Bachelor Thesis

Interpreting Anomaly Detection: Optimized Preprocessing for Ensemble Methods

Hsin Ping Tang

October 23, 2022

Supervisor:

Prof. Dr. Emmanuel Müller

MSc. Simon Klüttermann

Technische Universität Dortmund

FAKULTÄT FÜR INFORMATIK

Lehrstuhl IX - Data Science and Data Engineering

https://ls9-www.cs.tu-dortmund.de/

Abstract

Compared to other data mining problems, the study of the ensemble method for unsupervised anomaly detection is hardly touched on. With the paper of DEAN, we concluded the advantages of combining weak but deep anomaly detectors into an ensemble and its potential in terms of interpretability. The interpretability of a model plays an essential role in understanding how a model works and its prediction. Therefore, we exploited the feature bagging method in our ensemble model for better interpretability. Our work in this thesis is to enhance the interpretation ability of our algorithm by optimizing the feature bagging process and analyzing the results from different data sets. To better evaluate our model, we defined our evaluation system based on the kNN algorithm for quantifying interpretability improvement. The results exhibit the underlying difficulties and possibilities of further research.

Contents

1 Introduction

1.1 Motivation and Background

In the era of Big Data, extensive amounts of data are being produced and collected daily. This gives us an unprecedented opportunity to explore large databases and obtain valuable information. During the data analysis process, what stands out and is viewed as interesting are often those unexpected and rare data points, so-called anomalies. Anomalies are referred to as outliers, abnormalities, or deviants in the data mining and statistics literature. They can be simply regarded as noise when lacking interesting characteristics or significant boundaries from regular data points [2], and therefore be removed during the data cleaning process. On the other hand, detecting and observing strong anomalies which are noticeably different from the norm of a data set reveals useful information. This can be valuable depending on the domain setting. In combination with machine learning techniques, it has gained attention in various application domains in recent years, such as cyber security, financial fraud, medical diagnosis, and law enforcement. For instance, fraud detection inspects abnormal usage patterns on credit card transaction data sets to identify fraudulent activity. In the medical field, the pixels in mammogram images are classified as cancerous or not for detecting calcification within the breast tissue [13].

Depending on the application-specific scenario, different types of machine-learning methods come into play. Neural networks(also called artificial neural

networks, ANN) are a set of algorithms used in machine learning inspired by the function of neurons in the human brain. A deep neural network essentially consists of multiple processing layers and is designed for analyzing more complex data in greater depth. The larger the amount of data given to training a deep learning model, the more experienced it becomes. Studies have shown that deep learning-based algorithm outperforms conventional machine learning method [9]. A common data mining problem like classification usually employs supervised machine learning where training data are labeled. This is often unrealistic for anomaly detection tasks because the manual efforts required to prepare massive training data are enormous, time-consuming, and sometimes impossible. Instead, unsupervised machine learning is a preferred technique for detecting anomalies. Without the pre-made labels, the model can pick up noticeable representations by itself and derive novel insights for which they have not been trained.

As anomaly detection starts to be involved in multiple decision-making processes, the reliability of its prediction becomes increasingly essential. Explaining what a model does and understanding how its prediction is made helps users gain trust and confidence in the model [21]. However, despite the effort put into research, limited progress has been made in the field of (unsupervised deep learning model) Interpretability [8]. The rise of the complexity of a method comes with the loss of its interpretability. Neural networks, for example, are considered intrinsically uninterpretable because of their non-linear and multilayered structure [8]. Therefore, we rely on a post hoc, model-specific explanation method to generate an explanation of the prediction. One approach of such is saliency maps, also known as pixel attribution [18]. It is mainly used for interpreting the classification of images with feature importance. Each pixel is considered a feature, and neural networks will highlight relevant pixels that contribute to the prediction.

Since anomalies tend to be unpredictable and rare, one of the biggest chal-

lenges of anomaly detection is to clearly define a boundary between abnormal and normal events [9]. One of the effective approaches to tackling this problem is by using multiple training models, which are employed for detecting different types of anomalies. This technique of combing weak learners into strong ensembles has not been thoroughly researched for anomaly detection, even though it has been proven to improve performance and robustness on classification [25]. In this thesis, we carry on the work of DEAN, our deep ensemble anomaly detection algorithm. And implemented different extensions for its feature bagging method to improve the interpretability of the model. With DEAN, we concluded the advantages of combining deep anomaly detection models into an ensemble and its potential in interpretability [14].

1.2 Related Work

As mentioned before, anomaly detection has been an emerging topic among many domains. There are studies across different application areas, such as cyber intrusion detection in [15], medical diagnosis [22], or fraud detection in finance [6]. It includes a wide variety and size of data sets, repeatedly showing the versatility and capability of anomaly detection. As deep learning approaches have been shown to improve performance further, it has also raised the interest in researching different methods for anomaly detection, including studies of the nearest-neighbor method [12], deep one-class classification [20], or methods based on training generative adversarial networks(GAN) [10]. They all present fairly well-performance models compared to traditional anomaly detection methods. However, without the combination with the ensemble method, their robustness and quality are still not as ideal as DEAN [14].

