Unsupervised anomaly detection using Ensembles

NIKITHA RAO, Technical University Dortmund, Germany

We have large amounts of data in all areas, and exploring anomalies in these datasets can help us uncover a few insights that could have a significant impact. Ensembles have robust performance and give accurate results. We can therefore use it in the detection of anomalies. But in an unsupervised data environment, we will be unaware of the ground truth, and hence the use of item response theory (IRT), a class of models used in educational psychometrics to assess characteristics of students and test questions can be helpful. The hidden trait calculation of the IRT helps in anomaly detection because this hidden trait is useful in discovering the hidden class labels. With the use of a new IRT mapping for the anomaly detection problem, the paper builds an ensemble that provides better performance compared to other ensemble techniques. This report summarizes the concepts, experiments, and conclusions of Dr. Sevvandi Kandanaarachchi's publication "Unsupervised Anomaly Detection Ensembles using Item Response Theory" [11] and further critiques this publication and the analysis. We will also see the limitations of the Ensembles, IRT, and other factors explored in the form.

ACM Reference Format:
Nikitha Rao. 2022. Unsupervised anomaly detection using Ensembles. 1, 1 (February 2022), 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In data analysis, the detection of "abnormal" instances has many applications. This process of abnormal detection is commonly referred to as anomaly detection. Unsupervised Anomaly Detection (AD) is used in many applications such as fraud detection. In fraud detection, many data logs are scanned for suspicious anomalies that indicate fraud. Specifically, we can detect fraudulent activity on bank accounts by analyzing financial transactions, and misused or stolen cards can be detected using credit card payment records. There are different techniques for performing anomaly detection.

Due to improved results and robust performance of ensembles,[8] they can be used as one of the methods to detect anomalies. Ensembles combine several base models to produce one optimal model. Common examples of ensembles are Decision trees, Random forests, etc. The way the ensemble works for anomaly detection is that every base model gives an "anomaly score" for each observation in the dataset. The larger the scores, the more anomalous is the observation. Later an ensemble score is obtained as a combination of normalized anomaly scores from several methods. This ensemble score provides a final verdict on the anomalousness of the data. Here we can check if the data is correctly classified or not based on the ground labels. But our setting is an unsupervised anomaly detection and we do not have the class labels or ground truth. So that imposes a challenge. Because of this, the proven and tested ensemble frameworks that use the class labels cannot be applied for anomaly detection. So the paper proposes a new ensemble framework with better performance that is based on Item Response Theory (IRT), which models the ground truth as a hidden trait.

Author's address: Nikitha Rao, nikitha.rao@tu-dortmund.de, Technical University Dortmund, Germany.

Manuscript submitted to ACM

We now know that we must arrive at an ensemble score while we are using ensembles to find the anomalous data, and this can be done in three ways [3]:

- Subsampling: We take several subsamples of a data set and fit the same algorithm on it. Then we take the average of the scores of all the algorithms to arrive at an ensemble score.
- Feature bagging: We take several subsets of features of a data set and fit the same algorithm on it. Then we take the average of the scores of all the algorithms to arrive at an ensemble score.
- Combination function: We apply different algorithms on the same data sets and then apply another algorithm to combine the results from all the models.

So subsampling and feature bagging uses the same algorithm. The combination function uses different algorithms and then applies an algorithm on all the algorithms to arrive at an ensemble score. So IRT is one of the combination functions. Fig.1 is a diagram representing how a combination works for better understanding. We use 7 different anomaly detection methods like LOF, KNN, etc. to get the anomaly scores. We then use 7 different combination functions like average, greedy, IRT, and a few others. The performance of IRT is then compared with the other 6 combination functions.

