Journal of Statistical Modelling: Theory and Applications
Vol. 6, No. 1, 2025, pp. 105-114
Yazd University Press 2025

JSMTA

Research Paper

Blind label ratio estimation

Javad Kazemitabar*[1], Soheila Rahimi[2], Amir Rezaei-Ghadim[2]
[1]Department of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Tehran, Iran
[2]Son Corporate Group, Tehran, Iran

Received: February 06, 2025/ Revised: May 30, 2025/ Accepted: November 09, 2025

Abstract: Many anomaly detection algorithms require knowledge of the ratio of the two labels to operate. In real life, however, we may not have access to this value. As such, we often run anomaly detection packages with default values that may differ significantly from the actual value. Experiments on multiple datasets show that correctly determination of this ratio or at least obtaining a close estimate can makes a significant difference in the final performance of the anomaly detection algorithm. In this paper, we address the problem of estimating this ratio using both theoretical and heuristic techniques. In the theoretical method, we maximize the mutual information between features and labels to find the exact ratio. In the heuristic method, we sweep the [0,1] range in 0.01 steps to search for the ratio. On each iteration, we run the anomaly detection algorithm based on the ratio for that iteration and record the correlation coefficient between the features and the label generated by the algorithm. After the 100th iteration, we declare the ratio that provides the maximum correlation coefficient as our estimate of the label ratio. Our experiments on multiple datasets and several anomaly detection algorithms show that maximizing the correlation coefficient leads to the best results.

Keywords: Anomaly detection; Correlation coefficient; Mutual information; One-class support vector machine; Spectral ranking of anomalies.
Mathematics Subject Classification (2010): 68T01.

*Corresponding author: j.kazemitabar@nit.ac.ir

1 Introduction

Anomaly detection is an unsupervised learning technique for identifying data points that deviate significantly from the norm within a dataset. The objective of anomaly detection is to find outliers, unexpected changes, or errors that do not conform to established patterns or statistical models. This technique is valuable for uncovering potential issues, risks, or threats in various domains. While anomaly detection algorithms learn without the knowledge of the actual label, they do require some metadata about the labels such as the ratio of the two classes.

Many anomaly detection algorithms require the knowledge of the ratio of the two labels in order to operate. In real life, however, we may not have access to this value. As such, oftentimes we run anomaly detection packages using the default value, which may be quite distant from the actual value. Experiments on multiple datasets show that correct determination of this ratio -or at least a close estimate- makes a significant difference in the final performance of the anomaly detection algorithm.

However, to our surprise the amount of research on the actual methods that estimate the label ratio in a dataset is quite rare. As a matter of fact, in the literature, we could not find any work that directly proposes a method for estimating the ratio of the two classes. Thus, in here we will mention previous work that provide some insight towards our goal. In Quadrianto et al. (2009) the authors propose a method for learning from label proportions (LLP), a machine learning problem where the goal is to predict the labels of individual instances from the label proportions of groups of instances. The authors present consistent estimators that can recover the true labels with high probability in a uniform convergence sense, and prove their theoretical guarantees. They also present an algorithm based on convex optimization and kernel methods to implement their estimators. They conduct experiments on synthetic and real-world datasets, and show that their method outperforms existing methods in terms of accuracy and robustness. One should note however, that the method does not estimate label ratios; instead it takes label ratio as a known input for final classification. While the work is not directly related to the topic of this paper, it emphasizes the importance of our work. In Iyer et al. (2016) the authors introduce learning models for addressing the class ratio estimation problem, where the objective is to predict the proportions of instances in an unlabeled set belonging to different classes. Unlike existing models that rely on instance-level supervision, their approach directly estimates class ratios using set-level supervision. The authors propose a label privacy-preserving mechanism tailored for supervised class ratio estimation. The mechanism is designed to ensure efficient differential privacy, especially when dealing with sufficiently large counts per class. While the technique proposed in this work is related to the topic of our paper, it requires a labeled reference dataset in the beginning. In other words, they have a *supervised* method for learning class ratio estimation of sampled sets that are subset of the original reference dataset. In our paper, we are proposing a completely blind method where there is no information beforehand about the labels of the dataset. In Chen et al. (2022) a new distribution shift model, termed sparse joint shift (SJS), is proposed which considers the joint shift of both labels and a few features. This model unifies and generalizes existing shift models such as label shift and sparse covariate shift, where only marginal feature or label distribution shifts are considered. The authors show that SJS is identifiable under certain conditions, and propose shift

