ISSN 1346-5597

NII National Institute of Informatics

NII Technical Report

Intrinsic Dimensional Outlier Detection in High-Dimensional Data

Jonathan von Brünken, Michael E. Houle, and Arthur Zimek

NII-2015-003E
Mar. 2015

URL: http://www.nii.ac.jp/ Tokyo 101-8430, JAPAN +81-3-4212-2000 FAX: +81-3-3556-1916

Intrinsic Dimensional Outlier Detection in High-Dimensional Data

Jonathan von Brünken, Michael E. Houle, Arthur Zimek

Abstract—We introduce a new method for evaluating local outliers, by utilizing a measure of the intrinsic dimensionality in the vicinity of a test point, the continuous intrinsic dimension (ID), which has been shown to be equivalent to a measure of the discriminative power of similarity functions. Continuous ID can be regarded as an extension of Karger and Ruhl's expansion dimension to a statistical setting in which the distribution of distances to a query point is modeled in terms of a continuous random variable. The proposed local outlier score, IDOS, uses ID as a substitute for the density estimation used in classical outlier detection methods such as LOF. An experimental analysis is provided showing that the precision of IDOS substantially improves over that of state-of-the-art outlier detection scoring methods, especially when the data sets are large and high-dimensional.

1 Introduction

The goal of outlier detection, one of the fundamental data mining tasks, is to identify data objects that do not fit well in the general data distribution. Applications include areas as diverse as fraud detection, error elimination in scientific data, or sports data analysis. Examples of successful outlier detection could be the detection of stylistic elements of distinct origins in written work as hints of plagiarism; or deviations in scientific data as indicators of equipment malfunction, of human error in data processing, or of a sub-optimal experimental setup.

Algorithmic approaches to outlier detection are as diverse as their application scenarios. A major — and very successful — family of methods relies on density estimation based on k-nearest-neighbor distances, in which the points with the lowest density estimates are reported as the strongest outlier candidates [1], [2], [3]. As a refined technique, the well-known method LOF [4] compares density estimates in a local context. This notion of locality has led to a large number of variants adapted to different contexts [5].

One problem commonly faced in outlier detection is the deterioration of the quality of density estimates as the dimensionality of the data increases. A recent survey [6] discusses several aspects of the 'curse of dimensionality', such as the phenomenon of distance concentration [7], [8], [9]. This effect, however, is not directly connected to the representational data dimension (the number of attributes); rather, it is better explained by notion of intrinsic dimensionality, measures of which can account for the data complexity and resulting performance loss of many data mining tasks in high-dimensional settings [6], [10], [11], [12].

In this paper, we demonstrate how a measure of intrinsic dimensionality can be used to improve outlier detection in data with strongly varying intrinsic dimensionality. Based on an adaptation of LOF [4], we present an efficient new method for outlier ranking, IDOS (Intrinsic Dimensionality Outlier Score), that is especially suitable for large and high dimensional datasets. We show how the density estimation performed by LOF can be explained in terms of the intrinsic dimensionality of continuous distance distributions, thereby allowing us to demonstrate the advantages of our approach in discriminating between inliers and outliers. IDOS, in its reliance on estimates of intrinsic dimensionality rather than density, will be shown to better address the challenges of the 'curse of dimensionality' for outlier detection [6].

The paper is organized as follows. In the next section, we will discuss different approaches to outlier detection, and their applicability to large and high dimensional data. In Section 3, we provide an overview of a model of continuous intrinsic dimensionality, and develop and explain our proposed outlier ranking method, IDOS. In Section 4, we evaluate our algorithm on different data sets, comparing its performance to state-of-the-art competitors on data sets whose dimensions span more than four orders of magnitude. Concluding remarks are presented in Section 5.

2 Related work

Models for outlier detection have been categorized as 'global' or 'local', based on the scope of the background against which they perform their assessment. However, locality is not a binary predicate but rather a matter of degree [5]: 'global' distance-based outlier detection models such as DB-Outlier [1] or k-NN Outlier [2], [3] compare a local property of a data point — such as a local density estimate calculated as the number of points

- *Jonathan von Brünken and Arthur Zimek are with the Institute for Informatics, Ludwig-Maximilians-Universität München, Oettingenstr. 67, 80538 Munich, Germany*
  http://www.dbs.ifi.lmu.de
  joni@jgeg.de, zimek@dbs.ifi.lmu.de
- *Michael E. Houle is with the National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan*
  meh@nii.ac.jp

in a predefined ε-neighborhood, or as the distance to the k-th nearest neighbor — to the corresponding property of all other data points. That is, the outlier ranking is based on a comparison of a local model with the complete (global) dataset as a reference set.

So called 'local' methods, such as LOF [4] and its many variants [13], [14], [15], [16], [17], compare the local density estimate with a *local reference set*. Typically, the same set of neighbors (such as the ε-neighborhood or the k nearest neighbors) is used both as a context set for the density estimate and as a reference set for model comparison, but this is not necessarily so. Schubert et al. [5], [18] demonstrated the improved applicability of local outlier detection in a more general setting, using different sets for model context, and as a reference for model comparison. We follow this pattern of 'generalized' local outlier detection [5], as we will indeed require two neighborhood sets with considerably different roles.

There exist particular challenges for outlier detection in high-dimensional data [6]. One recent offshoot in the line of investigation was to treat the problem as 'subspace' outlier detection [6], [19], [20], [21], [22], [23], [24], [25], or to define models that are more stable with increasing data dimensionality, such as an angle-based outlier model [26]. Ensemble methods for outlier detection can also be seen as variants particularly useful for high-dimensional data [27]. Another branch of investigation is more concerned with efficiency issues, using sophisticated acceleration mechanisms for high-dimensional data, such as random projections [28], [29] or locality sensitive hashing (LSH) [30]. These approaches all tackle some aspects of the 'curse of dimensionality', but none provides a comprehensive and satisfying solution [6].