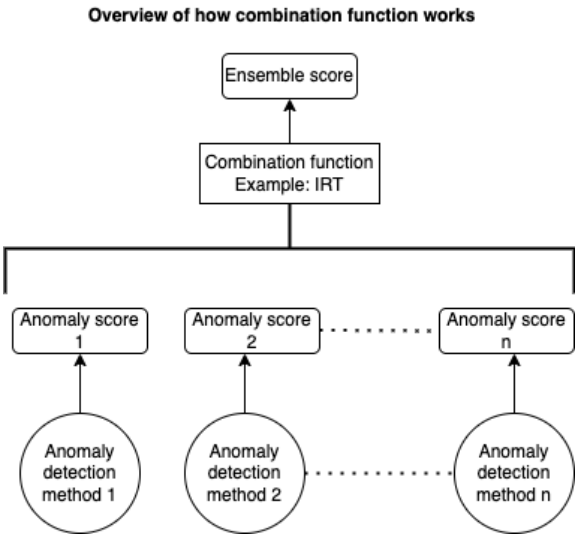

Fig. 1

The proposed IRT ensemble framework is a new combination function that can be used on a heterogeneous AD method. IRT models the unknown ground truth as a hidden/latent variable. Latent variables mean that they are not observed directly, but are inferred from a mathematical model. So they are hidden. For example, student intelligence can be modeled as a latent variable using student answers to a test. Similarly, the unknown label of anomalies can be modeled as a hidden trait using the scores we get from the different AD methods. We also use different combination functions and get the results to compare their performance with the IRT combination function. [11]

The IRT had already been studied and used. One of the examples was the IRT ensemble proposed by Chen and Ahn (2020). [4] But they used class labels to evaluate the model, so it couldn't be used directly in unsupervised environments. Therefore, the aforementioned paper proposed an IRT ensemble for unsupervised anomaly detection that used the latent

trait to discover the ground truth. In the next Section 2, we will get a better idea about IRT. We will also understand how IRT is mapped to machine learning problems in the paper. Then, we will compare different combination functions against IRT using the experiment conducted on huge publicly available anomaly detection datasets in Section 3. In Section 4, we will criticize some of the techniques used in the paper. Finally, we present our conclusions in Section 5.

2 ELABORATING THE PROBLEM STATEMENT AND IRT METHOD

2.1 Problem statement

Assume that we have X dataset with n observations. Assume that t denotes the ground truth vector of these observations. Ground truth is a real number. Now assume we have ℓ different anomaly methods that are giving 1 anomaly score each. So $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_\ell$ is the anomaly scores of ℓ different AD methods on X. Let $f_1, f_2, \ldots, f_k$ denote k ensemble combination functions with $\boldsymbol{v}_1 = f_1\left(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell\right), \ldots, \boldsymbol{v}_k = f_k\left(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell\right)$. Once these combination functions are applied on anomaly scores, we get k different anomaly scores for k combination functions. These ensemble scores are denoted as $\boldsymbol{v}_j \in \mathbb{R}^n$ for $j \in \{1, \ldots, k\}$. Now we want to check which combination function is better. So we take a performance parameter and check the better performing combination function. Let ξ denote an AD performance metric such that $\xi\left(t, v_j\right) \in \mathbb{R}$. If a combination function f_b with ensemble output $\boldsymbol{v}_b = f_b\left(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell\right)$ satisfies

$$\xi\left(\boldsymbol{t}, \boldsymbol{v}_b\right) > \xi\left(\boldsymbol{t}, \boldsymbol{v}_j\right)$$

for all $j \in \{1, \ldots, k\}$, $j \neq b$, this indicates that f_b is performing better than all other combination functions as it has the highest performance concerning the ground truth. [11]

2.2 Relating the problem statement to our methods

The Anomaly detection methods that give anomaly scores that are represented by $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_\ell$ in the problem statement are 7 different methods like density-based outlier LOF, distance-based outlier KNN and so on. The AD methods used are not the focus of the paper. The combination functions which are used on anomaly scores to get the ensemble scores, denoted by $\boldsymbol{v}_1 = f_1\left(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell\right), \ldots, \boldsymbol{v}_k = f_k$ are 7 different methods - IRT, Average, Greedy, Averaged greedy, Inverse Cluster Weighted Averaging (ICWA), Maximum Scoring (Max) and Threshold Sum (Thresh). Just to get an idea, the average function computes the average anomaly score to give an ensemble score. The advantage of the average combination function is that it is a benchmark, standard algorithm and it performs well on homogeneous data distribution. But the con is that the performance is not good for heterogeneous data. [1] The only combination function in focus is IRT. So this report will not explain the details of other combination functions. All these are existing combination functions. Further, IRT is explained in the next section. [11]

2.3 Item Response Theory explained

Item response theory models the ground truth as a hidden trait. It establishes the connection between hidden characteristics and their corresponding outcome or response. To elaborate, IRT establishes a link between 3 components

- Properties of an item
- The response we get for those items
- Underlying outcome or trait that is being observed

For better understanding, we will take the previous example of the student examination. Here the properties of an item are the test questions. The response we get for those items is the corresponding answers or score. The underlying

outcome or trait that is being observed is the intelligence of the student. There are 3 types of IRT based on the responses used in fitting The types are dichotomous IRT if the responses are binary. Polytomous IRT if the responses are discrete and continuous-valued IRT if the responses are continuous. We will look closely into Dichotomous IRT.