estimation and explanation under SJS (SEES), an algorithmic framework to estimate and explain the performance shift under SJS. SEES consists of three steps: (1) estimating the label proportions on the new data, (2) identifying the shifted features and estimating their conditional distributions, and (3) estimating the model performance and explaining the shift using Shapley values. The authors conduct extensive experiments on several real-world datasets with various maximum likelihood (ML) models, and demonstrate that SEES achieves significant improvement over existing methods in terms of shift estimation error and explanation quality. This work, too, is a supervised method and relies on a prior knowledge of label ratio in the reference dataset.

In Yang et al. (2019), the authors propose an outlier detection algorithm solely based on thresholding. They provide a two-state solution for outlier thresholding that is shown to perform better than the common confidence interval techniques, such as inter-quartile and median absolute deviation. While this work counts as a stand-alone outlier detection technique, it cannot be used for existing outlier detection algorithms. To be precise, the topic of our paper is a means to estimate the class ratios that pave the way for existing outlier detection algorithms such as one-class support vector machine (SVM) and local outlier factor (LOF).

Another attempt in class ratios is presented in Bootkrajang and Chaijaruwanich (2020) where label noise estimation is provided as a means for supervised learning algorithms. It goes without saying that this work too is not directly related to the problem we are trying to solve in our paper since it is not customized for unsupervised outlier detection algorithms.

In this paper, we tackle the problem of class ratio estimation to be used in an unsupervised learning problem with no prior knowledge. In so doing, we harness two powerful tools: mutual information and correlation, albeit by reverse engineering their regular functionality. In Section 2, we take a theoretical approach based on optimizing mutual information between features and labels. In Section 3, we explore a heuristic technique that are based on trial and error. In Section 4, we compare the results of these techniques. Finally, Section 5 concludes the paper.

2 Theoretical approach

Mutual information and correlation coefficient are two metrics used for feature selection in a supervised learning problem (Beraha et al., 2019; Hsu and Hsieh, 2010). We sort features based on their correlation or mutual information with the given label and choose those that stand higher for building a predictive model. The underlying assumption is that important features have a higher mutual information or correlation with the label. Now, imagine an unsupervised learning problem where we are not aware of the label. We still *know*, however, that the correct label -whatever that may be- has high mutual information (or correlation) with some of the existing features. Clearly, if we sum up the mutual information (or absolute value of correlations) of all the features with respect to the label, we will be capturing the most highly correlated features. We could use this knowledge to search for most likely statistics in the data. Consider the following optimization problem where we maximize mutual information between input variables, X, and the output label Y. We assume two labels, namely 0 and 1. We aim to find the best label ratio that maximizes the mutual information between the input

and output variables. Label ratio estimation is modeled here by finding $P\{Y = 1\}$ since $P\{Y = 0\} = 1 - P\{Y = 1\}$. In what follows we will be referring to $P\{Y = 1\}$ as $P(y_1)$. We will be considering $P(y_i)$ as our independent variables. We recall the definition of mutual information as first introduced by Shannon (1948)

$$I(X;Y) = -\sum_{y} p(y_j)\log(p(y_j)) + \sum_{y}\sum_{x} p(x)p(y|x)\log(p(y|x).$$

Mutual information is a concave function of $p(y)$ when $p(y|x)$ is considered a constant. This is easily verified by taking into consideration that $t\log(t)$ is a convex function of t. This paves the way to find the maximum of the mutual information in an optimization problem. We will form the optimization problem using an extra piece of information; that $P(y_1)$ is limited by an upper bound such as α_0. This could come from domain knowledge.

$$\begin{aligned} \min \quad & -I(X;Y), \\ & P(y_1) + P(y_0) = 1, \\ & -P(y_i) \leq 0, \\ & P(y_1) \leq \alpha_0. \end{aligned}$$

