

Improved Anomaly Detection by Using the Attention-Based Isolation Forest

Lev V. Utkin, Andrey Y. Ageev and Andrei V. Konstantinov

Peter the Great St.Petersburg Polytechnic University

St.Petersburg, Russia

e-mail: lev.utkin@gmail.com, andreyageev1@mail.ru, andrue.konst@gmail.com

Abstract

A new modification of Isolation Forest called Attention-Based Isolation Forest (ABIForest) for solving the anomaly detection problem is proposed. It incorporates the attention mechanism in the form of the Nadaraya-Watson regression into the Isolation Forest for improving solution of the anomaly detection problem. The main idea underlying the modification is to assign attention weights to each path of trees with learnable parameters depending on instances and trees themselves. The Huber's contamination model is proposed to be used for defining the attention weights and their parameters. As a result, the attention weights are linearly depend on the learnable attention parameters which are trained by solving the standard linear or quadratic optimization problem. ABIForest can be viewed as the first modification of Isolation Forest, which incorporates the attention mechanism in a simple way without applying gradient-based algorithms. Numerical experiments with synthetic and real datasets illustrate outperforming results of ABIForest. The code of proposed algorithms is available.

Keywords: anomaly detection, attention mechanism, Isolation Forest, Nadaraya-Watson regression, quadratic programming, contamination model

1 Introduction

One of the important machine learning problems is the novelty or anomaly detection problem which aims to detect abnormal or anomalous instances. This problem can be regarded as a challenging task because there is no a strong definition of anomalous instance and the anomaly itself depends on a certain application. Another difficulty which defines the challenge of the problem is that anomalies usually seldom appear and this fact leads to highly imbalanced training sets. Moreover, it is difficult to define a boundary between the normal and anomalous observations [1]. Due to importance of the anomaly detection problem in many applications, a huge amount of papers covering anomaly detection tasks and studying various aspects of the anomaly detection have been published in the last decades. Many approaches to solving the anomaly detection problem are analyzed in comprehensive survey papers [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11].

According to [1, 12], anomalies also referred to as abnormalities, deviants, or outliers can be viewed as data points which are located further away from the bulk of data points that are referred to as normal data.

Various approaches to solving the anomaly detection problem can be divided into several groups [10]. The first group consists of the probabilistic and density estimation models. It includes the classic density estimation models, energy-based models, neural generative models [10]. The second large group deals with the one-class classification models. This group includes the well-known one-class classification SVMs [13, 14, 15]. The third group includes reconstruction-based models which detect anomalies by reconstructing the data instances. The well-known models from this group are autoencoders which incorrectly reconstruct anomalous instances such that the distance between the instance and its reconstruction is larger than a predefined threshold which is usually regarded as a hyperparameter of the model.

The next group contains distance-based anomaly detection models. One of the most popular and effective models from the group is the Isolation Forest (iForest) [16, 17] which is a model for detecting anomalous points relative to a certain data distribution. According to iForest, anomalies are detected using isolation which measures how far an instance is from the rest of instances. iForest can be regarded as a tool implementing the isolation. It has the linear time complexity and works well with large amounts of data. The core idea behind iForest is the tendency for anomalous instances in a dataset to be more easily separated from the rest of the sample (isolated) compared to normal instances. To isolate a data point, the algorithm recursively creates sample partitions by randomly choosing an attribute and then randomly choosing a split value for the attribute between the minimum and maximum values allowed for that attribute. The recursive partition can be represented by a tree structure called an isolation tree, while the number of partitions needed to isolate a point can be interpreted as the length of the path within the tree to the end node, starting from the root. Anomalous instances are those with a shorter path length in the tree [16, 17].

In order to improve iForest, we propose to modify it by using the attention mechanism which can automatically distinguish the relative importance of instances and weigh them for improving the overall accuracy of iForest. The attention mechanism has been successfully applied to many applications, including the natural language processing models, the computer vision area, etc. Comprehensive surveys of properties and forms of the attention mechanism and transformers can be found in [18, 19, 20, 21, 22].

The idea to apply the attention mechanism to iForest stems from the attention-based random forest (ABRF) models proposed in [23] where attention is implemented in the form of the Nadaraya-Watson regression [24, 25] by assigning attention weights to leaves of trees in a specific way such that the weights depend on trees and instances. The attention learnable parameters in ABRF are trained by solving the standard quadratic optimization problem with linear constraints. It turns out that this idea to consider the random forest as the Nadaraya-Watson regression [24, 25] can be extended to iForest taking into account the iForest peculiarities which differ it from the random forest. According to the original iForest, the isolation measure is estimated as the mean value of the path lengths over all trees in the

forest. However, we can replace the averaging of the path lengths with the Nadaraya-Watson regression where the path length of an instance in each tree can be regarded as a prediction in the regression (the *value* in terms of the attention mechanism [27]), and weights (the attention weights) depend on the corresponding tree and the instance (the *query* in terms of the attention mechanism [27]). In other words, the final prediction of the expected path length in accordance with the Nadaraya-Watson regression is a weighted sum of path lengths over all trees. Weights of path lengths have learnable parameters (the learnable attention parameters) which can be computed by minimizing a loss function of a specific form. We aim to reduce the optimization problem to the quadratic programming problem or linear programming problem which has many algorithms for solving. In order to achieve this aim, the Huber’s ϵ -contamination model [26] is proposed to be used for computing the learnable attention parameters. The contamination model allows us to represent attention weights in the form of a linear combination of the softmax operation and learnable parameters with contamination parameter ϵ , which can be viewed as probabilities. As a result, the loss function for computing learnable parameters is linear with linear constraints on the parameters as probabilities. After adding the L_2 regularization term, the optimization problem for computing attention weights becomes to be quadratic one.

Our contributions can be summarized as follows:

1. A new modification of iForest called Attention-Based Isolation Forest (ABIForest) incorporating the attention mechanism in the form of the Nadaraya-Watson regression for improving solution of the anomaly detection problem is proposed.
2. The algorithm of computing attention weights is reduced to solving the linear or quadratic programming problems due to applying the Huber’s ϵ -contamination model. Moreover, we propose to use the hinge-loss function to simplify the optimization problem. Contamination parameter ϵ is regarded as a tuning hyperparameter.
3. Numerical experiments with synthetic and real datasets are performed for studying ABIForest. They demonstrate outperforming results for most datasets. The code of proposed algorithms can be found in <https://github.com/AndreyAgeev/Attention-based-isolation-forest>.

The paper is organized as follows. Related work can be found in Section 2. Brief introductions to the attention mechanism, the Nadaraya-Watson regression and iForest are given in Section 3. The proposed ABIForest model is considered in Section 4. Numerical experiments with synthetic and real datasets illustrating peculiarities of ABIForest and its comparison with iForest are provided in Section 5. Concluding remarks discussing advantages and disadvantages of ABIForest can be found in Section 6.

2 Related work

Attention mechanism. The attention mechanism can be viewed as an effective method for improving the performance of a large variety of machine learning models. Therefore,

there are many different types of attention mechanisms depending on their applications and models where attention mechanisms are incorporated. The term “attention” was introduced by Bahdanau et al. [27]. Following this paper, a huge amount of models based on the attention mechanism can be found in the literature. There are also several types of attention mechanisms [28], including soft and hard attention mechanisms [29], the local and global attention [30], self-attention [31], multi-head attention [31], hierarchical attention [32]. It is difficult to consider all papers devoted to the attention mechanisms and its applications. Comprehensive surveys [18, 19, 20, 21, 22, 33] cover a large part of available models and modifications of the attention mechanisms.

Most attention models are implemented as parts of neural networks. In order to extend a set of attention models, several random forest models incorporated with the attention mechanism were proposed in [23, 34, 35]. The gradient boosting machine added by the attention mechanism was presented in [36].