To summarize, the challenges that arise with increasing intrinsic dimensionality of datasets have been treated in different ways and to differing degrees by these specialized approaches, but none has been proved to excel with high-dimensional data. The classic LOF method [4] can still be seen as a competitive state-of-the-art method. In this paper, instead of attempting to merely cope with the curse of dimensionality, we show for the first time how one may make use of estimates of the local intrinsic dimensionality for outlier detection, simultaneously improving effectiveness and efficiency.

3 Outlier Detection and Intrinsic Dimensionality

In this section, we present a new outlier ranking strategy that makes use of a measure of intrinsic dimensionality of the data set in the vicinity of candidate objects. We begin by giving an overview of the various forms of intrinsic dimensionality that have been studied in the past. Next, we show how local estimates of intrinsic dimensionality can be used as a substitute for the density estimation traditionally used by the popular LOF outlier scoring function. We conclude the section by showing how local intrinsic dimensionality can be estimated from data samples, while accounting for stability problems when the sample size is small.

3.1 Intrinsic Dimensionality

Intrinsic dimensionality (ID) can be regarded as the number of latent variables or degrees of freedom needed to describe a given data set. It serves as a more realistic measure of the complexity of data, and as a predictor of the degradation in performance of fundamental data mining operations (such as search, classification, and clustering) for high dimensional applications, as well as determining an appropriate target dimension for dimensional reduction.

3.1.1 Overview

Many models of intrinsic dimensionality have been proposed over the past few decades. Using local data samples, topological models estimate the basis dimension of the tangent space of the data manifold [31], [32], [33], [34]. Projection models construct a subspace to which a data sample could be projected while minimizing the error of fit; the dimension of the resulting subspace is taken to be the intrinsic dimension. Examples include PCA and its variants [35], [36], manifold learning [37], [38], [39], and other non-linear extensions [40]. Multidimensional scaling attempts to determine a projection that preserves local pairwise distances within the data sample [37], [41]. Fractal models estimate an intrinsic dimension from the degree of self-similarity of the data, or the capacity of the data to fill the space within which it is contained [42], [43], [44]. Shattering models estimate dimensionality from the number of subconfigurations of data points that can be distinguished using a collection of splitters — a famous example is the Vapnik-Chervonenkis (VC) dimension [45]. Statistical estimators of intrinsic dimension can often be derived via parametric modeling and estimation of distribution [46], [47]. More recently proposed intrinsic dimensionality models, such as the expansion dimension [10] and the generalized expansion dimension [12], quantify intrinsic dimensionality in terms of the rate at which the number of encountered data objects grows as the range of distances expands. Expansion models of dimensionality have recently been applied to the design and analysis of index structures for similarity search [10], [48], [49], [50], and the analysis of a projection-based heuristic for LOF [28].

3.1.2 Continuous Intrinsic Dimensionality

Very recently, an expansion model of intrinsic dimensionality has been introduced for continuous distance distributions [51]. The distance from a given reference object is modeled in terms of an absolutely continuous random variable $\mathbf{X}$ with support $[0, \infty)$. Let $f_{\mathbf{X}}$ denote the probability density of $\mathbf{X}$, and $F_{\mathbf{X}}$ denote the corresponding cumulative density function. Whenever $\mathbf{X}$ is absolutely continuous at x, $F_{\mathbf{X}}$ is differentiable at x and its first-order derivative is $f_{\mathbf{X}}(x)$. With respect to this distribution, the continuous ID of $\mathbf{X}$ at distance x is defined to be

$$\mathrm{ID}_{\mathbf{X}}(x) :\equiv \lim_{\epsilon\to 0^+} \frac{\ln F_{\mathbf{X}}\left((1+\epsilon)x\right) - \ln F_{\mathbf{X}}(x)}{\ln(1+\epsilon)}$$

With respect to the generalized expansion dimension [12], the above definition can be regarded as the outcome of a dimensional test of neighborhoods of radii x and $(1+\epsilon)x$ in which the role of neighborhood cardinalities is filled by the expected number of neighbors. Continuous ID also turns out to be equivalent to a formulation of the (lack of) discriminative power of a distance measure at distance x from the reference object, as both formulations can be shown to have the same closed form [51]:

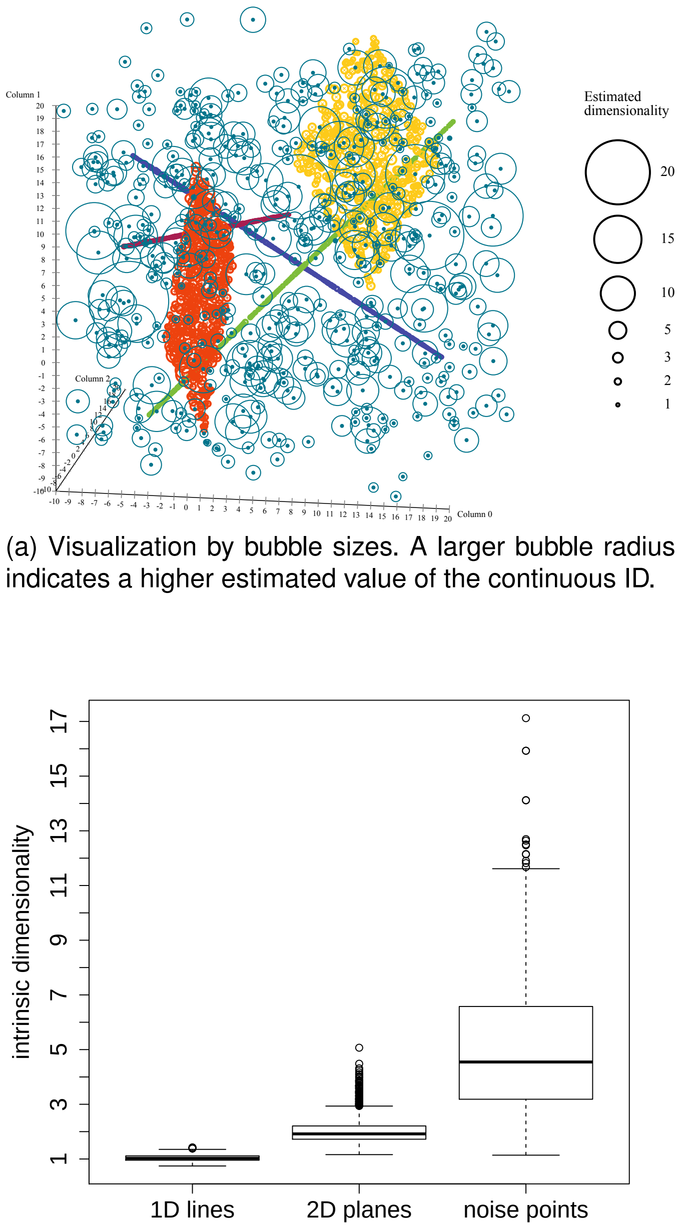
$$\mathrm{ID}_{\mathbf{X}}(x) = \frac{x f_{\mathbf{X}}(x)}{F_{\mathbf{X}}(x)}. \qquad (1)$$

Another quantity of interest is the limit of the continuous ID as the distance x tends to zero:

$$\mathrm{ID}_{\mathbf{X}} = \lim_{x\to 0^+} \mathrm{ID}_{\mathbf{X}}(x).$$

This quantity, simply referred to as the 'continuous intrinsic dimensionality' without mention of distance, has been shown [52] to be directly related to the shape parameter of the generalized Pareto distribution from extreme value theory, for the case of bounded distributional tails. Also shown is that for data distributed within any d-dimensional manifold within an m-dimensional space, $\mathrm{ID}_{\mathbf{X}} = d$.[1]

Figure 1 illustrates the ability of the continuous ID score of a reference point to reveal the dimension of the local manifold containing that point. With respect to a 3-dimensional domain, data points were generated within five manifolds, three of dimension 1 and two of dimension 2. A substantial proportion of noise points was also generated. Inlier points were associated with continuous ID scores that were in strong agreement with the dimensions of the manifolds to which they belonged. On the other hand, most of the outlier (noise) points had ID scores substantially larger than those of the inlier points. From the vantage of an outlier point q, the distribution of distances suffers from a lack of discriminability (or equivalently, a high continuous ID) due to the large numbers of inlier points (manifold points) having similar distances to q. The example shows that a relatively large continuous ID score has the potential to reveal an outlier in the vicinity of a cluster of inlier points.



(a) Visualization by bubble sizes. A larger bubble radius indicates a higher estimated value of the continuous ID.

(b) Visualization by box plots of the different types of clusters. The horizontal line in the middle of each box denotes the median intrinsic dimension value, while the limits of the box are at the upper and lower quartiles. The whiskers (dotted lines) denote the range of intrinsic dimensional values within 1.5 times the inter-quartile distance. All measurements falling outside this larger range are displayed as individual dots.

Fig. 1: Visualization of the continuous ID of a 3-dimensional data set using the estimator presented in Section 3.3.

3.2 Intrinsic Dimensionality Outlier Score

Our proposed outlier ranking function, IDOS, can be regarded as a reworking of the well-established Local Outlier Factor (LOF) score [4], in which the local density estimation performed by LOF is replaced by the local estimation of intrinsic dimensionality. Before presenting the details of IDOS, we first show how LOF can be reinterpreted in terms of the continuous intrinsic dimensionality.

[1] For more on properties of continuous ID, and on its connection to extreme value theory and the manifold structure of data, we refer the reader to [51], [52].

Based on the k-distance $\mathrm{dist}_k(p)$ — the radius of the smallest ball centered at p that captures its k nearest neighbors — the k-reachability distance from point p to point o is:

$$\mathrm{rdist}_k(p,o) \quad = \quad \max\{\mathrm{dist}_k(o), \mathrm{dist}(p,o)\}$$

LOF uses this asymmetric distance measure for the estimation of the data density in the vicinity of a point p, defined as the data mass within the context set $\mathrm{N}_k(p)$ of size k for p (this mass being k), divided by the average k-reachability distance $\mathrm{avg\_rdist}_k(p)$ over the context set:

$$\begin{aligned}\mathrm{lrd}_k(p) \quad &= \quad \frac{k}{\mathrm{avg\_rdist}_k(p)}, \quad \text{where} \\ \mathrm{avg\_rdist}_k(p) \quad &= \quad \frac{1}{k} \cdot \sum_{o \in \mathrm{N}_k(p)} \mathrm{rdist}_k(p,o)\,.\end{aligned}$$

This density measure, referred to as the local reachability density, provides more stability as compared to the often-used simple density $\mathrm{sd}_k(p) = k/\,\mathrm{dist}_k(p)$. It should be noted that in the original formulation of LOF [4], $\mathrm{lrd}_k(p)$ was defined as $1/\,\mathrm{avg\_rdist}_k(p)$ — justifiable whenever k is a fixed constant, as the omission of the mass factor k would lead to equivalent results. Since we wish to leave open the possibility of comparing results while varying the neighborhood size k, in this paper we have opted for the more traditional treatment of radial data density as the ratio between measures of data mass and radial distance.