From the above, we can form the Lagrangian as follows

$$\mathcal{L} = -I(X;Y) + \lambda(P(y_1) - \alpha_0) + \mu(P(y_1) + P(y_0) - 1) - \eta P(y_1) - \nu P(y_0|x_j).$$

The Karush-Kuhn-Tucker (KKT) conditions (Boyd and Vanderberghe, 2004) will then be as follows:

$$\frac{\partial \mathcal{L}}{\partial P(y_1)} = \log(p(y_1) + 1 + \lambda + \mu - \eta = 0, \tag{1}$$

$$\frac{\partial \mathcal{L}}{\partial P(y_0)} = \log(p(y_0) + \mu - \nu = 0, \tag{2}$$

$$p(y_1) + p(y_0) - 1 = 0,$$

$$p(y_1) + p(y_0) - \alpha_0 = 0, \tag{3}$$

$$\eta p(y_1) = 0, \tag{4}$$

$$\nu p(y_0) = 0. \tag{5}$$

From (4) and (5), we conclude that $\eta = \nu = 0$ since no label is supposed to have a probability of zero. From (3), we conclude that either $\lambda = 0$ or $p(y_1) = \alpha_0$. If the first is true, i.e. $\lambda = 0$, then from (1) and (2), we conclude that $p(y_1) = p(y_0) = 0.5$. Otherwise, we are stuck with the upper bound, i.e. $p(y_1) = \alpha_0$. To summarize, the optimization has two solutions; in the first solution the two labels are equi-probable. In the second solution, $p(y_1) = \alpha_0$ which was dictated as the upper bound in the optimization problem conditions. From the above analysis, we conclude that we are highly dependent on domain knowledge in the problem of label ratio estimation. Precisely speaking, mutual information maximization suggests either equal probability for the two labels, or choosing the upper bound. In the next section we will pursue a heuristic method to further fine-tune the ratio estimate.

3 Heuristic method

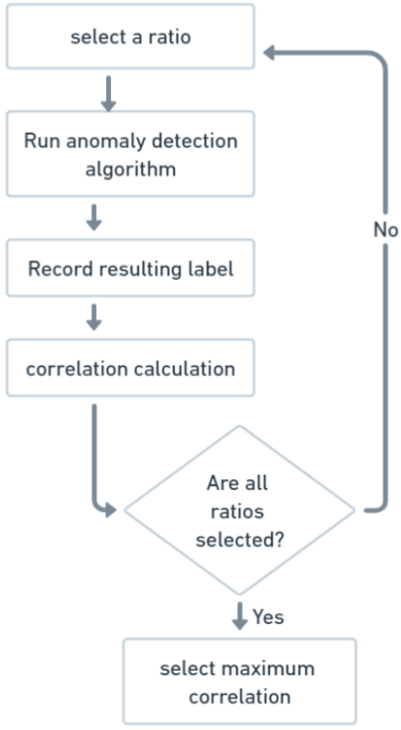

Figure 1: Block-diagram of the proposed heuristic method.

In this section, we use a heuristic method to estimate the most likely ratio of 0 and 1 labels by iterating over the range of [0,1]. At each iteration, we select a ratio, run the anomaly detection algorithm based on that ratio and then record the predicted label. Next, the predicted label is correlated with the features and the corresponding correlation coefficient is noted. Once all iterations are concluded, the ratio that leads to the highest correlation coefficient is declared as our estimate of the ratio. In other words, if the class ratio was chosen correctly, it would lead to a set of predicted labels that provide the highest correlation with features.

