Multi-criteria Anomaly Detection using Pareto Depth Analysis

Ko-Jen Hsiao, Kevin S. Xu, Jeff Calder, and Alfred O. Hero III
University of Michigan, Ann Arbor, MI, USA 48109
{coolmark,xukevin,jcalder,hero}@umich.edu

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.

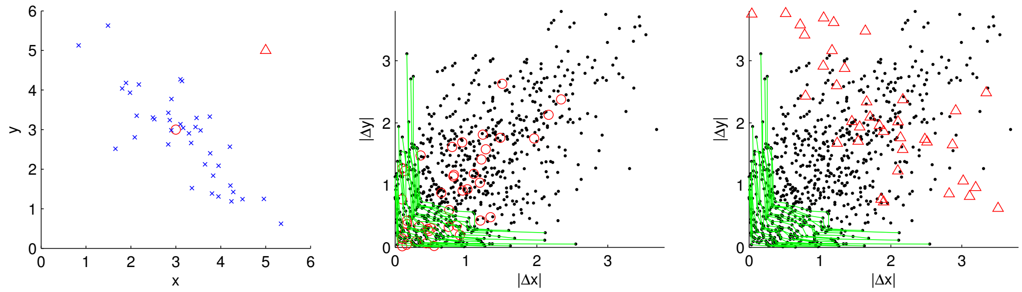

Figure 1: *Left:* Illustrative example with 40 training samples (blue x's) and 2 test samples (red circle and triangle) in $\mathbb{R}^2$. *Center:* Dyads for the training samples (black dots) along with first 20 Pareto fronts (green lines) under two criteria: $|\Delta x|$ and $|\Delta y|$. The Pareto fronts induce a partial ordering on the set of dyads. Dyads associated with the test sample marked by the red circle concentrate around shallow fronts (near the lower left of the figure). *Right:* Dyads associated with the test sample marked by the red triangle concentrate around deep fronts.

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 a 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 improve learning in another view. The problem of view disagreement, where samples take different classes in different views, has recently been investigated [7]. The views are similar to criteria in our problem setting. However, in our setting, different criteria may be orthogonal and could even give contradictory information; hence there may be severe view disagreement. Thus training in one view could actually worsen performance in another view, so the problem we consider differs from multi-view learning. A similar area is that of multiple kernel learning [8], which is typically applied to supervised learning problems, unlike the unsupervised anomaly detection setting we consider.

Finally, many other anomaly detection methods have previously been proposed. Hodge and Austin [1] and Chandola et al. [2] both provide extensive surveys of different anomaly detection methods and applications. Nearest neighbor-based methods are closely related to the proposed PDA approach. Byers and Raftery [9] proposed to use the distance between a sample and its kth-nearest neighbor as the anomaly score for the sample; similarly, Angiulli and Pizzuti [10] and Eskin et al. [11] proposed to the use the sum of the distances between a sample and its k nearest neighbors. Breunig et al. [12] used an anomaly score based on the local density of the k nearest neighbors of a sample. Hero [13] and Sricharan and Hero [14] introduced non-parametric adaptive anomaly detection methods using geometric entropy minimization, based on random k-point minimal spanning trees and bipartite k-nearest neighbor (k-NN) graphs, respectively. Zhao and Saligrama [15] proposed an anomaly detection algorithm k-LPE using local p-value estimation (LPE) based on a k-NN graph. These k-NN anomaly detection schemes only depend on the data through the pairs of data points (dyads) that define the edges in the k-NN graphs.

All of the aforementioned methods are designed for *single-criteria* anomaly detection. In the *multi-criteria* setting, the single-criteria algorithms must be executed multiple times with different weights, unlike the PDA anomaly detection algorithm that we propose in Section 4.

3 Pareto depth analysis

The PDA method proposed in this paper utilizes the notion of Pareto optimality, which has been studied in many application areas in economics, computer science, and the social sciences among others [16]. We introduce Pareto optimality and define the notion of a Pareto front.

