Abstract

We consider the problem of identifying patterns in a data set that exhibit anomalous behavior, often referred to as anomaly detection. In most anomaly detection algorithms, the dissimilarity between data samples is calculated by a single criterion, such as Euclidean distance. However, in many cases there may not exist a single dissimilarity measure that captures all possible anomalous patterns. In such a case, multiple criteria can be defined, and one can test for anomalies by scalarizing the multiple criteria using a linear combination of them. If the importance of the different criteria are not known in advance, the algorithm may need to be executed multiple times with different choices of weights in the linear combination. In this paper, we introduce a novel non-parametric multi-criteria anomaly detection method using Pareto depth analysis (PDA). PDA uses the concept of Pareto optimality to detect anomalies under multiple criteria without having to run an algorithm multiple times with different choices of weights. The proposed PDA approach scales linearly in the number of criteria and is provably better than linear combinations of the criteria.

1 Introduction

Anomaly detection is an important problem that has been studied in a variety of areas and used in diverse applications including intrusion detection, fraud detection, and image processing [1][2]. Many methods for anomaly detection have been developed using both parametric and non-parametric approaches. Non-parametric approaches typically involve the calculation of dissimilarities between data samples. For complex high-dimensional data, multiple dissimilarity measures corresponding to different criteria may be required to detect certain types of anomalies. For example, consider the problem of detecting anomalous object trajectories in video sequences. Multiple criteria, such as dissimilarity in object speeds or trajectory shapes, can be used to detect a greater range of anomalies than any single criterion. In order to perform anomaly detection using these multiple criteria, one could first combine the dissimilarities using a linear combination. However, in many applications, the importance of the criteria are not known in advance. It is difficult to determine how much weight to assign to each dissimilarity measure, so one may have to choose multiple weights using, for example, a grid search. Furthermore, when the weights are changed, the anomaly detection algorithm needs to be re-executed using the new weights.

In this paper we propose a novel non-parametric multi-criteria anomaly detection approach using Pareto depth analysis (PDA). PDA uses the concept of Pareto optimality to detect anomalies without having to choose weights for different criteria. Pareto optimality is the typical method for defining optimality when there may be multiple conflicting criteria for comparing items. An item is said to be Pareto-optimal if there does not exist another item that is better or equal in all of the criteria. An item that is Pareto-optimal is optimal in the usual sense under some combination, not necessarily linear, of the criteria. Hence, PDA is able to detect anomalies under multiple combinations of the criteria without explicitly forming these combinations.
The PDA approach involves creating dyads corresponding to dissimilarities between pairs of data samples under all of the dissimilarity measures. Sets of Pareto-optimal dyads, called Pareto fronts, are then computed. The first Pareto front (depth one) is the set of non-dominated dyads. The second Pareto front (depth two) is obtained by removing these non-dominated dyads, i.e. peeling off the first front, and recomputing the first Pareto front of those remaining. This process continues until no dyads remain. In this way, each dyad is assigned to a Pareto front at some depth (see Fig. 1 for illustration). Nominal and anomalous samples are located near different Pareto front depths; thus computing the front depths of the dyads corresponding to a test sample can discriminate between nominal and anomalous samples. The proposed PDA approach scales linearly in the number of criteria, which is a significant improvement compared to selecting multiple weights via a grid search, which scales exponentially in the number of criteria. Under assumptions that the multi-criteria dyads can be modeled as realizations from a smooth $K$-dimensional density we provide a mathematical analysis of the behavior of the first Pareto front. This analysis shows in a precise sense that PDA can outperform a test that uses a linear combination of the criteria. Furthermore, this theoretical prediction is experimentally validated by comparing PDA to several state-of-the-art anomaly detection algorithms in two experiments involving both synthetic and real data sets.

The rest of this paper is organized as follows. We discuss related work in Section 2. In Section 3 we provide an introduction to Pareto fronts and present a theoretical analysis of the properties of the first Pareto front. Section 4 relates Pareto fronts to the multi-criteria anomaly detection problem, which leads to the PDA anomaly detection algorithm. Finally we present two experiments in Section 5 to evaluate the performance of PDA.