In finding the correlation coefficient, since we need to combine all the coefficients between each feature and the label, we took three different approaches. In our first approach we found the correlation coefficient of each feature with the label and then added their absolute value. In the second approach we found the correlation coefficient of each feature with label and summed up their squares. As the third approach we used the technique described in Abdi (2007), namely *multiple correlation*. The framework of multiple regression enables the examination of the connection between a dependent variable and a set of independent variables, assessing how effectively the independent variables can forecast the dependent variable. The multiple coefficient of correlation

signifies the proportion of the dependent variable's variance explained by the independent variables. A high multiple coefficient of correlation indicates accurate prediction and relevance of the independent variables, while a low value suggests poor prediction and lack of relevance. Multiple regression framework, which is used to predict a dependent variable Y from a set of independent variables X assumes that the data consists of N observations, each described by J independent variables and one dependent variable. The data is collected in a matrix X and a vector y, as shown below

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,j} & \cdots & x_{n,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,j} & \cdots & x_{N,J} \end{bmatrix}, \text{ and } \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \\ y_N \end{bmatrix}.$$

The predicted values of the dependent variable are collected in a vector b and are obtained as

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b} \text{ with } \boldsymbol{b} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

The quality of the prediction is evaluated by computing the multiple coefficient of correlation $R^2_{Y.1,\ldots,J}$ which is equal to the squared correlation between the dependent variable (Y) and the predicted dependent variable (b). The multiple coefficient of correlation can be computed as

$$R^2_{Y.1,\ldots,J} = \frac{SS_{\text{regression}}}{SS_{\text{regression}} + SS_{\text{error}}} = \frac{SS_{\text{regression}}}{SS_{\text{total}}},$$

where $SS_{\text{regression}}$ is the regression sum of squares, SS_{total} is the total sum of squares, and SS_{error} is the residual (or error) sum of squares. These quantities are defined as

$$\begin{aligned} SS_{\text{regression}} &= \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{y} - \frac{1}{N}\left(\boldsymbol{1}^\top \boldsymbol{y}\right)^2, \\ SS_{\text{total}} &= \boldsymbol{y}^\top \boldsymbol{y} - \frac{1}{N}\left(\boldsymbol{1}^\top \boldsymbol{y}\right)^2, \\ SS_{\text{error}} &= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{y}. \end{aligned}$$

In the heuristic method, we will use $R_{Y.1,\ldots,J}$. This measure provides a means to sum up the corelation between a label and multiple features. We rely on the heuristic concept that features have high corelation with the true label. Conversely, if class ratios are chosen correctly, the anomaly detection algorithm will generate the most accurate labels that would then lead to highest corelation with the features.

4 Results

We used two publicly available datasets. The first dataset used for car insurance fraud detection includes 15420 samples and is made available by Automobile Insurance Bureau (AIB, Brockett et al., 2002). The dataset contains 31 features describing:
• driver details (gender, age, marital status, ratings, etc.),

• vehicle details (make, model, year, insurance type),
• accident details (day of week, date, location).

As a pre-processing step, We modified the features into numeric and categorical types to be fed into the outlier detection algorithms.

The second dataset used, also available to public, is for health insurance fraud detection and includes 5410 samples (Kaggle, 2019). The original dataset was in the form of multiple spreadsheets. We aggregated them all into a single spreadsheet with 38 features. These features provide:
• Claim details (Total claim amount, number of claims, Annual reimbursement details, etc.),
• Treatment details (Physician details, Chronic condition, etc.),
• Patient details (Age, race, etc.).

Our goal was to provide an estimate for the unsupervised learning algorithm to distinguish legitimate from fraudulent claims in each dataset. We tested four different unsupervised algorithms for each dataset:
1. One Class Support Vector Machine (OC-SVM),
2. LOF,
3. Isolation Forest (IF),
4. Spectral Ranking of Anomalies (Nian et al., 2020; Shaeiri and Kazemitabar, 2020).

We also used a supervised learning algorithm, namely random forest (RF), as an upper bound for the unsupervised algorithms. Table 1 shows the estimates found by the correlation technique. In order to evaluate the performance of our technique, we decided to measure the output of the anomaly detection algorithm rather than the output of the ratio estimation block. The percision-recall (PR) and receiver operating characteristic (ROC) curves for the above mentioned algorithms applied to car insurance dataset are shown in Figures 2 and 3 respectively. Figures 4 and 5 apply the same procedure on health insurance dataset. As can be seen in Figure 5, the performance of the SRA algorithm with the provided estimate, is close to the upper bound provided by the supervised algorithm (RF).