The scarcity of researching ensemble methods combined with deep learning models is still waiting to be filled. The ensemble method has been widely studied in other problem domains like classification and clustering, but it is

relatively new for anomaly detection [1]. For most, the issue lies in how to effectively measure the accuracy and the diversity of models, and the challenge of how to combine them [25]. And when the ensemble technique is combined with the deep learning method, the unsupervised nature makes it challenging for performance assessment and interpretability. This could explain the absence of an advanced state-of-the-art research [2]. Some methods are suitable for the realization of ensemble techniques for anomaly detection, such as feature bagging [16], subsampling [26], and pruning [19], which can successfully reduce the variance or bias of the model.

Another important research direction related to our work is deep learning interpretability. Although the term interpretability is universally used in machine learning, its definition is still vague [17]. The reasons why interpretability is crucial have been discussed [3] [17], e.g., understanding how a model works or how the prediction helps the user gain trust and confidence for the user; finding unnoticed causality; providing more ethical and fair decision-making, etc. To accurately evaluate a model's interpretability, we must first understand what exactly interpretability is. Some suggested that there are several types of model explanations, such as intrinsically interpretable models vs. post hoc explanations [21], or taxonomy of three dimensions: type of engagement, type of explanation, and the focus [24]. Different types of evaluation methods have also been introduced [11], such as application-grounded evaluation, human-grounded metrics, and functionally-grounded evaluation. However, the question of how to evaluate the quality of interpretability remains open. In this thesis, we proposed our novel way of constructing interpretability for the task of anomaly detection performance scores.

2 Contribution

Although the original DEAN algorithm achieved advanced results in many ways compared to other existing anomaly detection methods [14], there is still room for improvement. We focus on improving/reforming the feature bagging process by creating several extensions and experiments. The goal is to reach a better interpretation performance on our evaluation method. In section 2.1, we introduce a new evaluation method based on kNN. The details and results of each extension will be then discussed in the following sections. Section 2.7 also gives an example of how our model works on image data.

Table 2.1 lists all the data sets used in this thesis. Our ensemble consists of 2000 models. We used data sets with different sizes and dimensions to compare the outcomes, such as high-dimensional data set gas-drift with 128 features or low-dimensional data page-blocks and Magic Telescope with respectively 10 and 11 features only.

Name	X_train	X_test	Y_test
page-blocks	(4353, 10)	(1120, 10)	(1120,)
Magic Telescope	(8633, 11)	(7398, 11)	(7398,)
cardio	(1479, 21)	(352, 21)	(352,)
ionosphere	(135, 32)	(180, 32)	(180,)
satellite	(2640, 36)	(3518, 36)	(3518,)
waveform-5000	(1185, 40)	(1014, 40)	(1014,)
gas-drift	(1796, 128)	(1538, 128)	(1538,)
animals	(700, 150, 150)	(2203, 150, 150)	(2203,)

Table 2.1: List of data sets that were used in this paper and their size.

2.1 Evaluation

The unsupervised nature and small sample space pose a challenge to effectively evaluating the performance of anomaly detection [1]. One common way to achieve it is by using *ROC Area Under Curve* (AUC) [7]. In this thesis, we focus on the evaluation of the interpretability performance of our model. The definition of model interpretability lacks agreement among research papers, and it is also hard to objectively evaluate [17]. We created our own evaluation method based on the external measure. The main idea is to compare the importance score we generated(I_{DEAN}) for each feature to a reference importance score(I_{kNN}). The importance score of a feature refers to how the feature contributes to the determination of anomaly. Similar to the concept of individual conditional expectation (ICE) plot, in which each line indicates an instance and shows how its prediction is influenced by different features [18]. In other words, if a crucial(important) feature is removed, the output of the anomaly detection model should change. Based on this notion, we remove each feature iteratively and apply kNN to compute the anomaly score. Whether we remove an important or unimportant feature for a normal data point, the kNN algorithm will still consider it normal. As for an abnormal data point, removing a critical feature results in a lower kNN score. We then reverse the score since a lower score implies that the removal of a specific feature has a higher impact on the prediction.

kNN is a classification algorithm based on the Euclidean distance. As a "lazy algorithm," it has the advantage of simple implementation and no training period [23]. When we reduce the dimension of feature space, the distance between two points p, q (in the case of two dimensions) also reduces from $d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}$ to $d'(p,q) = (q_1 - p_1)$ or $(q_2 - p_2)$, depending on which dimension been removed. We know that $d'(p,q)$ is smaller than $d(p,q)$ based on Pythagorean theorem. It works particularly well in low-dimensional

feature space. Therefore, we employed kNN to compute the outlier scores after removing each feature to obtain a reference score. However, the major drawback of kNN method is the instability in high dimensional feature space [5]. The problem of "The Curse of Dimensionality" occurs as Euclidean distance loses its meaning when the data set has a high dimension [4]. This issue could restrain the ability of our evaluation method for high dimensional data sets and is also why I_{DEAN} stands out from the kNN method.

We calculate I_{DEAN} score by training our model and counting the times each feature is selected, then using the sum of their assigned anomaly scores to get an average of it. Each feature has a score based on how "important" they are.