2 Related work

Several machine learning methods utilizing Pareto optimality have previously been proposed; an overview can be found in [3]. These methods typically formulate machine learning problems as multi-objective optimization problems where finding even the first Pareto front is quite difficult. These methods differ from our use of Pareto optimality because we consider multiple Pareto fronts created from a finite set of items, so we do not need to employ sophisticated methods in order to find these fronts.

Hero and Fleury [4] introduced a method for gene ranking using Pareto fronts that is related to our approach. The method ranks genes, in order of interest to a biologist, by creating Pareto fronts of the data samples, i.e. the genes. In this paper, we consider Pareto fronts of dyads, which correspond to dissimilarities between pairs of data samples rather than the samples themselves, and use the distribution of dyads in Pareto fronts to perform multi-criteria anomaly detection rather than ranking.

Another related area is multi-view learning [5, 6], which involves learning from data represented by multiple sets of features, commonly referred to as “views”. In such case, training in one view helps to
For convenience, we say that a Pareto front $F$ is obtained by selecting items that are not strictly dominated by any of the remaining items, which are members of the set $S \setminus F$. A more powerful approach involves finding the set of Pareto-optimal items. An item $x$ is said to strictly dominate another item $x^{*}$ if $x$ is no greater than $x^{*}$ in each criterion and $x$ is less than $x^{*}$ in at least one criterion. This relation can be written as $x \succ x^{*}$ if $f_i(x) \leq f_i(x^{*})$ for each $i$ and $f_i(x) < f_i(x^{*})$ for some $i$. The set of Pareto-optimal items, called the Pareto front, is the set of items in $S$ that are not strictly dominated by another item in $S$. It contains all of the minimizers that are found using linear combinations, but also includes other items that cannot be found by linear combinations. Denote the Pareto front by $F_1$, which we call the first Pareto front. The second Pareto front can be constructed by finding items that are not strictly dominated by any of the remaining items, which are members of the set $S \setminus F_1$. More generally, define the $i$th Pareto front by

$$F_i = \text{Pareto front of the set } S \setminus \left( \bigcup_{j=1}^{i-1} F_j \right).$$

For convenience, we say that a Pareto front $F_i$ is deeper than $F_j$ if $i > j$.

### 3.1 Mathematical properties of Pareto fronts

The distribution of the number of points on the first Pareto front was first studied by Barndorff-Nielsen and Sobel in their seminal work [17]. The problem has garnered much attention since; for a
survey of recent results see [18]. We will be concerned here with properties of the first Pareto front that are relevant to the PDA anomaly detection algorithm and thus have not yet been considered in the literature. Let \( Y_1, \ldots, Y_n \) be independent and identically distributed (i.i.d.) on \( \mathbb{R}^d \) with density function \( f : \mathbb{R}^d \to \mathbb{R} \). For a measurable set \( A \subseteq \mathbb{R}^d \), we denote by \( F_A \) the points on the first Pareto front of \( Y_1, \ldots, Y_n \) that belong to \( A \). For simplicity, we will denote \( F_1 \) by \( F \) and use \( |F| \) for the cardinality of \( F \). In the general Pareto framework, the points \( Y_1, \ldots, Y_n \) are the images in \( \mathbb{R}^d \) of \( n \) feasible solutions to some optimization problem under a vector of objective functions of length \( d \).

In the context of this paper, each point \( Y_t \) corresponds to a dyad \( D_{ij} \), which we define in Section 4, and \( d = K \) is the number of criteria. A common approach in multi-objective optimization is linear scalarization [16], which constructs a new single criterion as a convex combination of the \( d \) criteria. Let \( \sum_{i=1}^{d} \alpha_i x_i \) be a linear function.