Anomaly detection with attention. A wide set of machine learning tasks include anomaly detection problems. Therefore, many methods and models have been developed to address them [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. One of the tools for solving the anomaly detection problems is the attention mechanism. Monotonic attention based autoencoders was proposed in [37] as an unsupervised learning technique to detect the false data injection attacks. Anomaly detection method based on the Siamese network with an attention mechanism for dealing with small datasets was proposed in [38]. The so-called residual attention network that employs the attention mechanism and residual learning to improve classification efficiency and accuracy was presented in [39]. The graph anomaly detection algorithm based on the attention-based deep learning to assist the audit process was provided in [40]. Madan et al. [41] presented a novel self-supervised masked convolutional transformer block that comprises the reconstruction-based functionality. Integration of the reconstruction-based functionality into a novel self-supervised predictive architectural building block was considered in [42]. Huang et al. [43] improved the efficiency and effectiveness of anomaly detection and localization at inference by using a progressive mask refinement approach that progressively uncovers the normal regions and finally locates the anomalous regions. A novel self-supervised framework for multivariate time-series anomaly detection via a graph attention network was proposed in [44]. It can be seen from the above works that the idea to apply attention in models solving the anomaly detection problem was successfully implemented. However, the attention was used in the form of components of neural networks. There are no forest-based anomaly detection models which use the attention mechanism.

iForest. iForest [16, 17] can be viewed as one of the important and effective methods for solving novelty and anomaly detection problems. Therefore, many modifications of the method have been developed [5] to improve it. A weighted iForest and Siamese Gated Recurrent Unit algorithm architecture which provides a more accurate and efficient method for outlier detection of data is considered in [45]. Hariri et al. [46] proposed an extension of the iForest, named Extended Isolation Forest, which resolves issues with assignment of anomaly score to given data points. A theoretical framework that describes the effectiveness of isolation-based approaches from a distributional viewpoint was studied in [47]. Lesouple

et al. [48] presented a generalized isolation forest algorithm which generates trees without any empty branch, which significantly improves the execution times. The k-Means-Based iForest was developed by Karczmarek et al. [49]. This modification of iForest allows to build a search tree based on many branches in contrast to the only two considered in the original method. Another modification, called the Fuzzy Set-Based Isolation Forest was proposed in [50]. A probabilistic generalization of iForest was proposed in [51], which is based on nonlinear dependence of a segment-cumulated probability from the length of segment. A robust anomaly detection method called the similarity-measured isolation forest was developed by Li et al. [52] to detect abnormal segments in monitoring data. A novel hyperspectral anomaly detection method with kernel Isolation Forest was proposed in [53]. The method is based on an assumption that anomalies rather than background can be more susceptible to isolation in the kernel space. An improved computational framework which allows us to seek the most separable attributes and spot corresponding optimized split points effectively was presented in [54]. Staerman et al. [55] introduced the so-called Functional Isolation Forest which generalizes iForest to the infinite dimensional context, i.e., the model deals with functional random variables that take their values in a space of functions. Xu et al. [56] proposed the Deep Isolation Forest which is based on an isolation method with arbitrary (linear/non-linear) partition of data implemented by using neural networks.

The above works is only a part of many extensions and modifications of iForest developed due to excellent properties of the method. However, to the best of our knowledge, there are no works considering approaches to incorporating the attention mechanism into iForest.

3 Preliminaries

3.1 Attention mechanism as the Nadaraya-Watson regression

If to consider the attention mechanism as a method for enhancing accuracy of iForest for the anomaly detection problem solution, then it allows us to automatically distinguish the relative importance of features, instances and isolation trees. According to [18, 57], the original idea of attention can be understood from the statistical point of view applying the Nadaraya-Watson kernel regression model [24, 25].

Given n instances $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, in which $\mathbf{x}_i = (x_{i1}, \dots, x_{id}) \in \mathbb{R}^d$ is a feature vector involving m features and $y_i \in \mathbb{R}$ represents the regression outputs, the task of regression is to construct a regressor $f : \mathbb{R}^m \rightarrow \mathbb{R}$ which can predict the output value \tilde{y} of a new observation \mathbf{x} , using available data S . The similar task can be formulated for the classification problem.

The original idea behind the attention mechanism is to replace the simple average of outputs $\tilde{y} = n^{-1} \sum_{i=1}^n y_i$ for estimating the regression output y , corresponding to a new input feature vector \mathbf{x} with the weighted average, in the form of the Nadaraya-Watson regression model [24, 25]:

$$\tilde{y} = \sum_{i=1}^n \alpha(\mathbf{x}, \mathbf{x}_i) y_i, \quad (1)$$

where weight $\alpha(\mathbf{x}, \mathbf{x}_i)$ conforms with relevance of the i -th instant to the vector \mathbf{x} , i.e., it is defined in agreement with the corresponding input \mathbf{x}_i locations relative to the input variable \mathbf{x} (the closer an input \mathbf{x}_i to the given variable \mathbf{x} , the greater $\alpha(\mathbf{x}, \mathbf{x}_i)$).

In terms of the attention mechanism [27], vectors \mathbf{x} , \mathbf{x}_i and outputs y_i are called as the *query*, *keys* and *values*, respectively. Weight $\alpha(\mathbf{x}, \mathbf{x}_i)$ is called as the attention weight.

The attention weights $\alpha(\mathbf{x}, \mathbf{x}_i)$ can be defined by a normalized kernel K as:

$$\alpha(\mathbf{x}, \mathbf{x}_i) = \frac{K(\mathbf{x}, \mathbf{x}_i)}{\sum_{j=1}^n K(\mathbf{x}, \mathbf{x}_j)}. \quad (2)$$

For the Gaussian kernel with parameter ω , the attention weights are represented through the softmax operation as:

$$\alpha(\mathbf{x}, \mathbf{x}_i) = \sigma\left(-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{\omega}\right). \quad (3)$$

In order to enhance the attention capability, weights are added by trainable parameters. Several definitions of attention weights and attention mechanisms have been proposed. The most popular definitions are the additive attention [27], the multiplicative or dot-product attention [58, 31].

3.2 Isolation forest

In this subsection, the main definitions of iForest are provided in accordance with results given in [16, 17]. Suppose that there is a dataset $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ consisting of n instances, where $\mathbf{x}_i = (x_{i1}, \dots, x_{id}) \in \mathbb{R}^d$ is a feature vector. The isolation tree is built by using a randomly generated subset \mathcal{D}^* of the dataset \mathcal{D} . The dataset \mathcal{D}^* splits into two subsets to define a random node as follows. A feature is randomly selected by generating the random value q from the set $\{1, \dots, d\}$. Then a split value p is randomly selected from interval $[\min_{i=1, \dots, n} x_{iq}, \max_{i=1, \dots, n} x_{iq}]$. Having p and q , the subset \mathcal{D}^* is recursively divided at each node by using the feature number q and the split value p into two parts: the left branch corresponding to the set with $x_{iq} \leq p$ and the right branch corresponding to the set with $x_{iq} > p$. Thus generated values q and p determine whether the data points at a node are sent down the left or the right branch. The above conditions determine the subsequent child nodes for a split node. The division stops in accordance with a rule, for example, when a branch contains a single point or when some limited depth of the tree is reached. The process of the isolation tree building begins again with a new random subsample to build another randomized tree. After building a forest consisting of T trees, the training process is complete.

In the k -th isolation tree, an instance \mathbf{x} is isolated on one of the outer nodes such that a path of length $h_k(\mathbf{x})$ can be associated with this instance, which is defined as a number of nodes that \mathbf{x} goes from the root node to the leaf. Anomalous instances are those with a shorter path length in the tree. This conclusion is motivated by the fact that normal instances are more concentrated than anomalies and thus require more nodes to be isolated. By having the trained T trees, i.e., the isolated forest, we can estimate the isolation measure

as the expected path length $E[h(\mathbf{x})]$ which is computed as the mean value of the path lengths over all trees in the forest. By having the expected path length $E[h(\mathbf{x})]$, an anomaly score is defined as

$$s(\mathbf{x}, n) = 2^{-\frac{E(h(\mathbf{x}))}{c(n)}}, \quad (4)$$

where $c(n)$ is the normalizing factor defined as the average value of $h(\mathbf{x})$ for a dataset of size n , which is computed as

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}. \quad (5)$$

Here $H(n)$ is the n -th harmonic number estimated from $H(n) = \ln(n) + \delta$, where $\delta \approx 0.577216$ is the Euler-Mascheroni constant. If $n = 2$, then $c(n) = 1$.