The next step is to compare the two importance scores sets I_{kNN} and I_{DEAN}. Our evaluation system can be represented with the following formula:

$$\frac{2}{n * (n-1)} \cdot \sum_{i=0}^{n} \sum_{j=i+1}^{n} (I_{kNN}(x_i) < I_{kNN}(x_j)) = (I_{DEAN}(x_i) < I_{DEAN}(x_j)) \tag{2.1}$$

where:

n = the number of features

$I_{DEAN}(x)$ = original importance scores of feature x, generated by DEAN algorithm

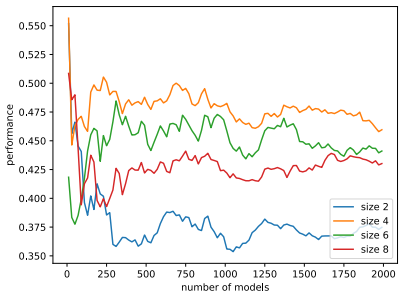
$I_{kNN}(x)$ = reference importance score of feature x, generated by our evaluation system based on kNN

We compared each score from the original and reference importance scores in pairs with one another. Each positive comparison, where score i and j have the same tendency in I_{kNN} and I_{DEAN}, will score one point. This score lies in the range of 0 to 1; if there is no relation between I_{DEAN} and I_{kNN}, the score is smaller. In the end, we averaged the accumulated points and received the evaluation result. This final score is invariant under monotonous transformations, e.g., scaling, translation, and exponential relationships.

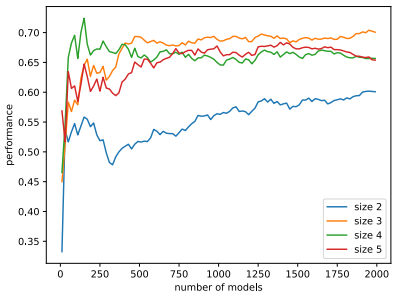
2.2 Feature Bagging Size

Feature bagging is a common ensemble learning method. Instead of training the models on the whole data set, each ensemble is assigned to a subset of randomly selected features and outputs individual outlier scores which can be combined into a final result. The feature bagging method has the benefit of reducing the variance within a learning model [14], thus resulting in better robustness. However, finding an optimal bagging size is tricky. A bigger bagging size increases the chance of each feature being selected, which provides more information to each model and allows more models to contribute to each I_{DEAN}. On the contrary, a smaller bagging size limits the number of features being considered, which gives us a clearer view of the impact of the particular feature on each model. In practice, it shows better interpretability performance. Based on our observation, the former improves the ROC score, while the latter increases interpretability.

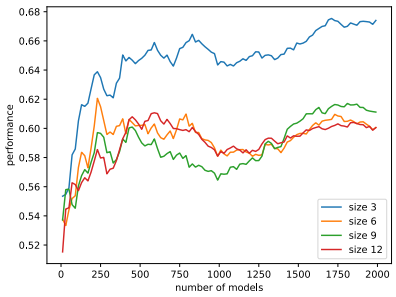
In this section, we compared each data set's interpretability performance of different bagging sizes. As shown in Figure 2.1, it is true that, in general, a smaller bagging size ($bag = 5$) generates better results. This indicates an advantage in reducing the processing time of the feature bagging method. For data sets like gas-drift with 128 features, a lower bag still brought a better performance than a much higher bag. One could also suppose that there is no clear correlation between the size of the data set and its optimal bagging size. For data sets such as page-blocks and MagicTelescope, $bag = 2$ is too small of a bagging size regarding the small sample dimensions, whereas $bag = 3$ yields the optimal interpretability performance. Some data sets have intrinsically worse performance than others, but the principle of choosing its ideal bagging size remains the same.