- Dichotomous IRT: Responses are binary
- Polytomous IRT: Responses are discrete
- Continuous-valued IRT: Responses are Continuouse values

We we will look into Dichotomous IRT to get a better understanding of IRT. Fig. 2 is the function for IRT. Assume that

Fig. 2. The diagram represents the 2PL model for IRT

there are N participants, that is $i = 1, 2, \ldots, N$ are participants and assume that there are n test items. So $j = 1, 2, \ldots, n$ are test items. Consider $y_{ij} \in \{0, 1\}$ denote the score or response of the i^{th} participant to the j^{th} test item. The discrimination parameter for test item j is denoted by α_j and the difficulty parameter by β_j. These give the probability of a correct response for each item given the ability level θ_i as

$$\Phi\left(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j\right) = \frac{1}{1 + \exp\left(-\alpha_j\left(\theta_i - \beta_j\right)\right)}.$$

Here the parameters are discrimination, difficulty, and ability of the student. The discrimination parameter for test item j is denoted by α_j and the difficulty parameter by β_j. β_j is the location parameter of the logistic function and can be seen as a measure of item difficulty. α_j indicates the steepness of the function at the location point. θ_i is the ability of the participant i. [11], [7], [5]

Interpreting the formula in terms of real-life example, considering the student-examination is again, We can see that when the difficulty parameter of the student is β_j and the ability is $theta_i$, the whole equation reduces to 0.5 indicating that there is a 50% possibility of a student answering the question correctly.

2.4 Mapping Item Response Theory to ensembles

2.4.1 ***In Psychometrics:*** [11] There are two possibilities of mapping the response in psychometrics. One is where y_{ij} is the score of the ith participant to the jth test item. Fig. 3 represents that mapping. Here the forward arrow indicates each participant doing each test item. From this IRT mapping, we obtain item discrimination and difficulty for test items, and participant ability for participants. The next is is where original responses are used in studies investigating

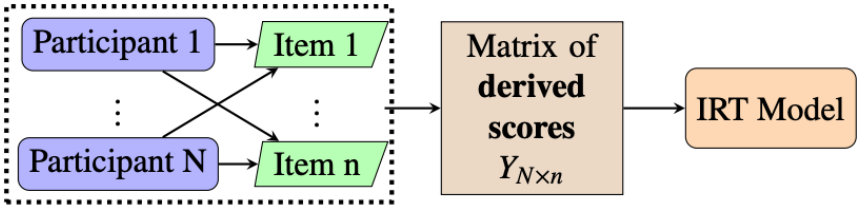

Fig. 3. The matrix of derived test scores/marks are used to fit the IRT model.

underlying characteristics instead of scores. For example these can be in scenarios where users are asked behavioral questions and there are no right answers to get scores. So here we do not derive score, but use the original response itself. We can see this in Fig. 4

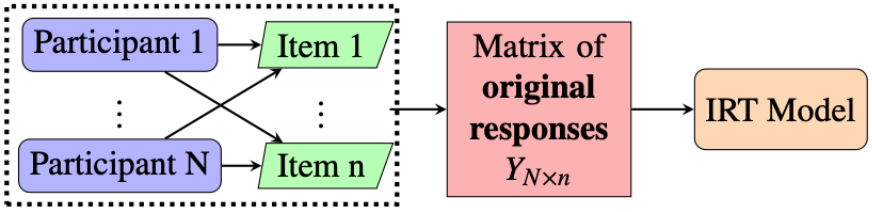

Fig. 4. The matrix of original responses are used to fit the IRT model.

2.4.2 ***In Algorithm evaluation:*** [11] Next is mapping in algorithm evaluation and even this is of two types. The standard approach is to map the participants to algorithms, and test items to dataset observations. y_{ij} denotes an accuracy measure such as classification accuracy of the ith algorithm on the jth observation. This can be seen in Fig.5 The second is an inverted mapping that uses datasets instead of observations to evaluate algorithm performance. In this mapping participants were mapped to datasets and test items were mapped to algorithms as shown in Fig. 6. 7.

2.4.3 . In the current model implemented: [11] Current research is the combination of psychometrics and the evaluation model. As the data set is unsupervised, it is obvious that we cannot use accuracy measures as per the mapping in 2.4.2. Thus the implementation uses standardized original responses instead of an accuracy measure in y_{ij}. Then the participants are mapped to dataset observations and test items to algorithms so that we obtain item discrimination and difficulty for test items, and participant abilities. This can be seen in Fig.7.