Consider the following problem: given n items, denoted by the set $\mathcal{S}$, and K criteria for evaluating each item, denoted by functions $f_1, \dots, f_K$, select $x \in \mathcal{S}$ that minimizes $[f_1(x), \dots, f_K(x)]$. In most settings, it is not possible to identify a single item x that simultaneously minimizes $f_i(x)$ for all $i \in \{1, \dots, K\}$. A minimizer can be found by combining the K criteria using a linear combination of the f_i's and finding the minimum of the combination. Different choices of (non-negative) weights in the linear combination could result in different minimizers; a set of items that are minimizers under some linear combination can then be created by using a grid search over the weights, for example.

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 $\mathcal{S}$ that are not strictly dominated by another item in $\mathcal{S}$. 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 $\mathcal{F}_1$, 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 $\mathcal{S} \setminus \mathcal{F}_1$. More generally, define the ith Pareto front by

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

For convenience, we say that a Pareto front $\mathcal{F}_i$ is *deeper* than $\mathcal{F}_j$ 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 $A \subset \mathbb{R}^d$, we denote by $\mathcal{F}_A$ the points on the first Pareto front of $Y_1, \ldots, Y_n$ that belong to A. For simplicity, we will denote $\mathcal{F}_1$ by $\mathcal{F}$ and use $|\mathcal{F}|$ for the cardinality of $\mathcal{F}$. 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_l 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. It is well-known, and easy to see, that linear scalarization will only identify Pareto points on the boundary of the convex hull of $\bigcup_{x\in\mathcal{F}}(x + \mathbb{R}^d_+)$, where $\mathbb{R}^d_+ = \{x \in \mathbb{R}^d \,|\, x_i \geq 0, i = 1 \ldots, d\}$. Although this is a common motivation for Pareto methods, there are, to the best of our knowledge, no results in the literature regarding how many points on the Pareto front are missed by scalarization. We present such a result here. We define

$$\mathcal{L} = \bigcup_{\alpha\in\mathbb{R}^d_+} \operatorname*{argmin}_{x\in S_n} \left\{ \sum_{i=1}^{d} \alpha_i x_i \right\}, \quad S_n = \{Y_1, \ldots, Y_n\}.$$

The subset $\mathcal{L} \subset \mathcal{F}$ contains all Pareto-optimal points that can be obtained by some selection of weights for linear scalarization. We aim to study how large $\mathcal{L}$ can get, compared to $\mathcal{F}$, in expectation. In the context of this paper, if some Pareto-optimal points are not identified, then the anomaly score (defined in section 4.2) will be artificially inflated, making it more likely that a non-anomalous sample will be rejected. Hence the size of $\mathcal{F} \setminus \mathcal{L}$ is a measure of how much the anomaly score is inflated and the degree to which Pareto methods will outperform linear scalarization.

Pareto points in $\mathcal{F} \setminus \mathcal{L}$ are a result of non-convexities in the Pareto front. We study two kinds of non-convexities: those induced by the geometry of the domain of $Y_1, \ldots, Y_n$, and those induced by randomness. We first consider the geometry of the domain. Let $\Omega \subset \mathbb{R}^d$ be bounded and open with a smooth boundary $\partial\Omega$ and suppose the density f vanishes outside of Ω. For a point $z \in \partial\Omega$ we denote by $\nu(z) = (\nu_1(z), \ldots, \nu_d(z))$ the unit inward normal to $\partial\Omega$. For $T \subset \partial\Omega$, define $T_h \subset \Omega$ by $T_h = \{z + t\nu \,|\, z \in T, 0 < t \leq h\}$. Given $h > 0$ it is not hard to see that all Pareto-optimal points will almost surely lie in $\partial\Omega_h$ for large enough n, provided the density f is strictly positive on $\partial\Omega_h$. Hence it is enough to study the asymptotics for $E|\mathcal{F}_{T_h}|$ for $T \subset \partial\Omega$ and $h > 0$.