The higher the value of $s(\mathbf{x}, n)$ (closer to 1), the more likely the instance \mathbf{x} is anomalous. If we introduce a threshold $\tau \in [0, 1]$, then condition $s(\mathbf{x}, n) > \tau$ indicates that instance \mathbf{x} is detected as an anomaly. If condition $s(\mathbf{x}, n) \leq \tau$ is valid, then instance \mathbf{x} is likely normal. The threshold τ in the original iForest is taken 0.5.

4 Attention-Based Isolation Forest

It should be noted that the expected path length $E[h(\mathbf{x})]$ in the original iForest is computed as the mean value of the path lengths $h_k(\mathbf{x})$ of trees:

$$E[h(\mathbf{x})] = \frac{1}{T} \sum_{k=1}^T h_k(\mathbf{x}). \quad (6)$$

This method for computing the expected path length does not take into account the possible relationship between an instance and each isolation tree, the possible difference between trees. Ideas behind the attention-based RF [23] can also be applied to iForest. Therefore, our next task is to incorporate the attention mechanism into iForest.

4.1 Keys-values and the query in iForests

First, we can point out that the outcome of each isolation tree is the path length $h_k(\mathbf{x})$, $k = 1, \dots, n$. This implies that this outcome can be regarded as the value in the attention mechanism. Second, we define the query and keys in iForest. Suppose that the feature vector \mathbf{x} falls into the i -th leaf of the k -th tree. Let $\mathcal{J}_i^{(k)}$ be a set of indices of $n_i^{(k)}$ training instances \mathbf{x}_j which also fall into the same leaf. A distance between vector \mathbf{x} and all vectors \mathbf{x}_j , $j \in \mathcal{J}_i^{(k)}$, shows how vector \mathbf{x} is in agreement with corresponding vectors \mathbf{x}_j , how it is close to vectors \mathbf{x}_j from the same leaf. If the distance is small, then we can conclude that vector \mathbf{x} is well performed by the k -th tree. The distance between vector \mathbf{x} and all vectors \mathbf{x}_j , $j \in \mathcal{J}_i^{(k)}$, can be represented as a distance between vector \mathbf{x} and the mean values of all vectors \mathbf{x}_j with indices $j \in \mathcal{J}_i^{(k)}$. The mean vector of \mathbf{x}_j with indices $j \in \mathcal{J}_i^{(k)}$ can be

viewed as a characteristic of the corresponding path, i.e., this vector characterizes a group of instances which fall into the corresponding leaf. Hence, the mean vector shows how vector \mathbf{x} is in agreement with this group. If we denote the mean value of \mathbf{x}_j , $j \in \mathcal{J}_i^{(k)}$ as $\mathbf{A}_k(\mathbf{x})$, then there holds

$$\mathbf{A}_k(\mathbf{x}) = \frac{1}{n_i^{(k)}} \sum_{j \in \mathcal{J}_i^{(k)}} \mathbf{x}_j. \quad (7)$$

We omit the index j in $\mathbf{A}_k(\mathbf{x})$ because the instance \mathbf{x} can fall only into one leaf of each tree.

Vectors $\mathbf{A}_k(\mathbf{x})$ and \mathbf{x} can be regarded as the key and the query, respectively. Then (6) can be rewritten by using the attention weights $\alpha(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \mathbf{w})$ as follows:

$$E[h(\mathbf{x})] = \sum_{k=1}^T \alpha(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \mathbf{w}) \cdot h_k(\mathbf{x}), \quad (8)$$

where $\alpha(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \mathbf{w})$ conforms with relevance of “mean instance” $\mathbf{A}_k(\mathbf{x})$ to vector \mathbf{x} and satisfies condition

$$\sum_{k=1}^T \alpha(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \mathbf{w}) = 1, \quad \alpha(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \mathbf{w}) \geq 0, \quad k = 1, \dots, T. \quad (9)$$

We have replaced the expected path length (6) with the weighted sum of path lengths (8) such that weights α depend on \mathbf{x} , mean vector $\mathbf{A}_k(\mathbf{x})$ and vector of parameters \mathbf{w} . Vector \mathbf{w} in attention weights represents trainable attention parameters. Their values depend on the dataset and on the isolation tree properties. If we return to the Nadaraya-Watson kernel regression model, then the expected path length $E[h(\mathbf{x})]$ can be viewed as the regression output, path lengths $h_k(\mathbf{x})$ of all trees for query \mathbf{x} are predictions (values in terms of the attention mechanism [27]).

Suppose that the trainable parameters \mathbf{w} belong to a set \mathcal{W} . Then they can be found by solving the following optimization problem:

$$\mathbf{w}_{opt} = \arg \min_{\mathbf{w} \in \mathcal{W}} \sum_{s=1}^n L(E[h(\mathbf{x}_s)], \mathbf{x}_s, \mathbf{w}). \quad (10)$$

Here $L(E[h(\mathbf{x}_s)], \mathbf{x}_s, \mathbf{w})$ is the loss function whose definition as well as the definition of $\alpha(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \mathbf{w})$ are the next tasks.

4.2 Loss function and attention weights

First, we reformulate the decision rule ($s(\mathbf{x}, n) > \tau$) for determining anomalous instances by establishing a similar condition for $E[h(\mathbf{x})]$. Suppose that γ is a threshold such that condition $E[h(\mathbf{x})] \leq \gamma$ indicates that instance \mathbf{x} is detected as an anomaly. Then it follows from (4) that γ can be expressed through threshold τ as:

$$\gamma = -c(n) \cdot \log_2(\tau). \quad (11)$$

Hence, we can write the decision rule about the anomaly as follows:

$$\text{decision} = \begin{cases} \text{anomalous}, & \text{if } E[h(\mathbf{x})] - \gamma \leq 0, \\ \text{normal}, & \text{otherwise.} \end{cases} \quad (12)$$

Introduce also the instance label y_s which is 1 if the training instance \mathbf{x}_s is anomalous, and -1 if it is normal. If labels are not known, then prior values of labels can be determined by using the original iForest.

We propose the following loss function:

$$L(h(\mathbf{x}_s), \mathbf{x}_s, \mathbf{w}) = \max(0, y_s(E[h(\mathbf{x}_s)] - \gamma)). \quad (13)$$

It can be seen from (13) that the loss function is 0 if $E[h(\mathbf{x}_s)] - \gamma$ and y_s have different signs, i.e., if the decision about an anomalous (normal) instance coincides with the corresponding label. Substituting (8) into (13), we rewrite optimization problem (10) as:

$$\mathbf{w}_{opt} = \arg \min_{\mathbf{w} \in \mathcal{W}} \left[\sum_{s=1}^n \max \left(0, y_s \left(\sum_{k=1}^T \alpha(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \mathbf{w}) \cdot h_k(\mathbf{x}) - \gamma \right) \right) \right]. \quad (14)$$

An important question is how to simplify the above problem to get a unique solution and how to define the attention weights $\alpha(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \mathbf{w})$ depending on the trainable parameters \mathbf{w} . It can be done by using the Huber's ϵ -contamination model.

4.3 Huber's contamination model

We propose to use a simple representation of attention weights presented in [23], which is based on applying the Huber's ϵ -contamination model [26]. The model is represented as a set of discrete probability distributions F of the form:

$$F = (1 - \epsilon) \cdot P + \epsilon \cdot R, \quad (15)$$

where $P = (p_1, \dots, p_T)$ is a discrete probability distribution contaminated by another probability distribution denoted $R = (r_1, \dots, r_T)$ under condition that the probability distribution R can be arbitrary; the contamination parameter $\epsilon \in [0, 1]$ controls the degree of the contamination.

The contaminating distribution R is a point in the unit simplex with T vertices denoted as $S(1, T)$. The distribution F is a point in a small simplex which belongs to the unit simplex. The size of the small simplex depends on the hyperparameter ϵ . If $\epsilon = 1$, then the small simplex coincides with the unit simplex. If $\epsilon = 0$, then the the small simplex is reduced to a single distribution P .