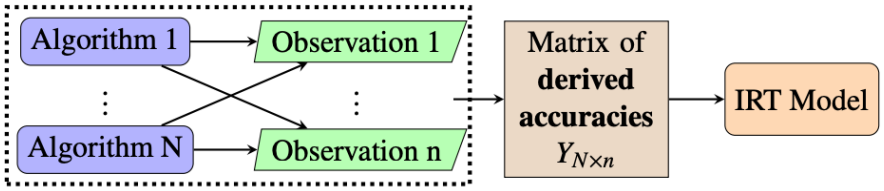

Fig. 5. Algorithms acting on data set observations in the standard algorithm evaluation setting

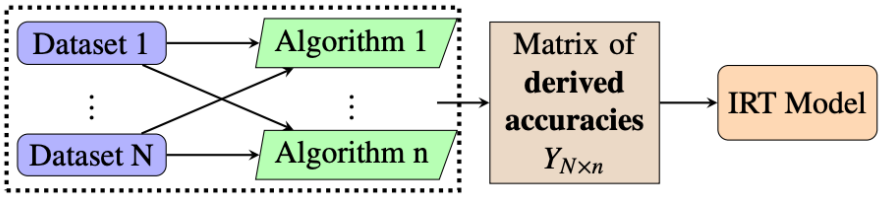

Fig. 6. Datasets acting on algorithms in the algorithm evaluation setting

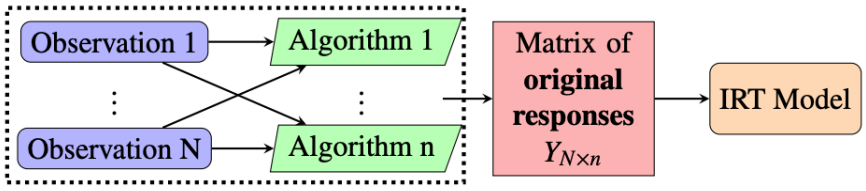

Fig. 7. Observations act on algorithms and y_{ij} is the original responses. The proposed IRT mapping for AD ensemble learning is visualised.

2.5 Interpretation of parameters after mapping

Now that we have mapped IRT to ensembles, we will check what would each parameter represent.

2.5.1 ***Ability parameter*** θ_i***:*** [11] As seen in Fig. 7, the mapping uses original responses in the IRT model. Doing this results in a latent trait continuum holding a hidden quality. For example, the latent trait continuum can be confidence in a behavioral questionnaire. After that, the participants are ordered in the latent trait continuum by this measured quality. We can also see in Fig. 7 that the participants are mapped to observations in the ensemble setting, so the parameter θ denotes a certain quality about the observations. The mapping also uses the standardized algorithm response instead of an accuracy measure, so the new trait parameter is proportional to the algorithm response. In our case, algorithm response is the anomaly score assigned to each observation. We can observe that θ increases with the anomaly score. Thus, the parameter θ_i indicates how anomalous an observation i is according to all the AD methods. Thus we can conclude that θ_i denotes ensemble score.

2.5.2 ***Difficulty parameter*** β_j***:*** [11] In the context of our mapping, β_j is the anomalousness threshold of algorithm j based on which it decides if a data point is anomalous or not. In general, if we want to avoid false positives i.e. normal data being identified as an anomaly, we need to set a higher β_j.

2.5.3 ***Discrimination parameter*** α_j***:*** [11] In the context of our mapping, α_j denotes the degree to which the algorithm can distinguish non-anomalous points from anomalous points. When we set a lower discrimination parameter, the algorithm gives similar scores for anomalies and non-anomalies. So we need a higher discrimination parameter.

3 EXPERIMENTAL RESULT

Now that we have understood IRT and how it is mapped to ensembles, we will look into a real-world experiment conducted in the paper to understand the performance of IRT against other combination functions. But we will not go into every detail of the experiment conducted as it is out of the scope of what we are trying to understand here. [11]

Consider 12433 publicly available anomaly detection datasets [12], which have been prepared from 119 source datasets. These data sets have been prepared by downsampling different classes. The anomaly in these data sets is the minority class of each downsampled dataset. Now we apply 7 different anomaly detection methods like LOF, KNN, and more. Then we get the anomaly scores. After that, we apply 7 different combination functions - - IRT, Average, Greedy, Averaged greedy, Inverse Cluster Weighted Averaging (ICWA), Maximum Scoring (Max), and Threshold Sum (Thresh) on these anomaly scores to obtain 7 ensemble scores. Lastly, the performance of these are measured by AUROC, which is a plot between "True positive rate" versus "False positive rate" providing an aggregate measure of performance.[9]

Each experiment is conducted twice with different parameter settings. k indicates the number of cluster points and below is the two parameter settings used:

- T1 uses the default values for $k = k\_min = 5$ and $k\_max = 10$, and
- T2 uses $k = k\_min = \max(N/10, 50)$ and $k\_max = k + 10$. N is the number of observations in the dataset.