Theorem 1. *Let $f \in C^1(\overline{\Omega})$ with $\inf_\Omega f > 0$. Let $T \subset \partial\Omega$ be open and connected such that*

$$\inf_{z\in T} \min(\nu_1(z), \ldots, \nu_d(z)) \geq \delta > 0, \quad \textit{and} \quad \{y \in \overline{\Omega} : y \preceq x\} = \{x\}, \text{ for } x \in T.$$

Then for $h > 0$ sufficiently small, we have

$$E|\mathcal{F}_{T_h}| = \gamma n^{\frac{d-1}{d}} + \delta^{-d-1} O\left(n^{\frac{d-2}{d}}\right) \text{ as } n \to \infty,$$

where $\gamma = d^{-1}(d!)^{\frac{1}{d}}\Gamma(d^{-1}) \int_T f(z)^{\frac{d-1}{d}} (\nu_1(z)\cdots\nu_d(z))^{\frac{1}{d}} dz$.

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 ν 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 $E|\mathcal{F} \setminus \mathcal{L}| \geq \gamma n^{\frac{d-1}{d}} + \delta^{-d-1}O(n^{\frac{d-2}{d}})$, as $n \to \infty$, where $\gamma \geq d^{-1}(d!)^{\frac{1}{d}}\Gamma(d^{-1})|T|\delta C^{\frac{d-1}{d}}$ 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 Ω 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 $d = 2$.

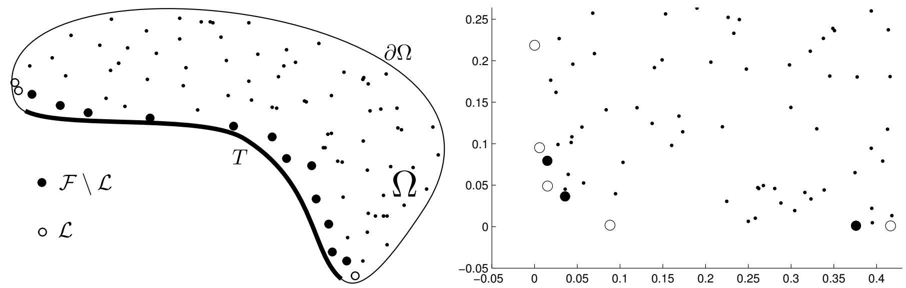

Figure 2: *Left:* Non-convexities in the Pareto front induced by the geometry of the domain Ω (Theorem 1). *Right:* Non-convexities due to randomness in the samples (Theorem 2). In each case, the larger points are Pareto-optimal, and the large black points *cannot* be obtained by scalarization.

Theorem 2. *Let* $Y_1, \ldots, Y_n$ *be independent and uniformly distributed on* $[0,1]^2$*. Then*

$$\frac{1}{2}\ln n + O(1) \le E|\mathcal{L}| \le \frac{5}{6}\ln n + O(1), \text{ as } n \to \infty.$$

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 $\frac{1}{6}$ (and likely more) of the Pareto points can only be obtained via Pareto methods even when Ω 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 $\mathcal{X}_N = \{X_1, \ldots, X_N\}$ 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 $\mathcal{X}_N$. 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 lth 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 $\mathbb{D}$. By the definition of strict dominance in Section 3, a dyad D_{ij} strictly dominates another dyad $D_{i^*j^*}$ if $d_l(i,j) \le d_l(i^*,j^*)$ for all $l \in \{1, \ldots, K\}$ and $d_l(i,j) < d_l(i^*,j^*)$ 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 $\mathcal{D} \setminus \mathcal{F}_1$ that are not strictly dominated by any other dyads from $\mathcal{D} \setminus \mathcal{F}_1$, and so on, as defined in Section 3. Recall that we refer to $\mathcal{F}_i$ 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

Algorithm 1 PDA anomaly detection algorithm.