We will be concerned here with properties of the first Pareto front that are relevant to the PDA anomaly detection algorithm and thus have not yet been considered in the literature. Let \( Y_1, \ldots, Y_n \) be independent and identically distributed (i.i.d.) on \( \mathbb{R}^d \) with density function \( f : \mathbb{R}^d \to \mathbb{R} \). For a measurable set \( A \subseteq \mathbb{R}^d \), we denote by \( F_A \) the points on the first Pareto front of \( Y_1, \ldots, Y_n \) that belong to \( A \). For simplicity, we will denote \( F_1 \) by \( F \) and use \( |F| \) for the cardinality of \( F \). In the general Pareto framework, the points \( Y_1, \ldots, Y_n \) are the images in \( \mathbb{R}^d \) of \( n \) feasible solutions to some optimization problem under a vector of objective functions of length \( d \).

Theorem 1. Let \( f \in C^1(\Omega) \) with inf_\Omega f > 0. Let \( T \subset \partial \Omega \) be open and connected such that
\[
\inf_{z \in T} \min(\nu_1(z), \ldots, \nu_d(z)) \geq \delta > 0, \quad \{y \in \Omega : y \preceq x\} = \{x\}, \quad \text{for} \quad x \in T.
\]

Then for \( h > 0 \) sufficiently small, we have
\[
E|F_{T_h}| = \gamma n^{d-1} + \delta^{-d-1} O \left( n^{d-2} \right) \quad \text{as} \quad n \to \infty,
\]
where \( \gamma = d^{-1} (d!)^{\frac{1}{2}} \Gamma(d^{-1}) \int_T f(z)^{d-1} \nu_1(z) \cdots \nu_d(z) \frac{dz}{\nu_1(z) \cdots \nu_d(z)} \left( \nu_2(z) \cdots \nu_d(z) \right)^{d-1} \).

The proof of Theorem 1 is postponed to Section 1 of the supplementary material. Theorem 1 shows asymptotically how many Pareto points are contributed on average by the segment \( T \subset \partial \Omega \). The number of points contributed depends only on the geometry of \( \partial \Omega \) through the direction of its normal vector \( \nu \) and is otherwise independent of the convexity of \( \partial \Omega \). Hence, by using Pareto methods, we will identify significantly more Pareto-optimal points than linear scalarization when the geometry of \( \partial \Omega \) includes non-convex regions. For example, if \( T \subset \partial \Omega \) is non-convex (see left panel of Figure 2) and satisfies the hypotheses of Theorem 1, then for large enough \( n \), all Pareto points in a neighborhood of \( T \) will be unattainable by scalarization. Quantitatively, if \( f \geq C \) on \( T \), then
\[
E|F \setminus L| \geq \gamma n^{d-1} + \delta^{-d-1} O(n^{d-2}), \quad \text{as} \quad n \to \infty,
\]
where \( \gamma \geq d^{-1} (d!)^{\frac{1}{2}} \Gamma(d^{-1}) |T| \delta C^{d-1} \) and \( |T| \) is the \( d-1 \) dimensional Hausdorff measure of \( T \). It has recently come to our attention that Theorem 1 appears in a more general form in an unpublished manuscript of Baryshnikov and Yukich [19].

We now study non-convexities in the Pareto front which occur due to inherent randomness in the samples. We show that, even in the case where \( \Omega \) is convex, there are still numerous small-scale non-convexities in the Pareto front that can only be detected by Pareto methods. We illustrate this in the case of the Pareto box problem for \( d = 2 \).
Let \( Y_1, \ldots, Y_n \) be independent and uniformly distributed on \([0, 1]^2\). Then
\[
\frac{1}{2} \ln n + O(1) \leq E|\mathcal{L}| \leq \frac{5}{6} \ln n + O(1), \quad \text{as } n \to \infty.
\]

The proof of Theorem 2 is also postponed to Section 1 of the supplementary material. A proof that \( E|\mathcal{F}| = \ln n + O(1) \) as \( n \to \infty \) can be found in [17]. Hence Theorem 2 shows that, asymptotically and in expectation, only between \( \frac{1}{2} \) and \( \frac{5}{6} \) of the Pareto-optimal points can be obtained by linear scalarization in the Pareto box problem. Experimentally, we have observed that the true fraction of points is close to 0.7. This means that at least \( \frac{1}{2} \) (and likely more) of the Pareto points can only be obtained via Pareto methods even when \( \Omega \) is convex. Figure 2 gives an example of the sets \( \mathcal{F} \) and \( \mathcal{L} \) from the two theorems.