We then plot the performance of each combination function for both parameter settings as per dataset (shown in Fig. 8) and the data source (shown in Fig. 9). We can see that IRT has the highest performance for most of the datasets and data sources in both parameter setting $T1$ and $T2$.

4 CRITICISM OF THE PAPER

This section highlights some criticism and improvements on our reference paper. [11]

4.0.1 ***The base models that give the anomaly scores are not explored:*** The foundation of ensembles is that it works by reducing variance between good base models to improve performance. But in anomaly detection using ensembles, the basic problem is getting these good models. The base models must be accurate and diverse to get a good ensemble score in the end. But the base classifiers were completely neglected by the paper. They just selected 7 base models without explaining why these models are optimal to calculate anomaly scores. Also, they did not explore if any other combination function would give better results for different base models. [10]

4.0.2 ***Ensembles create a black box:*** Ensembles use different base models to achieve different anomaly scores, and the combination of these scores then results in an ensemble score. So the downside is that the ensembles are more difficult to understand and it's challenging to understand the complex reasons why they choose a particular output. It is difficult to trace back the steps and calculate them as it is with single learners. We also have no explanation for the decisions made as it is with prototypes.[13] [2]

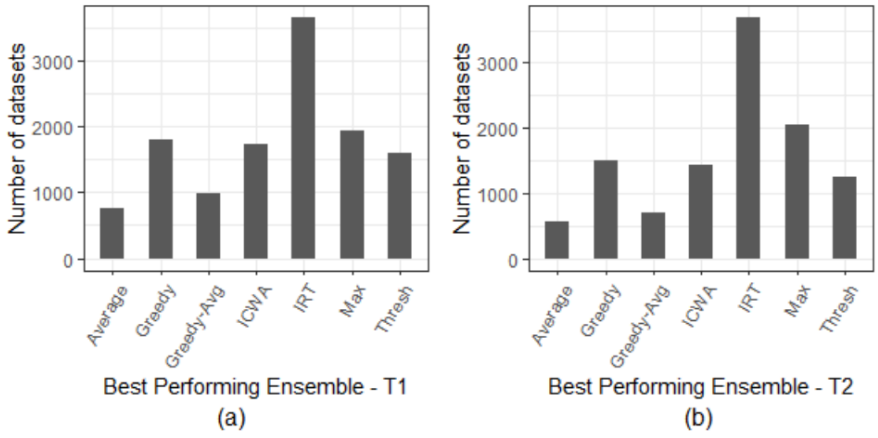

Fig. 8. Best performing ensemble for T1 and T2 as per datasets

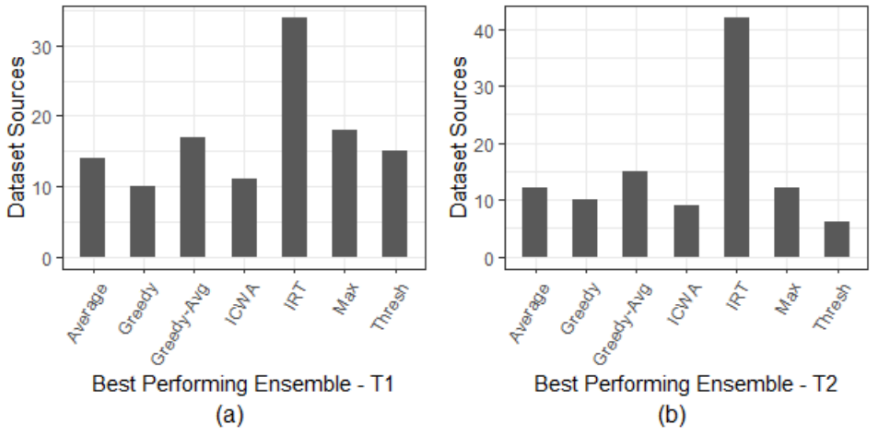

Fig. 9. Best performing ensemble for T1 and T2 as per data sources

4.0.3 ***AUROC is the only performance measure used:*** Research on outlier detection is currently focused on the development of new methods. The assessment, however, is quite repetitive, using only the area under the receiver operating characteristic curve. AUROC method calculates the true positive rate and false-positive rate, which is a higher-level representation of scores. But the method loses scoring information at a granular level, that is the degree to which each combination function provides a contrast between inliers and outliers. It completely neglects ensemble scores of each combination function and does not allow easy and clear comparison between these different methods or an overview of the similarity between their results. [6]