We propose to consider every element of P as a result of the softmax operation

$$p_k = \sigma \left(-\frac{\|\mathbf{x} - \mathbf{A}_k(\mathbf{x})\|^2}{\omega} \right), \quad (16)$$

that is

$$P = \left(\sigma \left(-\frac{\|\mathbf{x} - \mathbf{A}_1(\mathbf{x})\|^2}{\omega} \right), \dots, \sigma \left(-\frac{\|\mathbf{x} - \mathbf{A}_T(\mathbf{x})\|^2}{\omega} \right) \right).$$

Moreover, we propose to consider the distribution R as the vector of trainable parameters \mathbf{w} , that is

$$R = \mathbf{w} = (w_1, \dots, w_T).$$

Hence, the attention weight $\alpha(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \mathbf{w})$ can be represented for every $k = 1, \dots, T$ as follows:

$$\alpha(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \mathbf{w}) = (1 - \epsilon) \cdot \sigma \left(-\frac{\|\mathbf{x} - \mathbf{A}_k(\mathbf{x})\|^2}{\omega} \right) + \epsilon \cdot w_k. \quad (17)$$

An important property of the above representation is that the attention weight linearly depends on the trainable parameters, and the softmax operation depends only on the hyper-parameter ω . The trainable parameters $\mathbf{w} = (w_1, \dots, w_T)$ are restricted by the unit simplex $S(1, T)$ and, therefore, $\mathcal{W} = S(1, T)$. This implies that the constraints for \mathbf{w} are linear ($w_i \geq 0$ and $w_1 + \dots + w_T = 1$).

4.4 Loss function with the contamination model

Let us substitute the obtained expression (17) for the attention weight $\alpha(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \mathbf{w})$ into the objective function (14). We get after simplification

$$\min_{\mathbf{w} \in S(1, T)} \sum_{s=1}^n \max \left(0, D_s(\epsilon, \omega) + y_s \epsilon \sum_{k=1}^T h_k(\mathbf{x}_s) w_k \right) \quad (18)$$

where

$$D_s(\epsilon, \omega) = y_s (1 - \epsilon) \sum_{k=1}^T \sigma \left(-\frac{\|\mathbf{x}_s - \mathbf{A}_k(\mathbf{x}_s)\|^2}{\omega} \right) - \gamma T. \quad (19)$$

Let us introduce new variables

$$v_s = \max \left(0, D_s(\epsilon, \omega) + y_s \epsilon \sum_{k=1}^T h_k(\mathbf{x}_s) w_k \right). \quad (20)$$

Then problem (18) can be rewritten as follows:

$$\min \sum_{s=1}^n v_s, \quad (21)$$

subject to

$$v_s \geq D_s(\epsilon, \omega) + y_s \epsilon \sum_{k=1}^T h_k(\mathbf{x}_s) w_k, \quad (22)$$

Table 1: A brief introduction about datasets

Dataset	n_{norm}	n_{anom}	d
Circle (synthetic)	1000	200	2
Normal dataset (synthetic)	1000	50	2
Credit	1500	400	30
Ionosphere	225	126	33
Arrhythmia	386	66	18
Mulcross	1800	400	4
Http	500	50	3
Pima	500	268	8

$$v_s \geq 0, \quad s = 1, \dots, n, \quad (23)$$

$$w_1 + \dots + w_T = 1, \quad w_k \geq 0, \quad k = 1, \dots, T. \quad (24)$$

This is a linear optimization problem with the optimization variables w_1, \dots, w_T and v_1, \dots, v_n .

The optimization problem can be improved by adding a regularization term $\|\mathbf{w}\|^2$ with the hyperparameter λ which controls the strength of the regularization. In this case, the optimization problem becomes

$$\min \sum_{s=1}^n v_s + \lambda \|\mathbf{w}\|^2, \quad (25)$$

subject to (22), (23), (24).

We get the standard quadratic programming problem whose solution does not meet any difficulties.

5 Numerical experiments

The proposed attention-based iForest is studied by using synthetic and real data and is compared with the original iForest. A brief introduction about these datasets is given in Table 1 where d is the number of features, n_{norm} and n_{anom} are numbers of normal and anomalous instances, respectively.

Different values for hyper-parameters, including threshold τ , the number of trees in the forest, the contamination parameter ϵ , the kernel parameter ω have been tested, choosing those leading to the best results. In particular, hyperparameter ϵ in ABIForest takes values 0, 0.25, 0.5, 0.75, 1; hyperparameter γ changes from 0.5 to 0.7; hyperparameter ω takes values 0.1, 10, 20, 30, 40. F1-score is used as a measure of the anomaly detection accuracy. To evaluate the F1-score, a cross-validation with 100 repetitions is performed, where in each run, 66.7% of data for training ($2n/3$) and 33.3% for testing ($n/3$) are randomly selected. Numerical results are presented in tables where the best results are shown in bold.

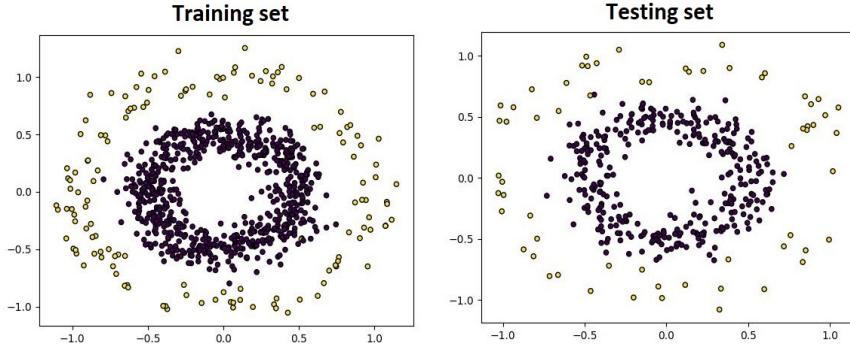


Figure 1: Points from the Circle dataset

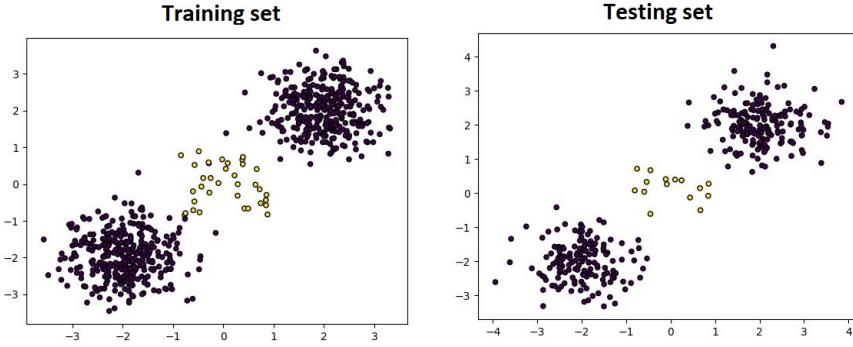


Figure 2: Points from the Normal dataset

5.1 Synthetic datasets

The first synthetic dataset used for numerical experiments is the *Circle* dataset. Its points are divided into two parts concentrated around small and large circles as it is shown in Fig. 1 where the training and testing sets are depicted in the left and rights pictures, respectively. In order to optimize the model parameters in numerical experiments, we perform a cross-validation. The Gaussian noise with the standard deviation 0.1 is added to data for all experiments.

The second synthetic dataset (the *Normal* dataset) contains points generated from the normal distributions with two expectations $(-2, -2)$ and $(2, 2)$. Anomalies are generated from the uniform distribution in interval $[-1, 1]$. Training and testing sets are depicted in Fig. 2.

First, we study the Circle dataset. F1-score measures obtained for ABIForest are shown

Table 2: F1-score of ABIForest consisting of $T = 150$ trees as a function of hyperparameters τ and ϵ for the Circle dataset by $\omega = 20$

ϵ	τ		
	0.5	0.6	0.7
0.0	0.276	0.973	0.236
0.25	0.2749	0.975	0.162
0.5	0.273	0.978	0.100
0.75	0.273	0.975	0.062
1.0	0.271	0.973	0.037

Table 3: F1-score of the original iForest as a function of the number T of trees and hyperparameter τ for the Circle dataset

τ	T				
	5	15	25	50	150
0.3	0.270	0.270	0.270	0.270	0.270
0.4	0.286	0.273	0.271	0.270	0.270
0.5	0.729	0.864	0.899	0.906	0.920
0.6	0.639	0.603	0.598	0.603	0.606

in Table 2 where the F1-score is presented as the function of hyperparameters ϵ and τ by the number of trees in the isolation forest $T = 150$. It is interesting to note that ABIForest is sensitive to changes of τ whereas ϵ does not significantly impact on results. For comparison purposes, F1-score measures of the original iForest as a function of the number T of trees and hyperparameter τ are shown in Table 3. It can be seen from Table 3 that the largest value of the F1-score is achieved by 150 trees in the forest and by $\tau = 0.5$. One can also see from Tables 2 and 3 that ABIForest provides results which outperform the same results of the original iForest.