4 Multi-criteria anomaly detection

Assume that a training set \( X_N = \{X_1, \ldots, X_N\} \) of nominal data samples is available. Given a test sample \( X \), the objective of anomaly detection is to declare \( X \) to be an anomaly if \( X \) is significantly different from samples in \( X_N \). Suppose that \( K > 1 \) different evaluation criteria are given. Each criterion is associated with a measure for computing dissimilarities. Denote the dissimilarity between \( X_i \) and \( X_j \) computed using the measure corresponding to the \( l \)th criterion by \( d_l(i, j) \).

We define a dyad by \( D_{ij} = [d_1(i, j), \ldots, d_K(i, j)]^T \in \mathbb{R}_+^K, i \in \{1, \ldots, N\}, j \in \{1, \ldots, N\} \setminus i \). Each dyad \( D_{ij} \) corresponds to a connection between samples \( X_i \) and \( X_j \). Therefore, there are in total \( \binom{N}{2} \) different dyads. For convenience, denote the set of all dyads by \( \mathcal{D} \) and the space of all dyads \( \mathbb{R}_+^K \) by \( \mathcal{D} \). By the definition of strict dominance in Section 3, a dyad \( D_{ij} \) strictly dominates another dyad \( D_{i'j'} \) if \( d_l(i, j) < d_l(i', j') \) for all \( l \in \{1, \ldots, K\} \) and \( d_l(i, j) \leq d_l(i', j') \) for some \( l \). The first Pareto front \( \mathcal{F}_1 \) corresponds to the set of dyads from \( \mathcal{D} \) that are not strictly dominated by any other dyads from \( \mathcal{D} \). The second Pareto front \( \mathcal{F}_2 \) corresponds to the set of dyads from \( \mathcal{D} \setminus \mathcal{F}_1 \), and so on, as defined in Section 3. Recall that we refer to \( \mathcal{F}_1 \) as a deeper front than \( \mathcal{F}_j \) if \( i > j \).

4.1 Pareto fronts of dyads

For each sample \( X_n \), there are \( N - 1 \) dyads corresponding to its connections with the other \( N - 1 \) samples. Define the set of \( N - 1 \) dyads associated with \( X_n \) by \( \mathcal{D}^n \). If most dyads in \( \mathcal{D}^n \) are located at shallow Pareto fronts, then the dissimilarities between \( X_n \) and the other \( N - 1 \) samples are small under some combination of the criteria. Thus, \( X_n \) is likely to be a nominal sample. This is the basic idea of the proposed multi-criteria anomaly detection method using PDA.

We construct Pareto fronts \( \mathcal{F}_1, \ldots, \mathcal{F}_M \) of the dyads from the training set, where the total number of fronts \( M \) is the required number of fronts such that each dyad is a member of a front. When a test sample \( X \) is obtained, we create new dyads corresponding to connections between \( X \) and training samples, as illustrated in Figure 1. Similar to many other anomaly detection methods, we connect each test sample to its \( k \) nearest neighbors. \( k \) could be different for each criterion, so we denote \( k_i \) as the choice of \( k \) for criterion \( i \). We create \( s = \sum_{i=1}^{K} k_i \) new dyads, which we denote by the set...
Testing phase:

1. Calculate pairwise dissimilarities \( d_l(i,j) \) between all training samples \( X_i \) and \( X_j \)
2. Create k-NNG using \( d_l(i,j) \)’s
3. Create dyads \( D_{ij} = [d_1(i,j), \ldots, d_K(i,j)] \) for all training samples
4. Construct Pareto fronts on set of all dyads until each dyad is in a front

PDA anomaly detection algorithm.