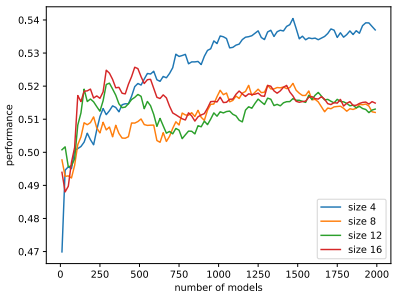
(a) page-blocks



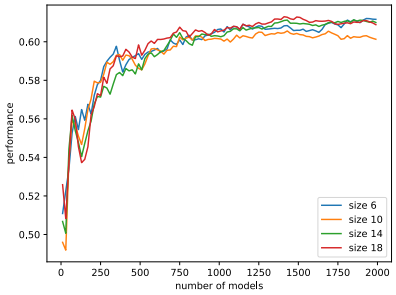
(b) Magic Telescope



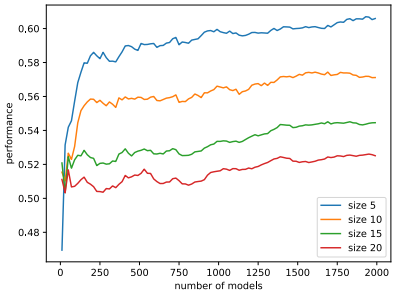
(c) cardio



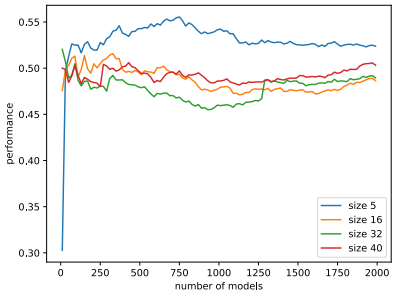
(d) ionosphere



(e) satellite



(f) waveform-5000



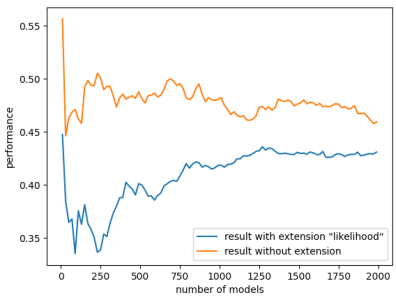
(g) gas-drift

Figure 2.1: Results of different feature bagging size in each data set.

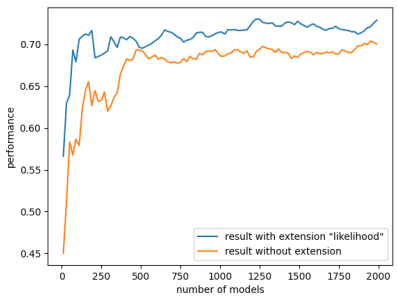
2.3 Reinforced Feature Importance

From the previous section, we obtained the optimal bagging size of each data set and the outlier scores of each of their features. The result is then reused in this section, where we train our models exclusively on certain selected features of each testing sample. Since the total number of selected features stays constant, we tested three different variations of sampling strategies in the following sections: reinforced feature importance, reversed feature importance, and equal probability of feature selection. These sampling adjustments create a feature space with lower variance, which creates more models with specific features, hopefully obtaining more accurate information about them and decreasing the uncertainty of their importance.

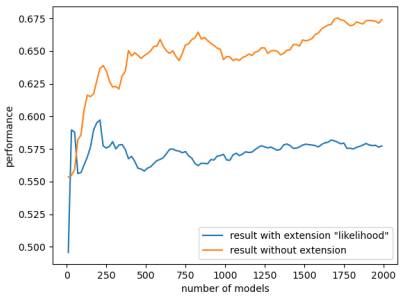
In this section, we manipulated the training data so that the top 50% features with higher outlier scores remained, and the rest were set to zero. We used the altered data set to train our models with the feature bagging method and compare the performance to the result before alteration in Figure 2.2. This extension only slightly improved the performance of a few data sets, such as satellite and Magic Telescope. As for data set waveform-5000, ionosphere, and gas-drift, the performances are similar with and without the extension. On the other hand, data set cardio has a big performance drop after applying the extension.



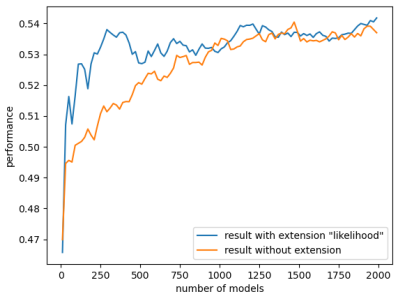
(a) page-blocks



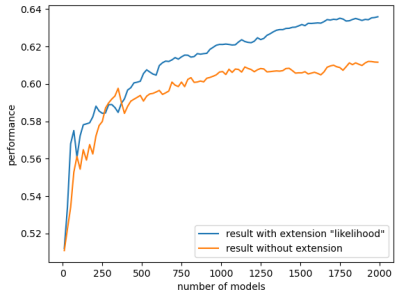
(b) Magic Telescope



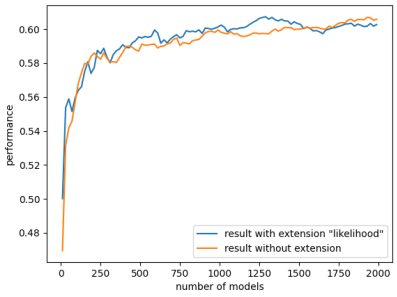
(c) cardio



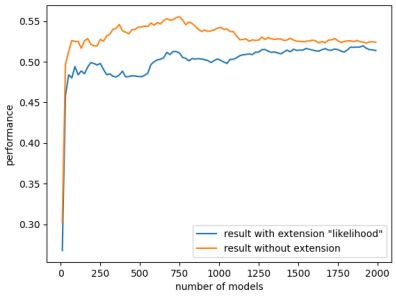
(d) ionosphere



(e) satellite



(f) waveform-5000



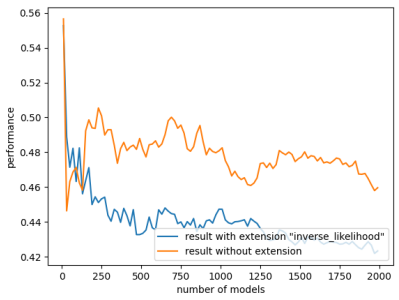
(g) gas-drift

Figure 2.2: Interpretability performance after reinforced feature importance adjustment compared to original result in each data set.