Similar numerical experiments with the Normal dataset are presented in Tables 4 and 5. We can again see that ABIForest outperforms the iForest, namely the best value of the F1-score provided by the iForest is 0.252 whereas the best value of the F1-score for ABIForest is 0.413, and this result is obtained by $\omega = 20$.

Fig. 3 illustrates how the F1-score depends on hyperparameter τ for the Circle dataset. The corresponding functions are depicted for different contamination parameters ϵ and obtained for the case of $T = 150$ trees in the iForest. It can be seen from Fig. 3 that the largest value of the F1-score is achieved by $\omega = 20$ and $\epsilon = 0.5$. It can also be seen from the results in Fig. 3 that the F1-score significantly depends on hyperparameter ω especially for small values of ϵ . F1-score measures as functions of the contamination parameter ω for different numbers of trees in the iForest T for the Circle dataset obtained by hyperparameters $\gamma = 0.6$ and $\omega = 20$ are depicted in Fig. 4.

Fig. 5 illustrates comparison results between the iForest and ABIForest on the basis of

Table 4: F1-score of ABIForest consisting of $T = 150$ trees as a function of hyperparameters τ and ϵ for the Normal dataset by $\omega = 20$

ϵ	τ		
	0.5	0.6	0.7
0.0	0.099	0.410	0.0
0.25	0.147	0.410	0.162
0.5	0.177	0.413	0.0
0.75	0.176	0.412	0.0
1.0	0.178	0.408	0.0

Table 5: F1-score of the original iForest as a function of the number T of trees and hyperparameter τ for the Normal dataset

τ	T				
	5	15	25	50	150
0.3	0.082	0.082	0.082	0.082	0.082
0.4	0.088	0.083	0.083	0.082	0.082
0.5	0.220	0.248	0.249	0.250	0.252
0.6	0.191	0.141	0.091	0.040	0.021

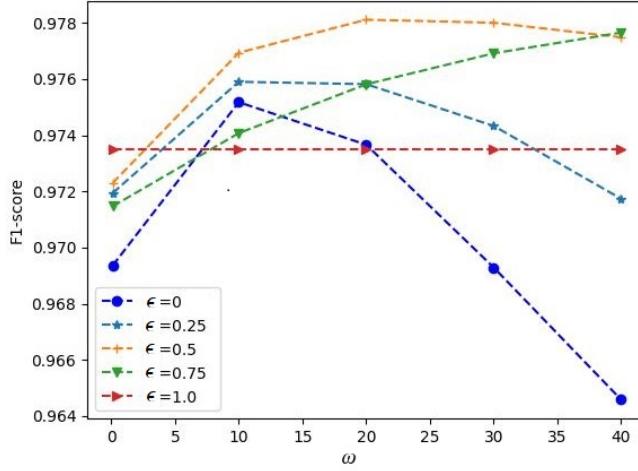


Figure 3: F1-score measures as functions of the softmax hyperparameter ω for different contamination parameters ϵ for the Circle dataset

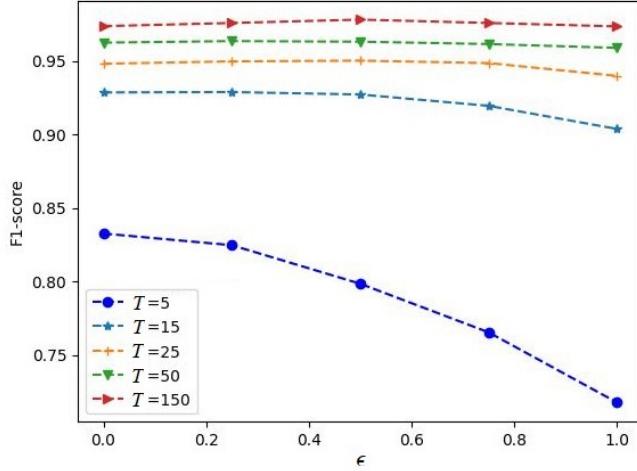


Figure 4: F1-score measures as functions of the contamination parameter ω for different numbers of trees in iForest T for the Circle dataset

the testing set which is depicted in the left picture of Fig. 5. Predictions obtained by the iForest consisting of 150 trees by $\tau = 0.5$ are depicted in the central picture. Predictions obtained by ABIForest by $\epsilon = 0.5$, $\tau = 0.6$ and $\omega = 0.1$ are shown in the right picture. One can see from Fig. 5 that some points in the central picture are incorrectly identified as anomalous ones whereas ABIForest is correctly classified them as normal instance. Fig. 5 should not be considered as the single realization which defines the F1-score. It is one of many cases corresponding to different generations of testing sets, therefore, the numbers of normal and anomalous instances can be different in each realization.

Similar dependencies for the Normal dataset are shown in Figs. 6 and 7. However, it follows from Fig. 6 that the largest values of the F1-score are achieved for $\epsilon = 0$. This implies that the main contribution into the attention weights is caused by the softmax operation. F1-score measures shown in Fig. 7 are obtained by hyperparameters $\gamma = 0.6$ and $\omega = 20$.

Comparison results between the iForest and ABIForest for the Normal dataset are shown in Fig. 8 where a realization of the testing set, predictions of the iForest and ABIForest are shown in the left, central, right pictures, respectively. Predictions are obtained by means of the iForest consisting of 150 trees by $\tau = 0.5$ and by ABIForest consisting of the same number of trees by $\epsilon = 0.5$, $\tau = 0.6$ and $\omega = 0.1$.

Another interesting question is how the prediction accuracy of ABIForest depends on the size of training data. The corresponding results for synthetic datasets are shown in Fig. 9 where the solid and dashed lines correspond to the F1-score of iForest and ABIForest, respectively. Numbers of trees in all experiments are taken $T = 150$. The same results in the numerical form are also given in Table 6. It can be seen from Fig. 9 for the Circle dataset that the F1-score of the iForest decreases with increase of the number of training data after

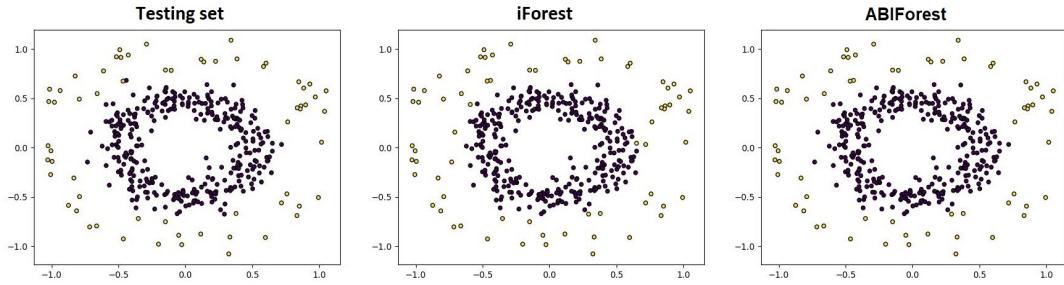


Figure 5: Comparison of the generated testing set for the Circle dataset (the left picture), predictions obtained by iForest (the central picture), predictions obtained by ABIForest (the right picture)

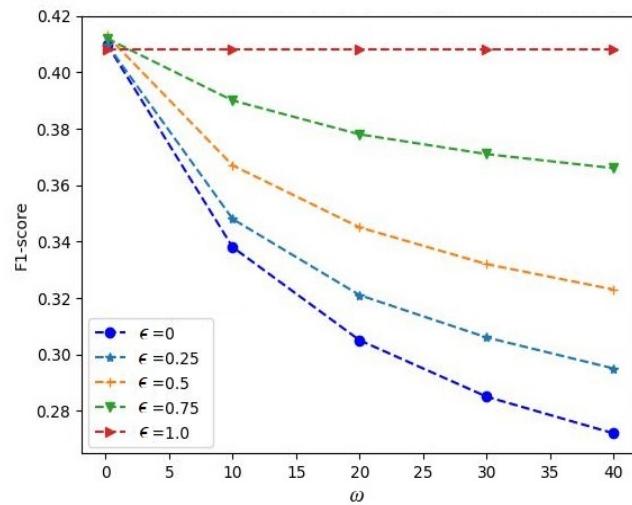


Figure 6: F1-score measures as functions of the softmax hyperparameter ω for different contamination parameters ϵ and for the Normal dataset