4.0.4 ***Computationally expensive:*** IRT techniques or ensembles, in general, may be computationally expensive. This is because we will have to fit IRT models for tens of millions of instances, which may be very common when

we have large datasets in almost all applications. The ensembles need to execute more than one learner, and thus the computational speed needed may significantly increase. [7]

5 CONCLUSION AND FUTURE WORK

In conclusion, anomaly detection is becoming popular with many applications. Robust Ensembles can be a good choice for anomaly detection. But in a real-world scenario, we might not have the ground truth of the labels. But ensembles might need ground truth to evaluate the performance. Thus constructing ensembles using heterogeneous AD methods was a challenge. So the paper introduced IRT which uses a latent trait to compute the ground truth. This was the first research that applied IRT to unsupervised ensemble anomaly detection. Evaluating the performance of IRT ensembles compared to 6 other ensemble techniques showed that IRT performs the best. But ensembles come with certain limitations like unexplainable decisions and expensive computation. We can also explore a better performance measure than ROC and explore different base models to see if it affects the conclusions of current experiments.

REFERENCES

[1] Chiang Alvin and Yeh Yi-Ren. 2015. Anomaly Detection Ensembles: In Defense of the Average. (2015), 207-210 pages. https://doi.org/10.1109/WI-IAT.2015.260

[2] Zimek Arthur, Campello Ricardo, and Sander Joerg. 2014. Ensembles for unsupervised outlier detection: challenges and research questions. A position paper. *ACM SIGKDD Explorations Newsletter* 15 (2014), 11-22 pages. https://doi.org/10.1145/2594473.2594476

[3] Jason Brownlee. 2021. *Understanding AUC - ROC Curve.* Retrieved February 2, 2022 from https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/

[4] Ziheng Chen and Hongshik Ahn. 2020. Item Response Theory Based Ensemble in Machine Learning. *International Journal of Automation and Computing* 17 (2020), 621 pages. https://doi.org/10.1007/s11633-020-1239-y

[5] S.E. Embretson and S.P. Reise. 2013. *Item Response Theory.* Taylor & Francis, New York, NY.

[6] Schubert Erich, Wojdanowski Remigius, Zimek Arthur, and Kriegel Hans-Peter. 2020. On Evaluation of Outlier Rankings and Outlier Scores. *Proceedings of the 12th SIAM International Conference on Data Mining, SDM 2012* (2020), 1047-1058 pages. https://doi.org/10.1137/1.9781611972825.90

[7] Plumed Fernando, Prudêncio Ricardo, Martínez-Usó Adolfo, and Hernandez-Orallo Jose. 2016. Making Sense of Item Response Theory in Machine Learning. (2016). https://doi.org/10.3233/978-1-61499-672-9-1140

[8] Valentini Giorgiovand Masulli Francesco. 2002. *Item Response Theory.* Springer Berlin Heidelberg, Berlin, Heidelberg.

[9] Sarang Narkhede. 2018. *A Gentle Introduction to Ensemble Learning Algorithms.* Retrieved February 2, 2022 from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[10] Melville Prem and Mooney Raymond. 2003. In Constructing Diverse Classifier Ensembles using Artificial Training Examples. *Proceedings of the 18th International Joint Conference on Artificial Intelligence* (2003). https://doi.org/10.5555/1630659.1630734

[11] Kandanaarachchi Sevvandi. 2021. Unsupervised anomaly detection ensembles using Item Response Theory. *RMIT University working paper* (2021). https://doi.org/10.13140/RG.2.2.18355.96801

[12] Kandanaarachchi Sevvandi and Smith-Miles Kate. 2020. Comprehensive Algorithm Portfolio Evaluation using Item Response Theory. (2020). https://doi.org/10.13140/RG.2.2.11363.09760

[13] Ming Yao, Xu Panpan, Qu Huamin, and Ren Liu. 2019. Interpretable and Steerable Sequence Learning via Prototypes. (2019). https://doi.org/10.1145/3292500.3330908