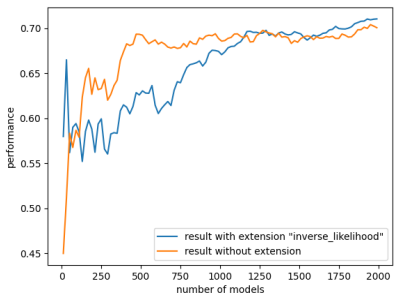
2.4 Reversed Feature Importance

In this section, we applied another extension to the feature bagging process. Reversed feature importance implies the opposite of what we did in section 2.4. This time, the 50% least important features remained, and the rest were set to zero. We used these less essential features to train our model. According to the results in Figure 2.3, for most data sets, the result without this extension either outperformed or stayed similar to the one with the extension, except for the data set ionosphere.

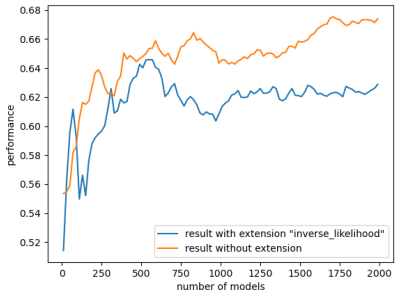
We observed that some data sets are more sensitive to the adjustment of feature bagging. In addition, the results from these two sections show that using more important features does not necessarily improve the performance of our model.



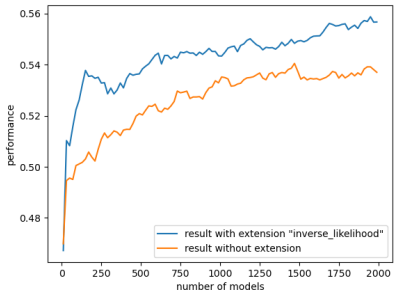
(a) page-blocks



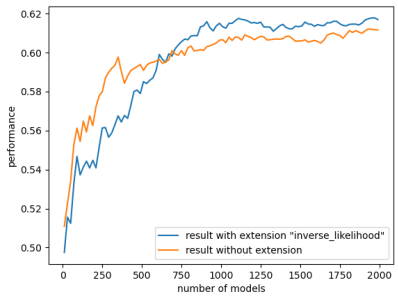
(b) Magic Telescope



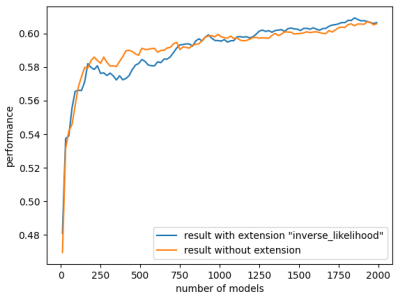
(c) cardio



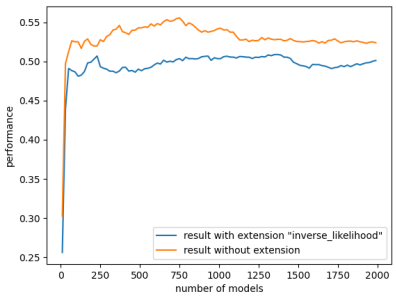
(d) ionosphere



(e) satellite



(f) waveform-5000

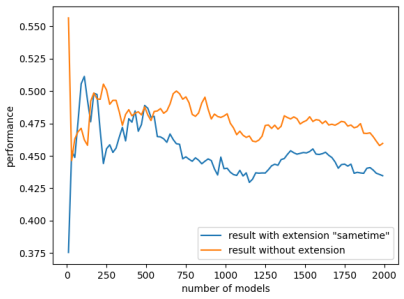


(g) gas-drift

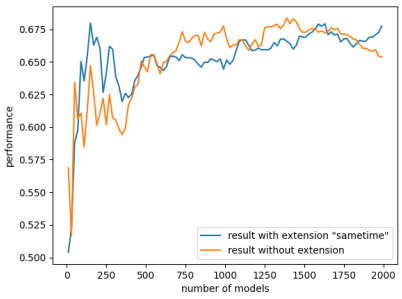
Figure 2.3: Interpretability performance after reversed feature importance adjustment compared to original result in each data set.

2.5 Equal Probability of Feature Selection

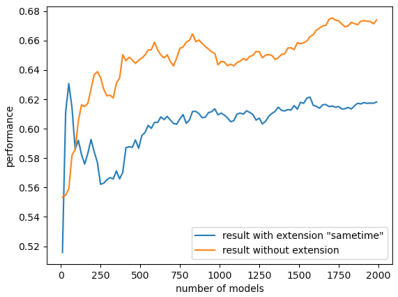
In the third variation of sampling, each feature was randomly selected for the same amount of time throughout the model training iteration. We maximized the usage of each feature and therefore reduced the bias. However, the result in Figure 2.4 showed no clear improvement after applying this extension. Data set ionosphere again yielded a positive outcome. We supposed this is due to the intrinsic quality of the data set ionosphere being better than the others.