Training phase:

for $l = 1 \to K$ **do**
  Calculate pairwise dissimilarities $d_l(i, j)$ between all training samples X_i and X_j
Create dyads $D_{ij} = [d_1(i, j), \ldots, d_K(i, j)]$ for all training samples
Construct Pareto fronts on set of all dyads until each dyad is in a front

Testing phase:

$nb \leftarrow [\,]$ {empty list}
for $l = 1 \to K$ **do**
  Calculate dissimilarities between test sample X and all training samples in criterion l
  $nb_l \leftarrow k_l$ nearest neighbors of X
  $nb \leftarrow [nb, nb_l]$ {append neighbors to list}
Create s new dyads D_i^{new} between X and training samples in nb
for $i = 1 \to s$ **do**
  Calculate depth e_i of D_i^{new}
Declare X an anomaly if $v(X) = (1/s) \sum_{i=1}^{s} e_i > \sigma$

$\mathcal{D}^{\text{new}} = \{D_1^{\text{new}}, D_2^{\text{new}}, \ldots, D_s^{\text{new}}\}$, 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[1] of X in any criterion i. We say that D_i^{new} is *below* a front $\mathcal{F}_l$ if $D_i^{\text{new}} \succ D_l$ for some $D_l \in \mathcal{F}_l$, i.e. D_i^{new} strictly dominates at least a single dyad in $\mathcal{F}_l$. Define the depth of D_i^{new} by

$$e_i = \min\{l \mid D_i^{\text{new}} \text{ is below } \mathcal{F}_l\}.$$

Therefore if e_i is large, then D_i^{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_i^{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

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[2].

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 shallower fronts than dyads corresponding to an anomalous test sample. Each test sample is associated with s new dyads, where the ith dyad D_i^{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 σ using the test

$$v(X) = \frac{1}{s} \sum_{i=1}^{s} e_i \underset{H_0}{\overset{H_1}{\gtrless}} \sigma.$$

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 and a heuristic for choosing the k_i's that performs well in practice. Both the training time and the

[1]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.

[2]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.

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.*

(a) Four-criteria simulation ($\pm$ standard error)

Method	AUC by weight	
	Median	Best
PDA	**0.948 $\pm$ 0.002**	
k-NN	0.848 $\pm$ 0.004	0.919 $\pm$ 0.003
k-NN sum	0.854 $\pm$ 0.003	0.916 $\pm$ 0.003
k-LPE	0.847 $\pm$ 0.004	0.919 $\pm$ 0.003
LOF	0.845 $\pm$ 0.003	0.932 $\pm$ 0.003

(b) Pedestrian trajectories

Method	AUC by weight	
	Median	Best
PDA	**0.915**	
k-NN	0.883	0.906
k-NN sum	0.894	0.911
k-LPE	0.893	0.908
LOF	0.839	0.863

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 1(a). 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

$$T = \begin{bmatrix} x_1 & x_2 & \dots & x_l \\ y_1 & y_2 & \dots & y_l \end{bmatrix},$$

where $[x_t, y_t]$ denote a pedestrian's position at time step t.