The use of reachability distance in the formulation of density leads to estimates with different tendencies, depending on whether the test point p is an inlier or an outlier. For outliers in the vicinity of some cluster of inliers, $\mathrm{dist}_k(o)$ can be expected to be smaller than $\mathrm{dist}(p,o)$ – in which case $\mathrm{lrd}_k(p)$ tends to $\mathrm{sd}'_k(p) = k/\,\mathrm{dist}'_k(p)$, where $\mathrm{dist}'_k(p)$ is the average of the distances from p to the members of its context set. On the other hand, for inliers located within a cluster, the local reachability density tends to $\mathrm{sd}''_k(p) = k/\,\mathrm{dist}''_k(p)$, where $\mathrm{dist}''_k(p)$ is the average of the k-distances of the members of the context set of p (which are very likely to be inliers).

The LOF score contrasts the density model of a query object q with those of a reference set local to q. More precisely, in LOF the reference sets are identical to the context sets within which the local reachability densities are measured. The LOF score is the ratio between the average density taken over the reference set, and the density based at q:

$$\mathrm{LOF}_k(q) \quad = \quad \frac{1}{k} \cdot \frac{\sum\limits_{p \in \mathrm{N}_k(q)} \mathrm{lrd}_k(p)}{\mathrm{lrd}_k(q)}$$

Inliers can be expected to have LOF scores near 1, whereas those having scores significantly higher than 1 would be deemed to be outliers. For many applications, due to the high cost of computing neighborhoods of context set items, it is often convenient to compute a simplified version of LOF in which the simple density sd_k is used in place of the local reachability density lrd_k [5].

By letting the context sets also serve as the reference sets for the comparison of density models, LOF ensures that the determination of outliers is purely with respect to the local context associated with the test object. This insistence on locality is the reason why LOF works well with data having more than one generation mechanism. However, since it is a distance measure subject to the 'curse of dimensionality', the local reachability density loses its explanatory power when the local intrinsic dimension is high.

The model of ID introduced in [51] establishes a relationship among the continuous intrinsic dimensionality $\mathrm{ID}_{\mathbf{X}}(x)$, and the probability density $f_{\mathbf{X}}(x)$ and cumulative density $F_{\mathbf{X}}(x)/x$ of the underlying distance distribution (as expressed in Equation 1). Given a sample of n data points, the simple density

$$\frac{1}{n}\,\mathrm{sd}_k(p) = \frac{k}{n \cdot \mathrm{dist}_k(p)}$$

can be viewed as an estimator for the distributional cumulative density at distance $\mathrm{dist}_k(p)$; this in turn implies that the cumulative density can be approximated by $\frac{1}{n}\,\mathrm{lrd}_k$. The LOF formula can then be regarded as an approximation of the ratio between the average cumulative density of the reference set objects and the cumulative density of the query object, all taken at the k-distances of the respective points involved:

$$\begin{aligned}\mathrm{LOF}_k(q) \quad &\approx \quad \frac{1}{k} \cdot \sum_{p \in \mathrm{N}_k(q)} \frac{F_p(\mathrm{dist}_k(p))}{\mathrm{dist}_k(p)} \Bigg/ \frac{F_q(\mathrm{dist}_k(q))}{\mathrm{dist}_k(q)} \\ &= \quad \frac{1}{k} \cdot \sum_{p \in \mathrm{N}_k(q)} \frac{f_p(\mathrm{dist}_k(p))}{\mathrm{ID}_p(\mathrm{dist}_k(p))} \Bigg/ \frac{f_q(\mathrm{dist}_k(q))}{\mathrm{ID}_q(\mathrm{dist}_k(q))}\end{aligned}$$

where ID_p is the continuous ID based at object p.

The reformulation of LOF in terms of continuous ID reveals the degree to which it is sensitive to variations in the probability density at various distances from the query and reference objects. When the reference set consists mainly of inlier points belonging to a common cluster manifold, significant variation in the probability density models would tend to reduce the contrast between the reference objects and the query object, in terms of their contributions to the LOF score. However, as indicated by the example in Figure 1, if the density models were replaced by estimations of the local intrinsic dimensionality, the contributions of the reference objects would tend to be more uniform, improving the contrast between the query object and reference objects.

Eliminating from LOF the variation due to distance suggests the following transformation:

$$\mathrm{LOF}_k(q) \quad \approx \quad \frac{1}{k} \cdot \sum_{p \in \mathrm{N}_k(q)} \frac{f_p(\mathrm{dist}_k(p))}{\mathrm{ID}_p(\mathrm{dist}_k(p))} \Bigg/ \frac{f_q(\mathrm{dist}_k(q))}{\mathrm{ID}_q(\mathrm{dist}_k(q))}$$

$$\longrightarrow \quad \frac{1}{k} \cdot \sum_{p \in \mathrm{N}_k(q)} \frac{1}{\mathrm{ID}_p} \Bigg/ \frac{1}{\mathrm{ID}_q} = \mathrm{ID}_q \cdot \left(\frac{1}{k} \sum_{p \in \mathrm{N}_k(q)} \frac{1}{\mathrm{ID}_p} \right)$$

This leads to our proposed outlier scoring function based on intrinsic dimensionality:

$$\mathrm{IDOS}(q) \;\; :\equiv \;\; \frac{\mathrm{ID}_q}{|\,\mathrm{N}_q^r\,|} \cdot \sum_{p \in \mathrm{N}_q^r} \frac{1}{\mathrm{ID}_p} \tag{2}$$

As with the original LOF score, IDOS would tend to assign inlier objects scores close to 1, while outliers would tend to receive substantially higher scores.