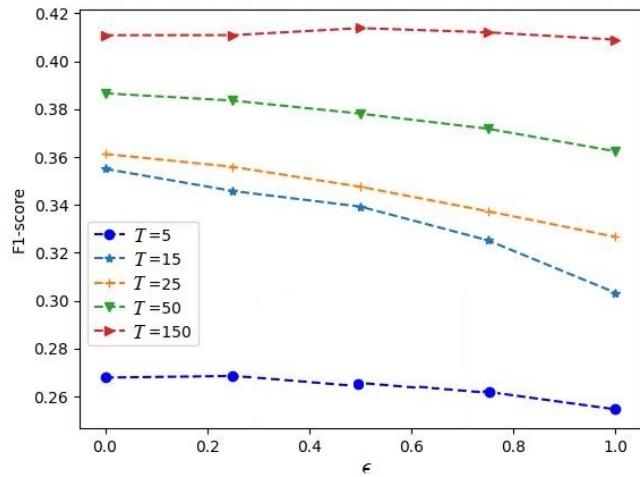


Figure 7: F1-score measures as functions of the contamination parameter ω for different numbers of trees in iForest T for the Circle dataset

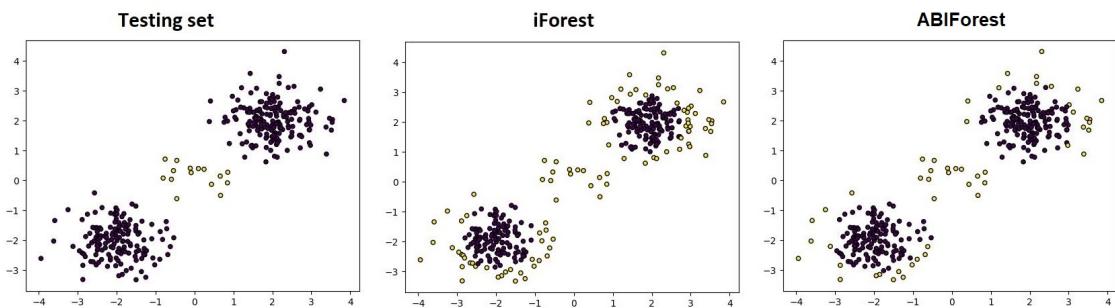


Figure 8: Comparison of the generated testing set for the Normal dataset (the left picture), predictions obtained by iForest (the central picture), predictions obtained by ABIForest (the right picture)

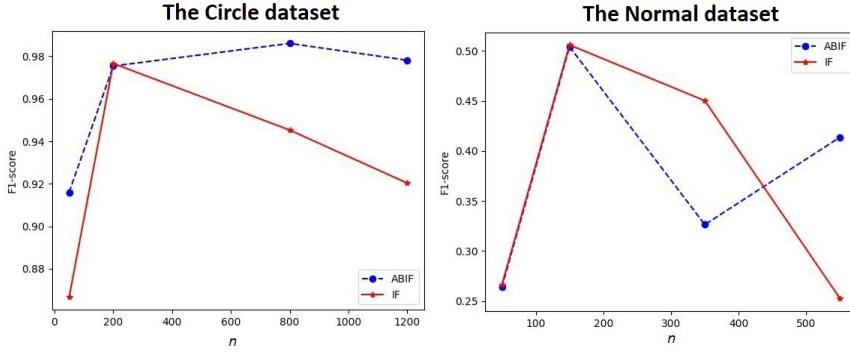


Figure 9: Illustration how the F1-score measures of iForest and ABIForest depend on numbers of training data for the Circle dataset (the left picture) and the Normal dataset (the right picture)

Table 6: F1-score measures of the original iForest and ABIForest as functions of the training data number n for the Circle and Normal datasets

The Circle dataset				
n	50	200	800	1200
iForest	0.867	0.977	0.945	0.920
ABIForest	0.916	0.975	0.986	0.978
The Normal dataset				
n	50	150	350	550
iForest	0.266	0.506	0.450	0.253
ABIForest	0.264	0.512	0.326	0.413

$n = 200$. This is because the number of trees ($T = 150$) is fixed and trees cannot be improved. This effect has been discussed in [17] where problems of swamping and masking were studied. Authors [17] considered the subsampling to overcome these problems. One can see from Fig. 9 that ABIForest copes with this difficulty. Another behavior of ABIForest can be observed for the Normal dataset which is characterized by two clusters of normal points. In this case, the F1-score decreases as n increases and then increases with n .

5.2 Real dataset

The first real dataset, used in numerical experiments and called the *Credit* dataset¹. According to the dataset description, it contains transactions made by credit cards in September 2013 by European cardholders with 492 frauds out of 284807 transactions. We use only 1500 normal instances and 400 anomalous ones, which are randomly selected from the whole Credit

¹<https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

Table 7: F1-score measures of ABIForest consisting of $T = 150$ trees for different real datasets by optimal values of τ , ϵ , ω and F1-score measures of iForest by optimal values of τ

Dataset	ABIForest			iForest		
	ϵ_{opt}	τ_{opt}	ω_{opt}	F1	τ_{opt}	F1
Credit	0.25	0.55	0.1	0.911	0.4	0.836
Ionosphere	0.0	0.4	0.1	0.693	0.45	0.684
Arrhythmia	1.0	0.45	—	0.481	0.4	0.479
Mulcross	0.0	0.6	0.1	0.507	0.5	0.516
Http	0.75	0.55	0.1	0.843	0.5	0.720
Pima	0.75	0.45	30	0.553	0.4	0.540

dataset. The second dataset, called the *Ionosphere* dataset², is a collection of radar returns from the ionosphere. The next dataset is called the *Arrhythmia* dataset³. The smallest classes with numbers 3, 4, 5, 7, 8, 9, 14, 15 are combined to form outliers in the Arrhythmia dataset. The *Mulcross* dataset⁴ is generated from a multi-variate normal distribution with two dense anomaly clusters. We use 1800 normal and 400 anomalous instances. The *Http* dataset⁵ is used in [17] for studying iForest. The *Pima* dataset⁶ aims to predict whether or not a patient has diabetes. Datasets Credit, Mulcross, Http are reduced to simplify experiments.

Numerical results are shown in Table 7. ABIForest is presented in Table 7 by hyperparameters ϵ , τ , ω and the F1-score. iForest is presented by hyperparameter τ and the corresponding F1-score. Hyperparameters leading to the largest F1-score are presented in Table 7. It can be seen from Table 7 that ABIForest provides outperforming results for five from six datasets. It is also interesting to point out that optimal values of hyperparameter ϵ for two datasets Ionosphere and Mulcross are equal to 0. This implies that attention weights are entirely determined by the softmax operation (see (17)). A contrary case is when $\epsilon_{opt} = 1$. In this case, the softmax operations as well as their parameter ω are not used, and the attention weights are entirely determined by parameters \mathbf{w} which can be regarded as weights of trees.