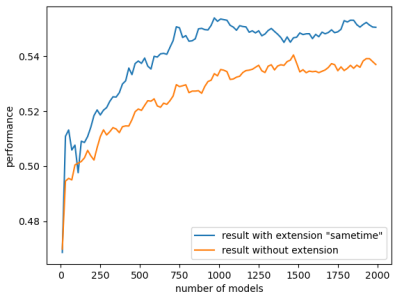
(a) page-blocks



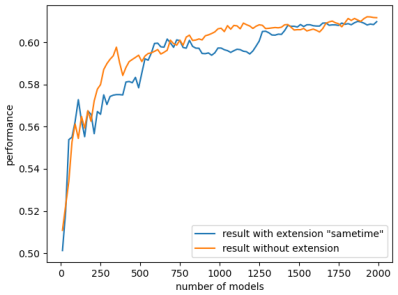
(b) Magic Telescope



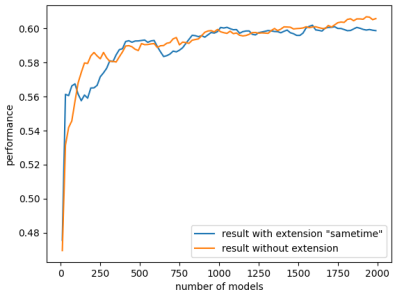
(c) cardio



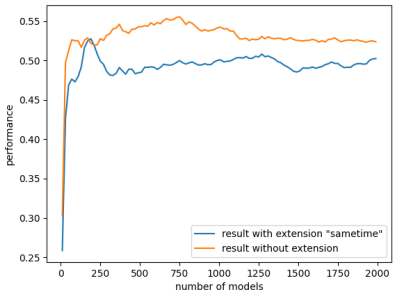
(d) ionosphere



(e) satellite



(f) waveform-5000



(g) gas-drift

Figure 2.4: Interpretability performance after equal probability adjustment compared to original result in each data set.

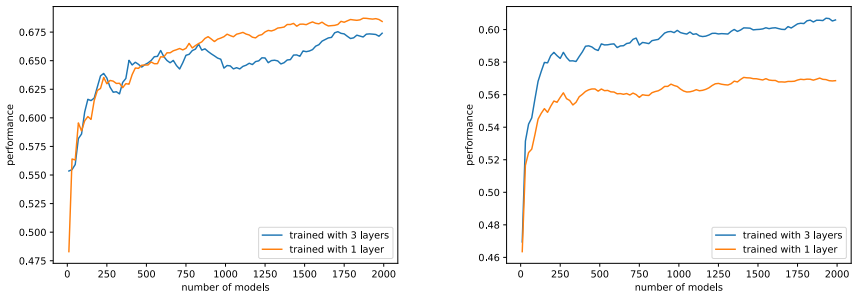
We conclude that our feature selection adjustment is not the optimal way to improve the sampling process. After implementing each extension, we can see no consistent improvement throughout the data sets. The results show that the effect is minor and limited to specific data sets. The first and third extensions help improve performance in some instances. It may have a more significant impact when fewer models or more features are involved.

2.6 Runtime

The paper of DEAN shows that our method achieved better results in different aspects, including runtime, compared to other algorithms [14]. In this section, we explored the possibility of reducing runtime furthermore without sacrificing the performance of interpretability. Two experiments were conducted: reducing the number of layers of the neural networks and finding convergence points.

2.6.1 Layer Reduction

The original algorithm is composed of two hidden layers and one output layer. We reduced it to a single-layer network and trained it on several data sets. In Figure 2.6. displays two different results. For specific data sets like cardio, the adjustment of hyperparameter *layer* from 3 to 1 maintains a good interpretability performance. It is yet not the case for other data sets. The result of data set waveform-5000 shows that a deep learning structure is still necessary for a robust anomaly detection model. The complex learning process of the deep learning model benefits interpretability. Considering the result of cardio, a single-layer network might still be helpful in special cases, such as simple anomaly detection tasks.



(a) cardio (b) waveform-5000

Figure 2.5: Results of interpretability performance on differenct layer size.

2.6.2 Convergence Point

Since our ensemble shows converged behavior, another possible way of reducing runtime is to find a convergence point regarding the interpretability performance. The earlier the convergence point is, the fewer models we need for training our ensemble to reach a comparable result. We trained each data set with 30000 models and attempted to find the first point where all the running standard deviations after it are constantly smaller than 0.01. Running standard deviations are calculated by subsetting the entire data set into different observation sets and obtaining a series of standard deviations. We set the size of the observation set to 50 and computed the standard deviations in each set.

Figure 2.6 displays the convergence point of each data set. The x-axis is sorted ascendingly by the number of features of the data set. For a high dimensional data set like gas-drift, the convergence point is similar to one of a low dimensional data set like page-blocks, around 1100 models. It is much less than the 30000 models we tested. This smaller convergence point represents that it takes less time than expected to reach an ideal interpretability performance. Therefore, the runtime can also be reduced by reducing the number of models needed.