Training phase:

1: \( \text{for } l = 1 \rightarrow K \text{ do} \)
2: \( \text{Calculate dissimilarities between test sample } X \text{ and all training samples in criterion } l \)
3: \( \text{Create } k \text{-NNG using } d_l(i,j) \text{'s} \)
4: \( \text{Create } s \text{ new dyads } D_{\text{new}} \text{ between } X \text{ and training samples in } nb \)
5: \( \text{for } i = 1 \rightarrow s \text{ do} \)
6: \( \text{Calculate depth } e_i \) of \( D_{\text{new}} \)
7: \( \text{Declare } X \text{ an anomaly if } v(X) = (1/s) \sum_{i=1}^{s} e_i > \sigma \)

Anomaly detection using depths of dyads

In k-NN based anomaly detection algorithms such as those mentioned in Section 2, the anomaly score is a function of the \( k \) nearest neighbors to a test sample. With multiple criteria, one could define an anomaly score by scalarization. From the probabilistic properties of Pareto fronts discussed in Section 3.1, we know that Pareto methods identify more Pareto-optimal points than linear scalarization methods and significantly more Pareto-optimal points than a single weight for scalarization.

This motivates us to develop a multi-criteria anomaly score using Pareto fronts. We start with the observation from Figure 1 that dyads corresponding to a nominal test sample are typically located near shallow fronts than dyads corresponding to an anomalous test sample. Each test sample is associated with \( s \) new dyads, where the \( i \)th dyad \( D_{\text{new}} \) has depth \( e_i \). For each test sample \( X \), we define the anomaly score \( v(X) \) to be the mean of the \( e_i \)’s, which corresponds to the average depth of the \( s \) dyads associated with \( X \). Thus the anomaly score can be easily computed and compared to the decision threshold \( \sigma \) using the test

\[
v(X) = \frac{1}{s} \sum_{i=1}^{s} e_i \begin{array}{c} H_3 \end{array} \geq \sigma.
\]

Algorithm 1 PDA anomaly detection algorithm.

**Training phase:**

1. **for** \( l = 1 \rightarrow K \) **do**
2. Calculate pairwise dissimilarities \( d_l(i,j) \) between all training samples \( X_i \) and \( X_j \)
3. Create \( k \)-NNG using \( d_l(i,j) \)’s
4. Create dyads \( D_{ij} = [d_1(i,j), \ldots, d_K(i,j)] \) for all training samples
5. Construct Pareto fronts on set of all dyads until each dyad is in a front

**Testing phase:**

1. \( nb \leftarrow \emptyset \) {empty list}
2. **for** \( l = 1 \rightarrow K \) **do**
3. Calculate dissimilarities between test sample \( X \) and all training samples in criterion \( l \)
4. \( nb_l \leftarrow k_l \) nearest neighbors of \( X \)
5. \( nb \leftarrow [nb, nb_l] \) {append neighbors to list}
6. Create \( s \) new dyads \( D_{\text{new}} \) between \( X \) and training samples in \( nb \)
7. **for** \( i = 1 \rightarrow s \) **do**
8. Calculate depth \( e_i \) of \( D_{\text{new}} \)
9. Declare \( X \) an anomaly if \( v(X) = (1/s) \sum_{i=1}^{s} e_i > \sigma \)

**Algorithm 1** PDA anomaly detection algorithm.

*Training phase:*

1. **for** \( l = 1 \rightarrow K \) **do**
2. Calculate pairwise dissimilarities \( d_l(i,j) \) between all training samples \( X_i \) and \( X_j \)
3. Create \( k \)-NNG using \( d_l(i,j) \)’s
4. Create dyads \( D_{ij} = [d_1(i,j), \ldots, d_K(i,j)] \) for all training samples
5. Construct Pareto fronts on set of all dyads until each dyad is in a front

**Testing phase:**

1. \( nb \leftarrow \emptyset \) {empty list}
2. **for** \( l = 1 \rightarrow K \) **do**
3. Calculate dissimilarities between test sample \( X \) and all training samples in criterion \( l \)
4. \( nb_l \leftarrow k_l \) nearest neighbors of \( X \)
5. \( nb \leftarrow [nb, nb_l] \) {append neighbors to list}
6. Create \( s \) new dyads \( D_{\text{new}} \) between \( X \) and training samples in \( nb \)
7. **for** \( i = 1 \rightarrow s \) **do**
8. Calculate depth \( e_i \) of \( D_{\text{new}} \)
9. Declare \( X \) an anomaly if \( v(X) = (1/s) \sum_{i=1}^{s} e_i > \sigma \)