Here, N_q^r denotes the reference set for query q. The context sets are not stated in this formulation, since it does not specify the intrinsic dimensionality to be estimated. In principle, any definition of intrinsic dimensionality could be considered, provided that stable estimates can be calculated in a reasonably efficient manner.

3.3 Estimation of Intrinsic Dimensionality

In our treatment we will make use of the continuous ID, not only for the theoretical properties discussed earlier, but also since several efficiently-computable statistical estimators have been proposed and evaluated [52]. For our implementation, we chose one of the best-performing estimators, a maximum-likelihood estimator (MLE) following the form of the well-known Hill estimator for the scaling exponent of a bounded power-law tail distribution [53]:

$$\widehat{\mathrm{ID}}_{\mathbf{X}} = -\left(\frac{1}{k} \sum\nolimits_{i=1}^{k} \ln \frac{x_i}{x_k} \right)^{-1} \tag{3}$$

We assume that the sample consists of a sequence $x_1, \ldots, x_k$ of observations of a random distance variable $\mathbf{X}$ with support $[0, w)$, in ascending order, i.e., $x_1 \leq x_2 \leq \cdots \leq x_k$.

The sample can be regarded as the context set for an ID-based outlier model building step analogous to the model building step [5]. If the context set is a neighborhood of q, the observations can be regarded as being derived from a distance tail distribution. While the MLE is known to be only asymptotically unbiased [54] for tail distributions, for our applications the bias may be neglected, provided that the reference set sizes are kept constant.

The asymptotic variance of the Hill estimator is known to increase with decreasing sample size, and with increasing intrinsic dimensionality [55]. Special attention must therefore be given when the sample sizes are small.

The experimental evaluation in [52] indicates that the MLE estimator generally produces stable estimates once the sample size reaches approximately 100. For samples of this size, the MLE estimator is used directly. However, all data points coinciding with the query object are first removed from the sample, as even a single distance value of zero would force the estimation of ID to be zero.

Algorithm 1: Algorithmic specification of IDOS.

```
Data: Database $D$
Input: context set size $k_c \geq 3$
Input: reference set size $k_r \geq 1$
Result: Outlier score $\mathrm{IDOS}(q)$ for each $q \in D$
// Preprocessing
forall the $p \in D$ do
    Compute reference set $\mathrm{N}_p^r$ of size $k_r$;
    Compute context set $\mathrm{N}_p^c$ of size $k_c$;
end
// Model building step
forall the $p \in D$ do
    if $k_c \geq 100$ then
        Compute intrinsic dimensionality $\widehat{\mathrm{ID}}_p$ directly
        from Equation 3;
    else
        Compute $\widehat{\mathrm{ID}}_{p,2}$;
        for $i$ = 3 to $k_c$ do
            compute $\widehat{\mathrm{ID}}_{p,i}$ from $\widehat{\mathrm{ID}}_{p,i-1}$ using Equation 5;
        end
        Compute $\widehat{\mathrm{ID}}_p$ as the weighted harmonic mean of
        all previously computed $\widehat{\mathrm{ID}}_{p,i}$, using Equation 4;
    end
end
// Model comparison step
forall the $q \in D$ do
    Compute $\widehat{\mathrm{ID}}_{\mathrm{N}}^{-1}$ as the arithmetic mean of $\widehat{\mathrm{ID}}_p^{-1}$ for all
    $p \in \mathrm{N}_q^r$;
    Compute $\mathrm{IDOS}(q) = \widehat{\mathrm{ID}}_q \cdot \widehat{\mathrm{ID}}_{\mathrm{N}}^{-1}$;
end
```

Estimation of continuous ID with sample size smaller than 100 can lead to unacceptably high variance. Smoothing the estimation by averaging over subsamples can help to lower this variance [54]. MLE (Equation 3) is used to obtain values of continuous ID for each subsample of the form $\mathrm{N}_{q,j} = \{x_1, \ldots, x_j\}$, for $1 \leq j \leq k$. A weighted harmonic mean of these ID values is recursively computed, using weights depending on the sample size, as follows:

$$\widehat{\mathrm{ID}}_q \;\; = \;\; \left(\sum_{j=1}^{k} \left(w_j \cdot \frac{1}{\widehat{\mathrm{ID}}_{q,j}} \right) \right)^{-1} \tag{4}$$

$$\widehat{\mathrm{ID}}_{q,j+1} \;\; = \;\; \frac{j+1}{j} \cdot \left(\frac{1}{\widehat{\mathrm{ID}}_{q,j}} + \ln\left(\frac{x_{j+1}}{x_j} \right) \right)^{-1} \tag{5}$$

where $\widehat{\mathrm{ID}}_{q,j}$ is the estimated intrinsic dimensionality computed over the subsample $\mathrm{N}_{q,j}^r$, and

$$w_j = (j-1) \Bigg/ \sum_{i=1}^{k-1} i = \frac{2j-2}{k^2-k}$$

is the weight of the j-th term.

Equation 4 can trivially be computed for all objects in $\mathcal{O}(k^2 \cdot |\,\mathrm{N}_q^r\,|)$ by computing each of the $k-1$ individual intrinsic dimensionalities in $\mathcal{O}(k)$ time. The execution cost can be reduced to $\mathcal{O}(k \cdot |\,\mathrm{N}_q^r\,|)$ by recursive exploitation of Equation 5. If the sizes of the reference sets and context sets are fixed at k_r and k_c, respectively, the cost of computing an outlier score for each database object is in $\mathcal{O}(n \cdot \max\{k_r, k_c\})$. To prevent unnecessary recomputations, the nearest neighbor distances and intrinsic dimensionalities need to be stored, resulting in $\mathcal{O}(n \cdot \max\{k_r, k_c\})$ complexity for the storage requirements as well. The full outlier detection process is summarized as Algorithm 1.