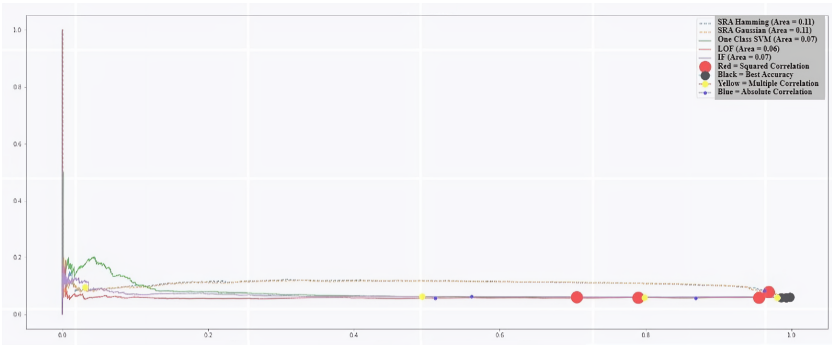

Figure 2: Precision-Recall curve for correlation technique applied to car insurance dataset.

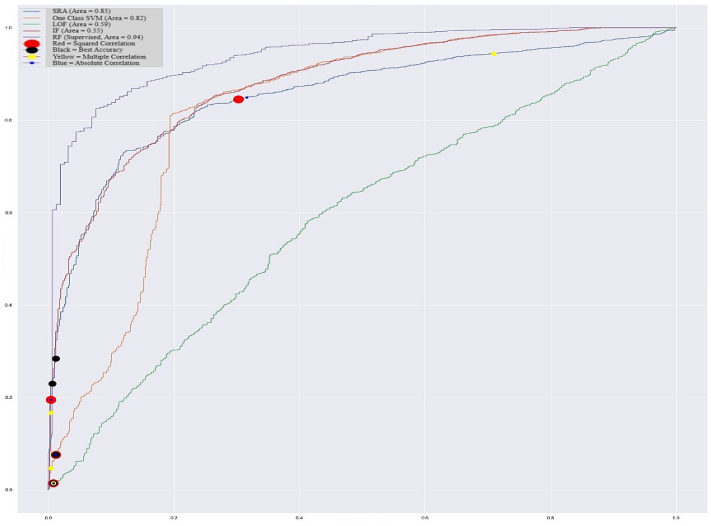

Figure 3: ROC curve for correlation technique applied to car insurance dataset.

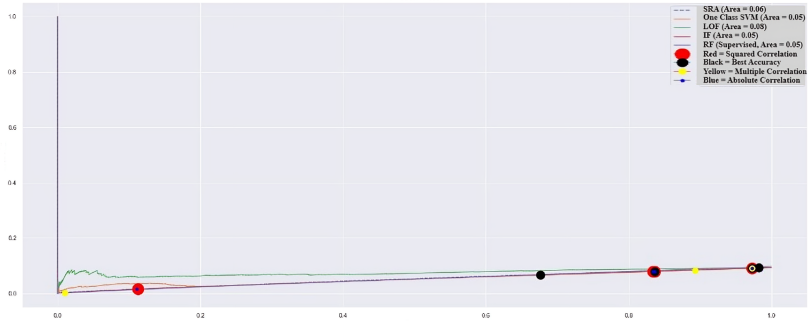

Figure 4: Precision-Recall curve for correlation technique applied to health insurance dataset.

Table 1: Estimated label ratios using different correlation techniques: Multiple correlation, sum of absolute value of single correlations and sum of square of single correlations.

Dataset	Automobile				Healthcare			
Method	Mult. Corr.	Abs. Corr.	Squared Corr.	Actual	Mult. Corr.	Abs. Corr.	Squared Corr.	Actual
OC-SVM	2%	9%	4%	6%	8%	2%	2%	9%
LOF	19%	46%	19%	6%	73%	1%	1%	9%
IF	56%	33%	19%	6%	6%	2%	3%	9%
SRA	33%	29%	25%	6%	7%	2%	3%	9%

5 Conclusion

Many anomaly detection algorithms require the knowledge of the ratio of the two labels in order to operate. In real life, however, we may not have access to this value. As