\( D_{\text{new}} = \{ D_{1\text{new}}, D_{2\text{new}}, \ldots, D_{s\text{new}} \} \), corresponding to the connections between \( X \) and the union of the \( k_i \) nearest neighbors in each criterion \( i \). In other words, we create a dyad between \( X \) and \( X_j \) if \( X_j \) is among the \( k_i \) nearest neighbors of \( X \) in any criterion \( i \). We say that \( D_{\text{new}} \) is below a front \( F_i \) if \( D_{\text{new}} \succ D_i \) for some \( D_i \in F_i \), i.e. \( D_{\text{new}} \) strictly dominates at least a single dyad in \( F_i \). Define the depth of \( D_{\text{new}} \) by

\[
e_i = \min \{ l \mid D_{\text{new}} \text{ is below } F_i \}.
\]

Therefore if \( e_i \) is large, then \( D_{\text{new}} \) will be near deep fronts, and the distance between \( X \) and the corresponding training sample is large under all combinations of the \( K \) criteria. If \( e_i \) is small, then \( D_{\text{new}} \) will be near shallow fronts, so the distance between \( X \) and the corresponding training sample is small under some combination of the \( K \) criteria.

**4.2 Anomaly detection using depths of dyads**

If a training sample is one of the \( k_i \) nearest neighbors in multiple criteria, then multiple copies of the dyad corresponding to the connection between the test sample and the training sample are created.

Theorems 1 and 2 require i.i.d. samples, but dyads are not independent. However, there are \( O(N^2) \) dyads, and each dyad is only dependent on \( O(N) \) other dyads. This suggests that the theorems should also hold for the non-i.i.d. dyads as well, and it is supported by experimental results presented in Section 2 of the supplementary material.

Pseudocode for the PDA anomaly detector is shown in Algorithm 1. In Section 3 of the supplementary material we provide details of the implementation as well as an analysis of the time complexity.
Table 1: AUC comparison of different methods for both experiments. Best AUC is shown in **bold**. PDA does not require selecting weights so it has a single AUC. The median and best AUCs (over all choices of weights selected by grid search) are shown for the other four methods. *PDA outperforms all of the other methods, even for the best weights, which are not known in advance.*

<table>
<thead>
<tr>
<th>Method</th>
<th>Median AUC</th>
<th>Best AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDA</td>
<td>0.948 ± 0.002</td>
<td>0.919 ± 0.003</td>
</tr>
<tr>
<td>k-NN</td>
<td>0.848 ± 0.004</td>
<td>0.919 ± 0.003</td>
</tr>
<tr>
<td>k-NN sum</td>
<td>0.854 ± 0.003</td>
<td>0.916 ± 0.003</td>
</tr>
<tr>
<td>k-LPE</td>
<td>0.847 ± 0.004</td>
<td>0.919 ± 0.003</td>
</tr>
<tr>
<td>LOF</td>
<td>0.845 ± 0.003</td>
<td>0.932 ± 0.003</td>
</tr>
</tbody>
</table>

and a heuristic for choosing the \( k \)'s that performs well in practice. Both the training time and the time required to test a new sample using PDA are *linear* in the number of criteria \( K \). To handle multiple criteria, other anomaly detection methods, such as the ones mentioned in Section 2, need to be re-executed multiple times using different (non-negative) linear combinations of the \( K \) criteria. If a grid search is used for selection of the weights in the linear combination, then the required computation time would be exponential in \( K \). Such an approach presents a computational problem unless \( K \) is very small. Since PDA scales *linearly* with \( K \), it does not encounter this problem.

5 Experiments

We compare the PDA method with four other nearest neighbor-based single-criterion anomaly detection algorithms mentioned in Section 2. For these methods, we use linear combinations of the criteria with different weights selected by grid search to compare performance with PDA.