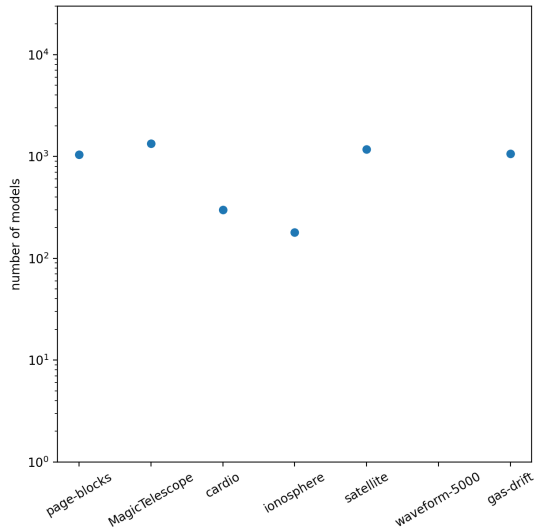


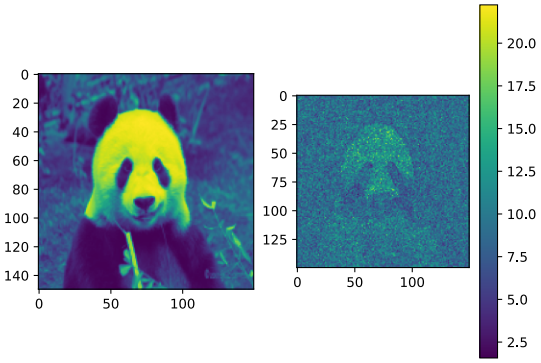
Figure 2.6: Convergence point of each data set(sorted by number of features).

This result indicates that we can reduce runtime by finding a smaller convergence point. In addition, there is no clear correlation between the data set's convergence point and feature dimension. A higher dimensional data does not necessarily require more models than a low dimensional data set.

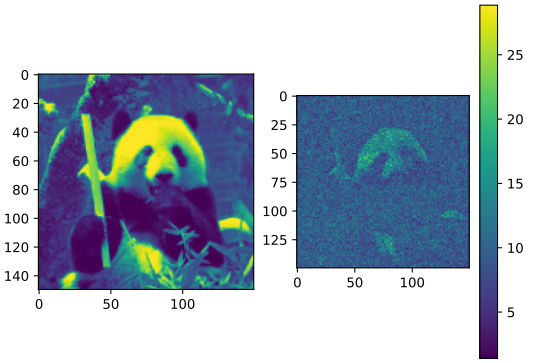
2.7 Results of image data

In this section, we share the result of our model used on image data sets. We used the data set animals and trained our model with dog images. In the images shown in Figure 2.8, our model has successfully detected panda images as abnormal and given us the reason for the anomaly. In the heap map, the lighter part has a higher anomaly score, representing higher feature importance in determining abnormality. As the model learned dog images to be normal, it shows why these images are not considered dogs. We can see clearly that the lighter-color pixels are primarily in the head area. We know that dogs do not have a head shape like a bear, so our model picked up this detail and marked

it as abnormal. The panda's iconic black and white fur color is also a good indicator to distinguish these two species. We rarely see a dog with a fur color contrasting head and body like a panda. This could be one of the reasons why our model marked the white fur part as abnormal since dogs have more consistent fur coloration throughout the entire body.



(a)



(b)

Figure 2.7: Results of interpretability on image data.

3 Conclusion and Future Work

In this thesis, we examined our DEAN algorithm with different extensions, aiming to improve the interpretability of our model. We also introduced a new evaluation system for assessing interpretability using kNN algorithm. Multiple real-life data sets were used for a comprehensive comparison. First, we determined the optimal feature bagging size for each data set with the conclusion that a smaller bag is universally ideal for the models, regardless of the dimension of the data set. We also introduced a new evaluation system for assessing interpretability. Although the experiments with extensions did not return optimal results on every data set, there is still potential utility for performance improvement, depending on the component and quality of the features. In the last part, we found the convergence point for each data set, and the result shows that the convergence point does not depend on the size of the data set. In summary, our work in this thesis manifests the capability and scalability of the ensemble method on anomaly detection tasks and their potential for the interpretability of deep learning models.

In the future, we expect more effort to be put into the field of deep ensemble anomaly detection and further improvement to the feature bagging process. One method to increase the anomaly detection quality could be applying averaged feature importance to filter the features. Another possibility would be observing the correlations between features after slightly altering their values. Since evaluating interpretability has long been a challenge in high-dimensional data, we are also intrigued to see further development on such a topic.

List of Figures

List of Tables

Bibliography

[1] Charu C Aggarwal. 2013. Outlier ensembles: position paper. *ACM SIGKDD Explorations Newsletter* 14, 2 (2013), 49–58.

[2] Charu C Aggarwal. 2017. An introduction to outlier analysis. In *Outlier analysis*. Springer, 1–34.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[4] Richard Bellman and Robert Kalaba. 1959. On adaptive control processes. *IRE Transactions on Automatic Control* 4, 2 (1959), 1–9.

[5] Aditya D. Bhat, Harshith R. Acharya, and Srikanth H.R. 2019. A Novel Solution to the Curse of Dimensionality in Using KNNs for Image Classification. In *2019 2nd International Conference on Intelligent Autonomous Systems (ICoIAS)*. 32–36. https://doi.org/10.1109/ICoIAS.2019.00012

[6] Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, and J Christopher Westland. 2011. Data mining for credit card fraud: A comparative study. *Decision support systems* 50, 3 (2011), 602–613.