4 Evaluation

4.1 Competitors and Evaluation Measures

We compare against state-of-the-art competitors, namely the Local Outlier Factor (LOF) [4], as well as two different models targeting high-dimensional data: SOD [19], as a representative of subspace outlier detection algorithms, and FastABOD [26], the quadratic-time approximation variant of angle-based outlier detection (ABOD) [26].

All algorithms and experiments were implemented and executed in the ELKI framework [56]. We compare the results based on the commonly-used Euclidean distance, as well as on the arc cosine distance that is often preferred for sparse, high-dimensional data. We report the precision of the outlier ranking, as assessed by the ratio of outliers correctly assigned to the top-t listed objects, where t is the total number of outliers in the respective dataset. In addition, we report the performance for each dataset by the area under the curve of the receiver operating characteristic (ROC AUC). ROC curves plot the true positive rate against the false positive rate. The resulting monotone curves are converted into a performance measure by computing the area under this curve (AUC). This allows several results to be displayed in a single graph, as well as a numerical comparison of the results. ROC curves and ROC AUC analysis inherently treat the class imbalance problem by using relative frequencies, which makes them popular for the evaluation of outlier detection strategies.

4.2 Datasets

For benchmarking our method, we designed four datasets, varying the representational dimensionality over four orders of magnitude (see the summary in Table 1).

The ALOI dataset is based on the Amsterdam Library of Object Images [57], a collection of 110,250 images of 1000 objects, taken under systematically varied conditions. We used the Object View Collection as source for the inlier set, resulting in 1000 small clusters, each containing 72 photographs of the associated object taken from different angles. To produce the local outlier set, for each object we selected images taken under markedly different illumination conditions. The final vectors were generated as a combination of RGB, HSV, YCbCr and edge-direction histograms. Despite the high representational dimensionality, the objects of the resulting dataset are associated with relatively low intrinsic dimensionality scores: typically less than 5 for inliers, and between 10 and 20 for outliers, as indicated in Figure 2a. The intrinsic dimensionalities of inliers and of outliers are well distinguished.

The FMA dataset was created from 190 public-domain mp3 files, downloaded from freemusicarchive.org in January 2013. The files were divided into 16 groups according to their meta tags and featured instrumentation. Inliers were created by concatenating segments randomly selected from files within a common group, whereas outliers were created by concatenating segments from different groups. The resulting sound files were converted into 1380-dimensional *rhythm patterns* [59]. This construction resulted in a dataset of relatively high intrinsic dimensionality, exhibiting a considerable overlap between the range of intrinsic dimensional values for inliers and the range for outliers (Figure 2b).

The Amazon dataset is based on the Amazon Commerce reviews set from the UCI Machine Learning Repository [58]. We used convex combinations of all pairs of distinct points belonging to the same reviewer to form the inlier set. To create the outlier set, for each pair of users, a convex combination was created from two reviews chosen at random (one per user). The typical range of intrinsic dimensional values is between 10 and 20, with an even stronger overlap between the ranges for inliers and outliers (Figure 2c).

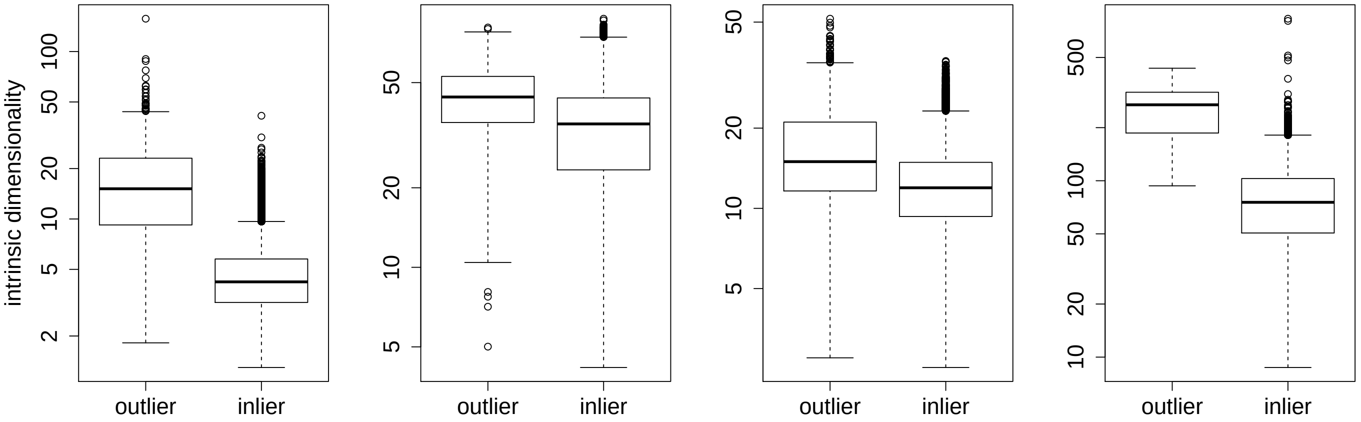
The Reuters dataset was created from the Reuters RCV1-RCV2 Multilingual data corpus [60] available from the UCI Machine Learning Repository [58]. It contains 111,740 vectorized documents in 5 languages from the RCV1 and RCV2 collections of Reuters news agency articles. Also included in the corpus are document vectors representing machine translations of each original document into all four other languages. Each document is assigned to one of six possible classes. To generate an inlier data set, for each article we concatenated the vectors for each of its 5 language versions into a single vector. The resulting inlier set possesses a high correlation between corresponding words in different translations. Outliers were generated by the combination of five vectors corresponding to 5 documents, each of a different language, and from distinct article classes. Both inlier and outlier document vectors were normalized, so as to allow meaningful application of the Euclidean distance measure as well as the more traditional cosine similarity [61]. The resulting set is represented by 107,756 dimensions; however, the representation is very sparse, with the maximum number of non-zero dimensions being 5,387 for inliers, and 751 for the outliers. As shown in Figure 2d, while being the highest over our four datasets, the median intrinsic dimensionality is still surprisingly low considering the very high representational dimension.