5.1 Simulated data with four criteria

First we present an experiment on a simulated data set. The nominal distribution is given by the uniform distribution on the hypercube \([0, 1]^4\). The anomalous samples are located just outside of this hypercube. There are four classes of anomalous distributions. Each class differs from the nominal distribution in one of the four dimensions; the distribution in the anomalous dimension is uniform on \([1, 1.1]\). We draw 300 training samples from the nominal distribution followed by 100 test samples from a mixture of the nominal and anomalous distributions with a 0.05 probability of selecting any particular anomalous distribution. The four criteria for this experiment correspond to the squared differences in each dimension. If the criteria are combined using linear combinations, the combined dissimilarity measure reduces to weighted squared Euclidean distance.

The different methods are evaluated using the receiver operating characteristic (ROC) curve and the area under the curve (AUC). The mean AUCs (with standard errors) over 100 simulation runs are shown in Table 1a. A grid of six points between 0 and 1 in each criterion, corresponding to \(6^4 = 1296\) different sets of weights, is used to select linear combinations for the single-criterion methods. Note that PDA is the best performer, outperforming even the best linear combination.

5.2 Pedestrian trajectories

We now present an experiment on a real data set that contains thousands of pedestrians’ trajectories in an open area monitored by a video camera [20]. Each trajectory is approximated by a cubic spline curve with seven control points [21]. We represent a trajectory with \( l \) time samples by

\[
T = \begin{bmatrix}
  x_1 & x_2 & \cdots & x_l \\
  y_1 & y_2 & \cdots & y_l
\end{bmatrix},
\]

where \([x_t, y_t]\) denote a pedestrian’s position at time step \( t \).
Figure 3: Left: ROC curves for PDA and attainable region for k-LPE over 100 choices of weights. PDA outperforms k-LPE even under the best choice of weights. Right: A subset of the dyads for the training samples along with the first 100 Pareto fronts. The fronts are highly non-convex, partially explaining the superior performance of PDA.

We use two criteria for computing the dissimilarity between trajectories. The first criterion is to compute the dissimilarity in walking speed. We compute the instantaneous speed at all time steps along each trajectory by finite differencing, i.e. the speed of trajectory $T$ at time step $t$ is given by $\sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}$. A histogram of speeds for each trajectory is obtained in this manner. We take the dissimilarity between two trajectories to be the squared Euclidean distance between their speed histograms. The second criterion is to compute the dissimilarity in shape. For each trajectory, we select 100 points, uniformly positioned along the trajectory. The dissimilarity between two trajectories $T$ and $T'$ is then given by the sum of squared Euclidean distances between the positions of $T$ and $T'$ over all 100 points.

The training sample for this experiment consists of 500 trajectories, and the test sample consists of 200 trajectories. Table 1(b) shows the performance of PDA as compared to the other algorithms using 100 uniformly spaced weights for linear combinations. Notice that PDA has higher AUC than the other methods under all choices of weights for the two criteria. For a more detailed comparison, the ROC curve for PDA and the attainable region for k-LPE (the region between the ROC curves corresponding to weights resulting in the best and worst AUCs) is shown in Figure 3 along with the first 100 Pareto fronts for PDA. k-LPE performs slightly better at low false positive rate when the best weights are used, but PDA performs better in all other situations, resulting in higher AUC. Additional discussion on this experiment can be found in Section 4 of the supplementary material.

6 Conclusion

In this paper we proposed a new multi-criteria anomaly detection method. The proposed method uses Pareto depth analysis to compute the anomaly score of a test sample by examining the Pareto front depths of dyads corresponding to the test sample. Dyads corresponding to an anomalous sample tended to be located at deeper fronts compared to dyads corresponding to a nominal sample. Instead of choosing a specific weighting or performing a grid search on the weights for different dissimilarity measures, the proposed method can efficiently detect anomalies in a manner that scales linearly in the number of criteria. We also provided a theorem establishing that the Pareto approach is asymptotically better than using linear combinations of criteria. Numerical studies validated our theoretical predictions of PDA's performance advantages on simulated and real data.

Acknowledgments

We thank Zhaoshi Meng for his assistance in labeling the pedestrian trajectories. We also thank Daniel DeWoskin for suggesting a fast algorithm for computing Pareto fronts in two criteria. This work was supported in part by ARO grant W911NF-09-1-0310.
References