It is interesting to study how hyperparameter τ impacts on the performance of ABIForest and iForest. The corresponding dependencies are depicted in Figs. 10-12. The comparison results are obtained under condition of optimal values of ϵ and ω given in Table 7. One can see from Fig. 10 that τ differently impacts on performances of ABIForest and iForest for the Credit dataset whereas the corresponding dependencies scarcely differ for the Ionosphere dataset. This peculiarity is caused by the optimal values of the contamination parameter ϵ . It can be seen from Table 7 that $\tau_{opt} = 0$ for the Ionosphere dataset. This implies that the attention weights are determined only by the softmax operations which weakly impact on

²<https://www.kaggle.com/datasets/prashant111/ionosphere>

³<https://www.kaggle.com/code/medahmedkrichen/arrhythmia-classification>

⁴<https://github.com/dple/Datasets>

⁵<http://odds.cs.stonybrook.edu/http-kddcup99-dataset/>

⁶<https://github.com/dple/Datasets>

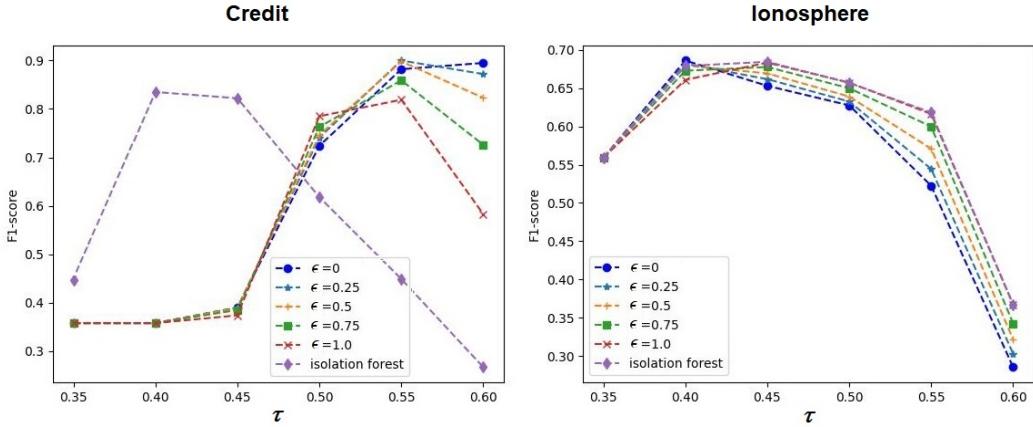


Figure 10: Comparison of iForest and ABIForest by different thresholds τ and by different contamination parameter ϵ for the Credit (the left picture) and Ionosphere (the right picture) datasets

the model performance and whose values are close to $1/T$. Moreover, the Ionosphere dataset is one of the smallest datasets with a large number of anomalous instances (see Table 1). Therefore, additional learnable parameters may lead to overfitting. This is a reason why the optimal hyperparameter ϵ does not impact on the model performance. It is also interesting to note that the optimal value of the contamination parameter for the Mullcross dataset is 0 (see Table 7). However, one can see quite different dependencies from the right picture of Fig. 11. This is caused by a large impact of the softmax operations whose values are far from $1/T$, and they provide results different from iForest.

Generally, one can see from Figs. 10-12 that models strongly depend on hyperparameters τ and ϵ . Most dependencies illustrate that there is an optimal value of τ for each case, which is close to 0.5 for iForest as well as for ABIForest. The same can be said about contamination parameter ϵ .

6 Concluding remark

A new modification of iForest using the attention mechanism has been proposed. Let us focus on advantages and disadvantages of the modification.

Advantages:

1. ABIForest is very simple from the computation point of view because, in contrast to the attention-based neural network, the attention weights in ABIForest are trained by solving the standard quadratic optimization problem. The modification avoids gradient-based algorithms to compute optimal learnable attention parameters.

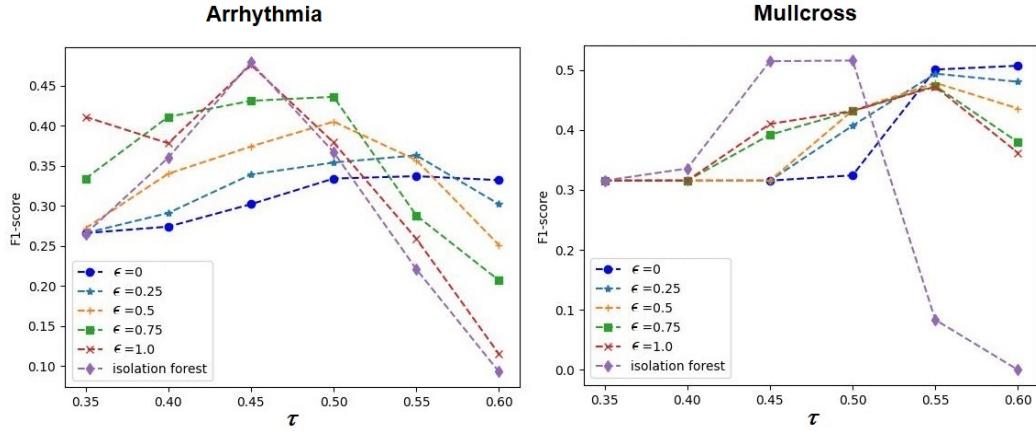


Figure 11: Comparison of iForest and ABIForest by different thresholds τ and by different contamination parameter ϵ for the Arrhythmia (the left picture) and Mullcross (the right picture) datasets

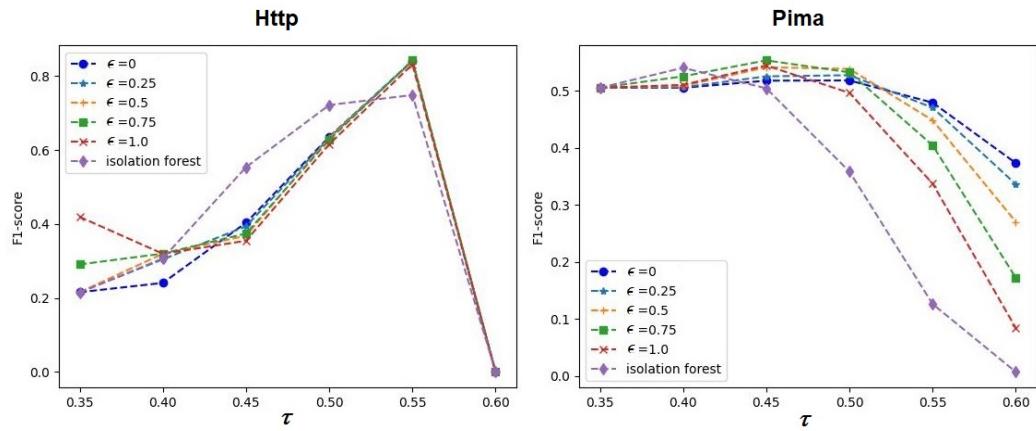


Figure 12: Comparison of iForest and ABIForest by different thresholds τ and by different contamination parameter ϵ for the Http (the left picture) and Pima (the right picture) datasets

2. ABIForest is a flexible model which can be simply modified. There are several components of ABIForest, which can be changed to improve the model performance. First, different kernels can be used instead of the Gaussian kernel considered above. Second, there are statistical models [59] different from the Huber’s ϵ -contamination model, which can also be used in ABIForest. Third, the attention weights can be associated with some subsets of trees, including intersecting subsets. In this case, the number of trainable parameters can be reduced to avoid overfitting. Fourth, paths in trees can be also attended, for example, by assigning attention weights to each branch in every path. Fifth, the multi-head attention can be applied to iForest in order to improve the model, for example, by changing hyperparameter ω of the softmax. Sixth, the distance between the instance \mathbf{x} and all instances, which fall in the same leaf as \mathbf{x} , can be defined differently. The above improvements can be regarded as directions for further research.
3. The attention model is trained after the forest building. This implies that we do not need to rebuild iForest to achieve a higher accuracy. Hyperparameters are tuned without rebuilding iForest. Moreover, we can apply various modifications and extensions of iForest and incorporate the attention mechanism in the same way as it is carried out with the original iForest.
4. ABIForest allows us to get interpretation answering the question why an instance is anomalous. This can be done by analyzing isolation trees with the largest attention weights.
5. ABIForest is perfectly deals with tabular data.
6. It follows from numerical experiments that ABIForest improves the iForest performance for many datasets.

Disadvantages:

1. The main disadvantage is that ABIForest has additionally three hyperparameters: contamination parameter ϵ , hyperparameter of the softmax operation ω , regularization hyperparameter λ . We do not include threshold τ which is also used in iForest. Additional hyperparameters lead to significant increase of the validation time.
2. Some additional time is required to solve the optimization problem (14).
3. In contrast to iForest, ABIForest is a supervised model. It requires to have labels of data (normal or anomalous) in order to determine a criteria of the optimization, in particular, to determine the optimization problem (14).

In spite of the disadvantages, ABIForest can be viewed as the first version for incorporating the attention mechanism into iForest which has illustrated outperforming results. The following modifications resolving the above disadvantages are interesting directions for further research.