4.3 Results

Much like the k_r parameter of IDOS, the competing methods all require that a parameter value k be supplied

TABLE 1: Summary of the datasets. The datasets are available at
`https://drive.google.com/file/d/0BwdAgJeKIbgEc1BJaFM5cF9kMEU/edit?usp=sharing`.

Name	Dimension	Instances	Outliers	Proportion of Outliers	Representation
ALOI [57]	612	73,000	1,000	1.37%	color & edge histograms
FMA	1,380	77,000	1,000	1.30%	rhythm patterns
Amazon [58]	10,000	22,975	1,225	5.33%	combined occurrence counts
Reuters [58]	107,756	106,701	720	0.67%	combined word vectors



(a) ALOI, k_c=50 (b) FMA, k_c=300 (c) Amazon, k_c=300 (d) Reuters, k_c=5000

Fig. 2: Box plots for the intrinsic dimensionality of inliers and outliers in the evaluation datasets, as measured at the context set size k_c for which IDOS performed best (in the experimentation of Section 4). The horizontal line in the middle of each box denotes the median intrinsic dimension value, while the limits of the box are at the upper and lower quartiles. The whiskers (dotted lines) denote the range of intrinsic dimensional values within 1.5 times the inter-quartile distance. All measurements falling outside this larger range are displayed as individual dots.

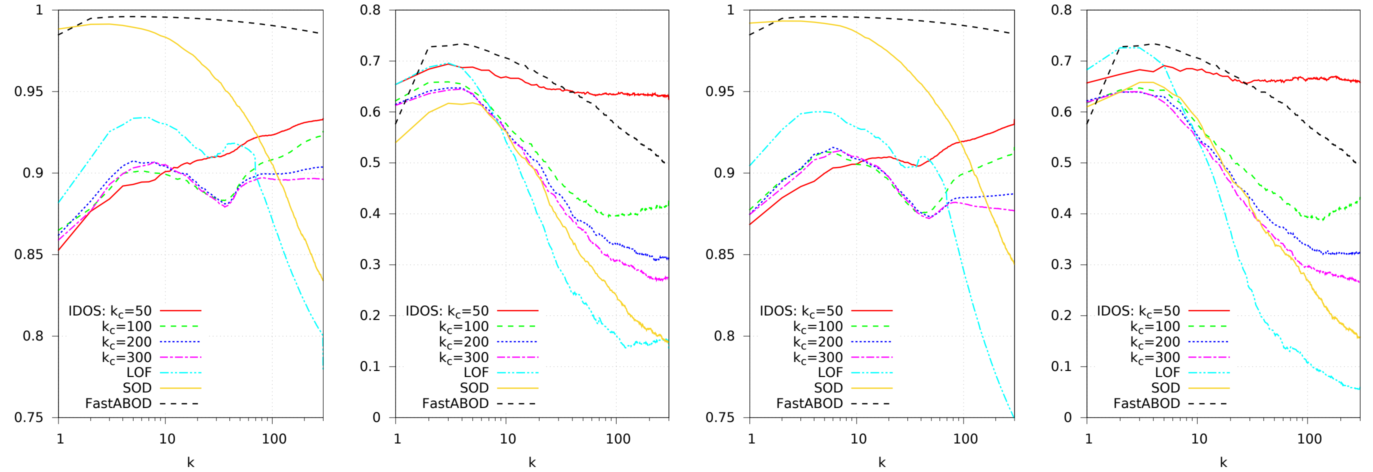
to determine the size of neighborhoods used as reference sets. We tested as parameter values for k the complete range from 1 to 300. In the result plots, the changing value for $k = k_r$ spans the x-axis. The parameter of IDOS defining the size of the context set (neighborhood) used for the estimation of continous ID, k_c, was set to values between 50 and 5000 depending on the data set: for ALOI, FMA, and Amazon we computed results for neighborhood sizes of 50, 100, 200, and 300, while for the very high dimensional Reuters dataset we chose values of 500, 1000, 2000, and 5000. For each choice of k_c, we plot a different curve.

In general, the estimation of continuous ID performed by IDOS can be expected to yield good results for sufficiently large k_c, subject to the following two restrictions: k_c should be less than 5% of the dataset, to ensure that the resulting sample retains the properties of a distribution tail; and less than the average inlier cluster size, to minimize spurious impact from unrelated clusters. We note that the latter restriction was discussed in the original LOF publication [4], and is a restriction for virtually all neighborhood-based outlier detection models. For the ALOI and Amazon sets, the limiting condition is the cluster size (72 and 435, respectively). For FMA and Reuters, the 5% limit on the sample proportion is more restrictive.

For the ALOI dataset, we display the results using Euclidean distance in Figures 3a and 3b, while Figures 3c and 3d show the performance for the arc cosine distance. FastABOD performs excellently on ALOI, while SOD detects outliers especially well when k is smaller than the cluster sizes. As one might expect due to the relatively small cluster sizes impeding the accurate estimation of intrinsic dimensionality, IDOS does not outperform FastABOD, and deteriorates with increasing k_c. On the other hand, IDOS does perform comparably to LOF, especially in terms of precision. SOD exhibits very good ROC AUC values, but its precision is not competitive to that of other algorithms, especially when using Euclidean distance.