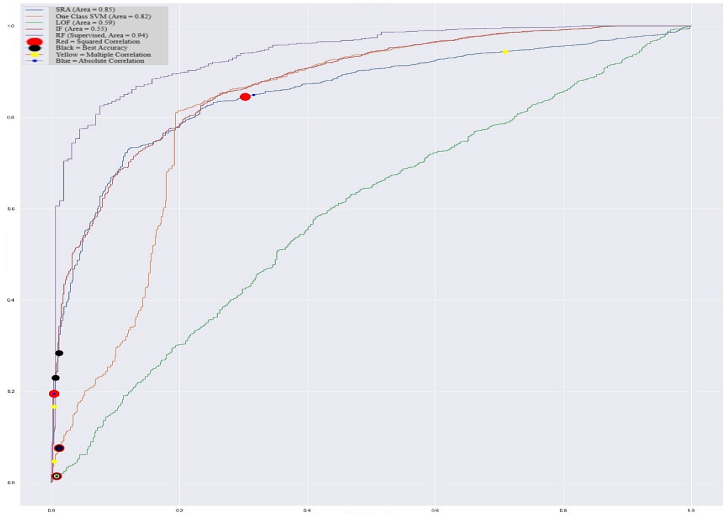

Figure 5: ROC curve for correlation technique applied to health insurance dataset.

such, oftentimes we run anomaly detection packages using the default value that may be quite distant from the actual value. Experiments on multiple datasets show that the correct determination of this ratio-or at least a close estimate-makes a significant difference in the final performance of the anomaly detection algorithm. In this paper we investigated a heuristic method that provided an acceptable estimate for several unsupervised algorithms. The results show that the area under curve in the ROC plots approach the upper bound provided by the supervised algorithm.

Acknowledgement

This work was supported by research project No. P/M/1126 between Son Corporate Group and Babol Noshirvani University of Technology.

References

Abdi, H. (2007). Multiple correlation coefficient. *Encyclopedia of Measurement and Statistics*, **648**(651):19.

Beraha, M., Metelli, A.M., Papini, M., Tirinzoni, A. and Restelli, M. (2019). Feature selection via mutual information: New theoretical insights. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-9. IEEE.

Bootkrajang, J. and Chaijaruwanich, J. (2020). Towards an improved label noise proportion estimation in small data: A Bayesian approach. *International Journal of Machine Learning and Cybernetics*, **13**(4):851–867.

Boyd, S. and Vanderberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Brockett, P.L., Derrig, R.A., Golden, L.L., Levine, A. and Alpert, M. (2002). Fraud classification using principal component analysis of RIDITs. *Journal of Risk and Insurance*, **69**(3):341–371.

Chen, L., Zaharia, M. and Zou, J. (2022). Estimating and explaining model performance when both covariates and labels shift. *Advances in Neural Information Processing Systems*, **35**:11467–11479.

Hsu, H.-H. and Hsieh, C.-W. (2010). Feature selection via correlation coefficient clustering. *Journal of Software*, **5**(12):1371–1377.

Iyer, A.S., Nath, J.S. and Sarawagi, S. (2016). Privacy-preserving class ratio estimation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 925–934.

Kaggle. (n.d.). Medical provider fraud detection. Available at https://www.kaggle.com/code/rohitrox/medical-provider-fraud-detection/input

Nian, K., Zhang, H., Tayal, A., Coleman, T. and Li, Y. (2016). Unsupervised spectral ranking for anomaly and application to auto insurance fraud detection. *Journal of Finance and Data Science*, **2**(1):1–28.

Quadrianto, N., Smola, A.J., Caetano, T.S. and Le, Q.V. (2009). Estimating labels from label proportions. *Journal of Machine Learning Research*, **10**:2349–2374.

Shaeiri, Z. and Kazemitabar, S.J. (2020). Fast unsupervised automobile insurance fraud detection based on spectral ranking of anomalies. *International Journal of Engineering*, **33**(7):1240–1248.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**(3):379–423.

Yang, J., Rahardja, S. and Fränti, P. (2019). Outlier detection: How to threshold outlier scores? In *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, 1–6.