References

- [1] R Chalapathy and S Chawla. Deep learning for anomaly detection: A survey. *ACM Computing Surveys*, 54(38):1–38, 2022.
- [2] A. Boukerche, Lining Zheng, and O. Alfandi. Outlier detection: Methods, models, and classification. *ACM Computing Surveys*, 53(3):1–37, Article 55 2020.
- [3] M. Braei and S. Wagner. Anomaly detection in univariate time-series: A survey on the state-of-the-art. arXiv:2004.00433, Apr 2020.
- [4] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(15):1–58, 2009.
- [5] W.S. Al Farizi, I. Hidayah, and M.N. Rizal. Isolation forest based anomaly detection: A systematic literature review. In *8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, pages 118–122. IEEE, 2021.
- [6] M. Fauss, A.M. Zoubir, and H.V. Poor. Minimax robust detection: Classic results and recent advances. *IEEE Transactions on Signal Processing*, 69:2252–2283, 2021.
- [7] Guansong Pang, Chunhua Shen, Longbing Cao, and A. Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(38):1–38, 2022.
- [8] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. arXiv:2110.11334v2, Aug 2022.
- [9] Guansong Pang, Longbing Cao, and Charu Aggarwal. Deep learning for anomaly detection: Challenges, methods, and opportunities. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1127–1130, 2021.
- [10] L. Ruff, J.R. Kauffmann, R.A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T.G. Dietterich, and K.R. Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.
- [11] M Hammad H Wang, MJ Bah. Progress in outlier detection techniques: A survey. *IEEE Access.*, 7:107964–108000, 2019.
- [12] C.C. Aggarwal. *An introduction to outlier analysis*, chapter Outlier analysis, pages 1–40. Springer, 2013.
- [13] C. Campbell and K.P. Bennett. A linear programming approach to novelty detection. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 395–401. MIT Press, 2001.
- [14] B. Scholkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

- [15] D. Tax and R. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, 2004.
- [16] Fei Tony Liu, Ming Ting Kai, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth IEEE international conference on data mining*, pages 413–422. IEEE, 2008.
- [17] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1–39, 2012.
- [18] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath. An attentive survey of attention models. arXiv:1904.02874, Apr 2019.
- [19] A.S. Correia and E.L. Colombini. Attention, please! A survey of neural attention models in deep learning. arXiv:2103.16775, Mar 2021.
- [20] A.S. Correia and E.L. Colombini. Neural attention models in deep learning: Survey and taxonomy. arXiv:2112.05909, Dec 2021.
- [21] T. Lin, Y. Wang, X. Liu, and X. Qiu. A survey of transformers. arXiv:2106.04554, Jul 2021.
- [22] Z. Niu, G. Zhong, and H. Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.
- [23] L.V. Utkin and A.V. Konstantinov. Attention-based random forest and contamination model. *Neural Networks*, 154:346–359, 2022.
- [24] E.A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.
- [25] G.S. Watson. Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.
- [26] P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.
- [27] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473, Sep 2014.
- [28] S.P. Sawant and S. Singh. Understanding attention: In minds and machines. arXiv:2012.02659, Dec 2020.
- [29] Kelvin Xu, Jimmy Ba, R. Kiros, Kyunghyun Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
- [30] Thang Luong, Hieu Pham, and C.D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015.

- [31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- [32] Zichao Yang, Diyi Yang, C. Dyer, Xiaodong He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.
- [33] F. Liu, X. Huang, Y. Chen, and J.A. Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. arXiv:2004.11154v5, Jul 2021.
- [34] L.V. Utkin and A.V. Konstantinov. Attention and self-attention in random forests. arXiv:2207.04293, Jul 2022.
- [35] L.V. Utkin and A.V. Konstantinov. Random survival forests incorporated by the nadaraya-watson regression. *Informatics and Automation*, 21(5):851–880, 2022.
- [36] A.V. Konstantinov, L.V. Utkin, and S.R. Kirpichenko. AGBoost: Attention-based modification of gradient boosting machine. In *31st Conference of Open Innovations Association (FRUCT)*, pages 96–101. IEEE, 2022.
- [37] A. Kundu, A. Sahu, E. Serpedin, and K. Davis. A3d: Attention-based auto-encoder anomaly detector for false data injection attacks. *Electric Power Systems Research*, 189(106795), 2020.
- [38] H. Takimoto, J. Seki, S.F. Situju, and A. Kanagawa. Anomaly detection using siamese network with attention mechanism for few-shot learning. *Applied Artificial Intelligence*, 36(1):2930–2946, 2022.
- [39] Xiaoming Lei, Ye Xia, Ao Wang, Xudong Jian, Huaqiang Zhong, and Limin Sun. Mutual information based anomaly detection of monitoring data with attention mechanism and residual learning. *Mechanical Systems and Signal Processing Volume 182, 1 January 2023, 109607*, 182(109607):1–18, 2023.
- [40] Yangming Yu, Zhiyong Zha, Bo Jin, Geng Wu, and Chenxi Dong. Graph-based anomaly detection via attention mechanism. In *International Conference on Intelligent Computing*, pages 401–411. Springer, Cham, 2022.
- [41] N. Madan, N.-C. Ristea, R.T. Ionescu, K. Nasrollahi, F.S. Khan, T.B. Moeslund, and M. Shah. Self-supervised masked convolutional transformer block for anomaly detection. arXiv:2209.12148, Sep 2022.
- [42] N.-C. Ristea, N. Madan, R.T. Ionescu, K. Nasrollahi, F.S. Khan, T.B. Moeslund, and M. Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13576–13586, 2022.

- [43] Chaoqin Huang, Qinwei Xu, Yanfeng Wang, Yu Wang, and Ya Zhang. Self-supervised masking for unsupervised anomaly detection and localization. arXiv:2205.06568, May 2022.
- [44] Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Multivariate time-series anomaly detection via graph attention network. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 841–850. IEEE, 2020.
- [45] Junfeng Wang, YanJia, Dongbo Wang, Wenjing Xiao, and Zhenfei Wang. Weighted iforest and siamese gru on small sample anomaly detection in healthcare. *Computer Methods and Programs in Biomedicine*, 218:106706, 2022.
- [46] S. Hariri, M.C. Kind, and R.J. Brunner. Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1479–1489, 2021.
- [47] S. Buschjager, P.-J. Honysz, and K. Morik. Generalized isolation forest: Some theory and more applications extended abstract. In *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 793–794 7. IEEE, 2020.
- [48] J. Lesouple, C. Baudoin, M. Spigai, and J.-Y. Tourneret. Generalized isolation forest for anomaly detection. *Pattern Recognition*, 149:109–119, 2021.
- [49] P. Karczmarek, A. Kiersztyn, W. Pedrycz, and E. Al. K-means-based isolation forest. *Knowledge-Based Systems*, 195(105659):1–15, 2020.
- [50] P. Karczmarek, A. Kiersztyn, and W. Pedrycz. Fuzzy set-based isolation forest. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE, 2020.
- [51] M. Tokovarov and P. Karczmarek. A probabilistic generalization of isolation forest. *Information Sciences*, 584(433–449), 2022.
- [52] Changgen Li, Liang Guo, Hongli Gao, and Yi Li. Similarity-measured isolation forest: anomaly detection method for machine monitoring data. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021.
- [53] Shutao Li, Kunzhong Zhang, Puhong Duan, and Xudong Kang. Hyperspectral anomaly detection with kernel isolation forest. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):319–329, 2020.
- [54] Zhen Liu, Xin Liu, Jin Ma, and Hui Gao. An optimized computational framework for isolation forest. *Mathematical Problems in Engineering*, 2018(Article ID 2318763):1–14, 2018.
- [55] G. Staerman, P. Mozharovskyi, S. Clemenccon, and F. d’Alche Buc. Functional isolation forest. *Proceedings of The Eleventh Asian Conference on Machine Learning*, PMLR 101:332–347, 2019.

- [56] Hongzuo Xu, Guansong Pang, Yijie Wang, and Yongjun Wang. Deep isolation forest for anomaly detection. arXiv:2206.06602, Jun 2022.
- [57] A. Zhang, Z.C. Lipton, M. Li, and A.J. Smola. Dive into deep learning. *arXiv:2106.11342*, Jun 2021.
- [58] T. Luong, H. Pham, and C.D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. The Association for Computational Linguistics, 2015.
- [59] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.