[7] Andrew P Bradley. 1997. The use of the area under the ROC curve in the

evaluation of machine learning algorithms. *Pattern recognition* 30, 7 (1997), 1145–1159.

[8] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. 2017. Interpretability of deep learning models: A survey of results. In *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*. IEEE, 1–6.

[9] Raghavendra Chalapathy and Sanjay Chawla. 2019. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407* (2019).

[10] Federico Di Mattia, Paolo Galeone, Michele De Simoni, and Emanuele Ghelfi. 2019. A survey on gans for anomaly detection. *arXiv preprint arXiv:1906.11632* (2019).

[11] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[12] Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. 2019. Statistical analysis of nearest neighbor methods for anomaly detection. *Advances in Neural Information Processing Systems* 32 (2019).

[13] Rui Hou, Yifan Peng, Lars J Grimm, Yinhao Ren, Maciej A Mazurowski, Jeffrey R Marks, Lorraine M King, Carlo C Maley, E Shelley Hwang, and Joseph Y Lo. 2021. Anomaly Detection of Calcifications in Mammography Based on 11,000 Negative Cases. *IEEE Transactions on Biomedical Engineering* 69, 5 (2021), 1639–1650.

[14] Simon Klüttermann and Emmanuel Müller. [n. d.]. DEAN: Deep Ensemble Anomaly Detection. ([n. d.]). https://github.com/psorus/DEAN.

[15] Donghwoon Kwon, Hyunjoo Kim, Jinoh Kim, Sang C Suh, Ikkyun Kim, and Kuinam J Kim. 2019. A survey of deep learning-based network anomaly detection. *Cluster Computing* 22, 1 (2019), 949–961.

[16] Aleksandar Lazarevic and Vipin Kumar. 2005. Feature bagging for outlier detection. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. 157–166.

[17] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[18] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

[19] Shebuti Rayana and Leman Akoglu. 2016. Less is more: Building selective anomaly ensembles. *Acm transactions on knowledge discovery from data (tkdd)* 10, 4 (2016), 1–33.

[20] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *International conference on machine learning*. PMLR, 4393–4402.

[21] Jonas Herskind Sejr and Anna Schneider-Kamp. 2021. Explainable outlier detection: What, for Whom and Why? *Machine Learning with Applications* 6 (2021), 100172.

[22] Weng-Keen Wong, Andrew W Moore, Gregory F Cooper, and Michael M Wagner. 2003. Bayesian network anomaly pattern detection for disease outbreaks. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 808–815.

[23] Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Ruili Wang. 2018. Efficient kNN Classification With Different Numbers of Nearest

Neighbors. *IEEE Transactions on Neural Networks and Learning Systems* 29, 5 (2018), 1774–1785. https://doi.org/10.1109/TNNLS.2017.2673241

[24] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. 2021. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2021).

[25] Arthur Zimek, Ricardo JGB Campello, and Jörg Sander. 2014. Ensembles for unsupervised outlier detection: challenges and research questions a position paper. *Acm Sigkdd Explorations Newsletter* 15, 1 (2014), 11–22.

[26] Arthur Zimek, Matthew Gaudet, Ricardo JGB Campello, and Jörg Sander. 2013. Subsampling for efficient and effective unsupervised outlier detection ensembles. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 428–436.

Eidesstattliche Versicherung

(Affidavit)

Tang, Hsin Ping — 216568

Name, Vorname (surname, first name) — Matrikelnummer (student ID number)

☒ Bachelorarbeit (Bachelor's thesis) ☐ Masterarbeit (Master's thesis)

Titel (Title)

Interpreting Anomaly Detection: Optimized Preprocessing for Ensemble Methods

Ich versichere hiermit an Eides statt, dass ich die vorliegende Abschlussarbeit mit dem oben genannten Titel selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

I declare in lieu of oath that I have completed the present thesis with the above-mentioned title independently and without any unauthorized assistance. I have not used any other sources or aids than the ones listed and have documented quotations and paraphrases as such. The thesis in its current or similar version has not been submitted to an auditing institution before.

Bochum, 23.10.22 — Hsin Ping Tang

Ort, Datum (place, date) — Unterschrift (signature)

Belehrung:
Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz - HG -).

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird ggf. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin") zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:
Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offense can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offenses of this type is the Chancellor of TU Dortmund University. In the case of multiple or other serious attempts at deception, the examinee can also be unenrolled, Section 63 (5) North Rhine-Westphalia Higher Education Act (*Hochschulgesetz, HG*).

The submission of a false affidavit will be punished with a prison sentence of up to three years or a fine.

As may be necessary, TU Dortmund University will make use of electronic plagiarism-prevention tools (e.g. the "turnitin" service) in order to monitor violations during the examination procedures.

I have taken note of the above official notification:*

Bochum, 23.10.22 — Hsin Ping Tang

Ort, Datum (place, date) — Unterschrift (signature)

***Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the Bachelor's/ Master's thesis is the official and legally binding version.**