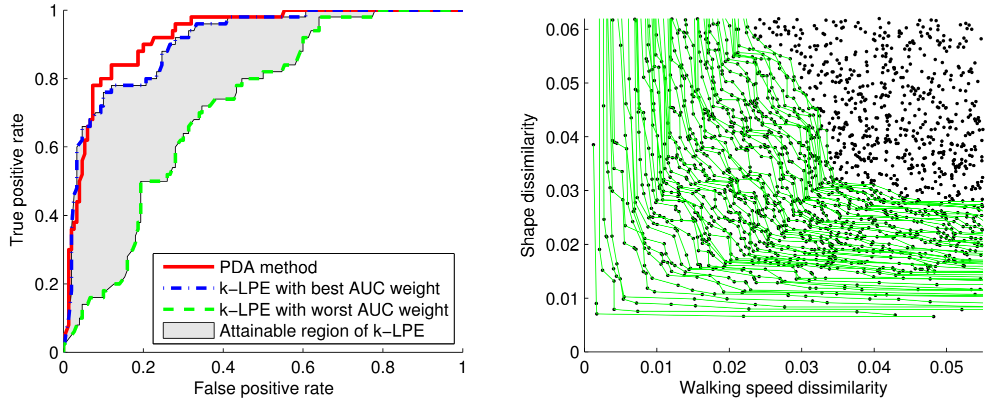

Figure 3: *Left:* ROC curves for PDA and attainable region for k-LPE over 100 choices of weights. PDA outperforms k-LPE even under the best choice of weights. *Right*: A subset of the dyads for the training samples along with the first 100 Pareto fronts. The fronts are highly non-convex, partially explaining 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

[1] V. J. Hodge and J. Austin (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review* **22**(2):85–126.

[2] V. Chandola, A. Banerjee, and V. Kumar (2009). Anomaly detection: A survey. *ACM Computing Surveys* **41**(3):1–58.

[3] Y. Jin and B. Sendhoff (2008). Pareto-based multiobjective machine learning: An overview and case studies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* **38**(3):397–415.

[4] A. O. Hero III and G. Fleury (2004). Pareto-optimal methods for gene ranking. *The Journal of VLSI Signal Processing* **38**(3):259–275.

[5] A. Blum and T. Mitchell (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*.

[6] V. Sindhwani, P. Niyogi, and M. Belkin (2005). A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of the Workshop on Learning with Multiple Views, 22nd International Conference on Machine Learning*.

[7] C. M. Christoudias, R. Urtasun, and T. Darrell (2008). Multi-view learning in the presence of view disagreement. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

[8] M. Gönen and E. Alpaydın (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research* **12**(Jul):2211–2268.

[9] S. Byers and A. E. Raftery (1998). Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association* **93**(442):577–584.

[10] F. Angiulli and C. Pizzuti (2002). Fast outlier detection in high dimensional spaces. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*.

[11] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo (2002). A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In *Applications of Data Mining in Computer Security*. Kluwer: Norwell, MA.

[12] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander (2000). LOF: Identifying density-based local outliers. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*.

[13] A. O. Hero III (2006). Geometric entropy minimization (GEM) for anomaly detection and localization. In *Advances in Neural Information Processing Systems 19*.

[14] K. Sricharan and A. O. Hero III (2011). Efficient anomaly detection using bipartite k-NN graphs. In *Advances in Neural Information Processing Systems 24*.

[15] M. Zhao and V. Saligrama (2009). Anomaly detection with score functions based on nearest neighbor graphs. In *Advances in Neural Information Processing Systems 22*.

[16] M. Ehrgott (2000). *Multicriteria optimization*. Lecture Notes in Economics and Mathematical Systems 491. Springer-Verlag.

[17] O. Barndorff-Nielsen and M. Sobel (1966). On the distribution of the number of admissible points in a vector random sample. *Theory of Probability and its Applications*, **11**(2):249–269.

[18] Z.-D. Bai, L. Devroye, H.-K. Hwang, and T.-H. Tsai (2005). Maxima in hypercubes. *Random Structures Algorithms*, **27**(3):290–309.

[19] Y. Baryshnikov and J. E. Yukich (2005). Maximal points and Gaussian fields. Unpublished. URL `http://www.math.illinois.edu/~ymb/ps/by4.pdf`.

[20] B. Majecka (2009). *Statistical models of pedestrian behaviour in the Forum*. Master's thesis, University of Edinburgh.

[21] R. R. Sillito and R. B. Fisher (2008). Semi-supervised learning for anomalous trajectory detection. In *Proceedings of the 19th British Machine Vision Conference*.