 63(2):455–467
Dawid AP (1982) The well-calibrated bayesian. J Am Stat Assoc 77(379):605–610
DeGroot MH, Fienberg SE (1982) Assessing probability assessors: calibration and refinement. Statist Decis Theory Relat Top III 1:291–314
DeGroot MH, Fienberg SE (1983) The comparison and evaluation of forecasters. J R Statist Soc: Ser D (The Statistician) 32(1–2):12–22
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. J Roy Stat Soc: Ser B (Methodol) 39(1):1–22
Flach PA, Matsubara ET (2007) A simple lexicographic ranker and probability estimator. In: ECML, Lecture Notes in Computer Science, vol 4701. Springer, pp 575–582
Fung K (2023a) Equal-area histograms. https://junkcharts.typepad.com/junk_charts/2023/04/equal-area-histograms.html, accessed: 2024-05-24
Fung K (2023b) More on equal-area histograms. https://junkcharts.typepad.com/junk_charts/2023/05/more-on-equal-area-histograms.html, Accessed: 2024-05-24
Gao J, Tan PN (2006) Converting output scores from outlier detection algorithms into probability estimates. In: Sixth International Conference on Data Mining (ICDM'06), IEEE, pp 212–221
Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. J R Stat Soc Ser B Stat Methodol 69(2):243–268
Goldstein M, Dengel A (2012) Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. KI-2012: poster and demo track 1:59–63
Gupta K, Rahimi A, Ajanthan T, et al. (2020) Calibration of neural networks using splines. In: International Conference on Learning Representations
Hastie T, Tibshirani R, Friedman JH (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. Springer Series in Statistics, Springer
Hawkins DM (1980) Identification of Outliers. Springer, Monographs on Applied Probability and Statistics
Hernández-Orallo J, Flach PA, Ramirez CF (2011) Brier curves: a new cost-based visualisation of classifier performance. In: ICML. Omnipress, pp 585–592
Hernández-Orallo J, Flach PA, Ferri C (2012) A unified view of performance metrics: translating threshold choice into expected classification loss. J Mach Learn Res 13:2813–2869
Hoffmann H (2007) Kernel PCA for novelty detection. Pattern Recognit 40(3):863–874
Kriegel H, Kröger P, Schubert E, et al. (2009) LoOP: local outlier probabilities. In: Cheung DW, Song I, Chu WW, et al (eds) Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009. ACM, pp 1649–1652
Kriegel H, Kröger P, Schubert E, et al. (2011) Interpreting and unifying outlier scores. In: Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA. SIAM / Omnipress, pp 13–24
Kriegel H, Kröger P, Schubert E, et al. (2012) Outlier detection in arbitrarily oriented subspaces. In: ICDM. IEEE Computer Society, pp 379–388
Kull M, Silva Filho TM, Flach P (2017) Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. Electron J Statist 11:5052–5080
Li Z, Zhao Y, Hu X et al (2023) ECOD: unsupervised outlier detection using empirical cumulative distribution functions. IEEE Trans Knowl Data Eng 35(12):12181–12193
Liu FT, Ting KM, Zhou Z (2012) Isolation-based anomaly detection. ACM Trans Knowl Discov Data 6(1):3:1–3:39
MacKay DJC (2003) Information theory, inference, and learning algorithms. Cambridge University Press, Cambridge
Marques HO, Campello RJ, Sander J et al (2020) Internal evaluation of unsupervised outlier detection. ACM Trans Knowl Discov Data (TKDD) 14(4):1–42
Marques HO, Zimek A, Campello RJGB, et al. (2022) Similarity-based unsupervised evaluation of outlier detection. In: SISAP, Lecture Notes in Computer Science, vol 13590. Springer, pp 234–248
Muhr D, Affenzeller M, Küng J (2023) A probabilistic transformation of distance-based outliers. Mach Learn Knowl Extr 5(3):782–802
Murphy AH (1972) Scalar and vector partitions of the probability score: Part i. two-state situation. J Appl Meteorol 1962–1982:273–282

Murphy AH (1973) A new vector partition of the probability score. J Appl Meteorol Climatol 12(4):595–600
Murphy AH, Winkler RL (1970) Scoring rules in probability assessment and evaluation. Acta Physiol (Oxf) 34:273–286. https://doi.org/10.1016/0001-6918(70)90023-5
Naeini MP, Cooper G, Hauskrecht M (2015) Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the AAAI conference on artificial intelligence
Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning, pp 625–632
Nixon J, Dusenberry MW, Zhang L, et al. (2019) Measuring calibration in deep learning. In: CVPR workshops
Perini L, Vercruyssen V, Davis J (2021) Quantifying the confidence of anomaly detectors in their example-wise predictions. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III, Springer, pp 227–243
Pevný T (2016) Loda: Lightweight on-line detector of anomalies. Mach Learn 102(2):275–304
Platt J et al (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classifiers 10(3):61–74
Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large data sets. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp 427–438
Ramos D, Franco-Pedroso J, Lozano-Diez A et al (2018) Deconstructing cross-entropy for probabilistic binary classifiers. Entropy 20(3):208
Rayana S, Akoglu L (2016) Less is more Building selective anomaly ensembles. ACM Trans Knowl Discov Data 10(4):1–33
Röchner P, Rothlauf F (2023) Unsupervised anomaly detection of implausible electronic health records: a real-world evaluation in cancer registries. BMC Med Res Methodol 23(1):125
Ruff L, Kauffmann JR, Vandermeulen RA et al (2021) A unifying review of deep and shallow anomaly detection. Proc IEEE 109(5):756–795
Shuford EH Jr, Albert A, Edward Massengill H (1966) Admissible probability measurement procedures. Psychometrika 31(2):125–145
Shyu ML, Chen SC, Sarinnapakorn K, et al. (2003) A novel anomaly detection scheme based on principal component classifier. In: Proceedings of the IEEE foundations and new directions of data mining workshop, IEEE Press, pp 172–179
Sotiris VA, Tse PW, Pecht MG (2010) Anomaly detection through a bayesian support vector machine. IEEE Trans Reliab 59(2):277–286
Sugiyama M, Borgwardt K (2013) Rapid distance-based outlier detection via sampling. Advances in neural information processing systems 26
Tang J, Chen Z, Fu AW, et al. (2002) Enhancing effectiveness of outlier detections for low density patterns. In: PAKDD, Lecture Notes in Computer Science, vol 2336. Springer, pp 535–548
Vaicenavicius J, Widmann D, Andersson CR, et al. (2019) Evaluating model calibration in classification. In: AISTATS, Proceedings of Machine Learning Research, vol 89. PMLR, pp 3459–3467
Wallace BC, Dahabreh IJ (2014) Improving class probability estimates for imbalanced data. Knowl Inf Syst 41(1):33–52
wrkyle F (2016) Matplotlib: How to make a histogram with bins of equal area? https://stackoverflow.com/questions/37649342/matplotlib-how-to-make-a-histogram-with-bins-of-equal-area, Accessed: 2024-05-24
Zhao Y, Nasrullah Z, Li Z (2019) Pyod: A python toolbox for scalable outlier detection. J Mach Learn Res 20(96):1–7. http://jmlr.org/papers/v20/19-011.html

Authors and Affiliations

Philipp Röchner[1] · Henrique O. Marques[2] · Ricardo J. G. B. Campello[2] · Arthur Zimek[2]

✉ Philipp Röchner
roechner@uni-mainz.de

Henrique O. Marques
oli@sdu.dk

Ricardo J. G. B. Campello
campello@imada.sdu.dk

Arthur Zimek
zimek@imada.sdu.dk

[1] Johannes Gutenberg University, Mainz, Germany

[2] University of Southern Denmark, Odense, Denmark