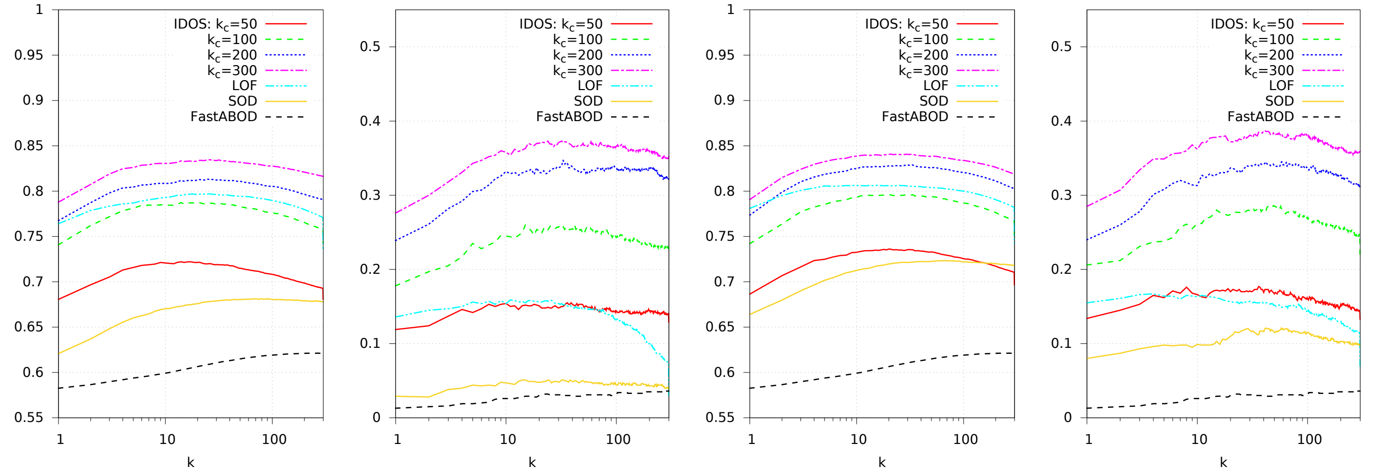
On the FMA dataset, from the results shown in Figures 4a to 4d, we see that LOF performs comparably to IDOS when the continuous ID is estimated over context sets of size k_c between 100 and 200. When this size is increased to more than 200 objects, IDOS outperforms all other competitors. The gain between k_c=200 and k_c=300 indicates that IDOS has the potential for even better performance for higher values. SOD and FastABOD both appear to struggle over this dataset, with the latter performing barely better than would a random ranking. The precision values achieved by IDOS are also significantly higher than those achieved by its competitors.

On the 10,000-dimensional Amazon dataset, in accordance with our expectations due to the size of its clusters (435), IDOS performs best for the larger choices of context set size k_c. The ROC AUC results in Figures 5a and 5c show that only SOD achieves comparable results,



(a) Euclidean and ROC AUC (b) Euclidean and precision (c) Arc cosine and ROC AUC (d) Arc cosine and precision

Fig. 3: Performance of IDOS, LOF, SOD, and FastABOD on the ALOI dataset for Euclidean and arc cosine distances, in terms of precision and ROC AUC.

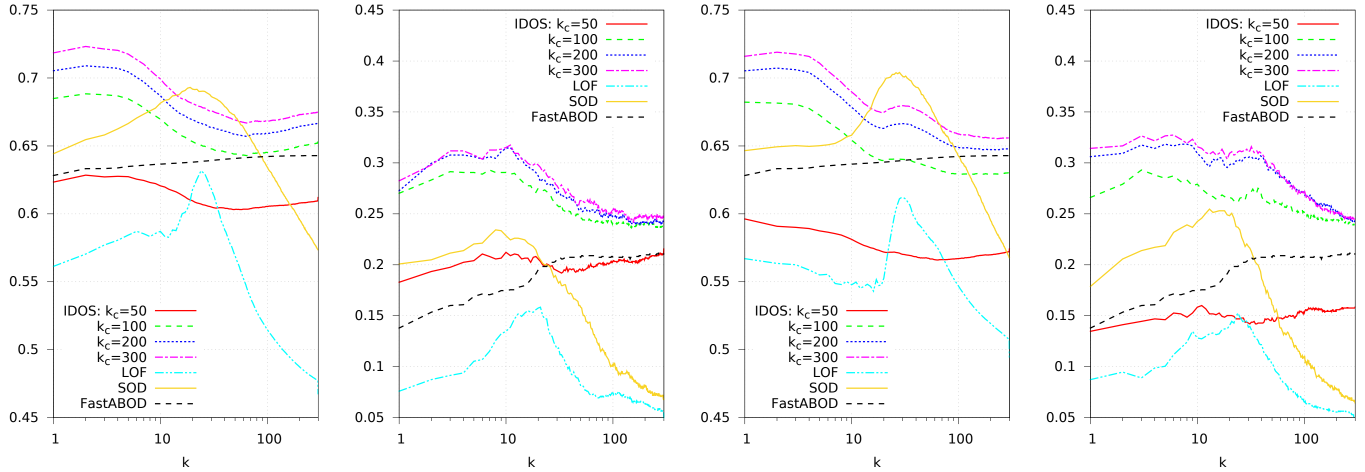


(a) Euclidean and ROC AUC (b) Euclidean and precision (c) Arc cosine and ROC AUC (d) Arc cosine and precision

Fig. 4: Performance of IDOS, LOF, SOD, and FastABOD on the FMA dataset for Euclidean and arc cosine distances, in terms of precision and ROC AUC.

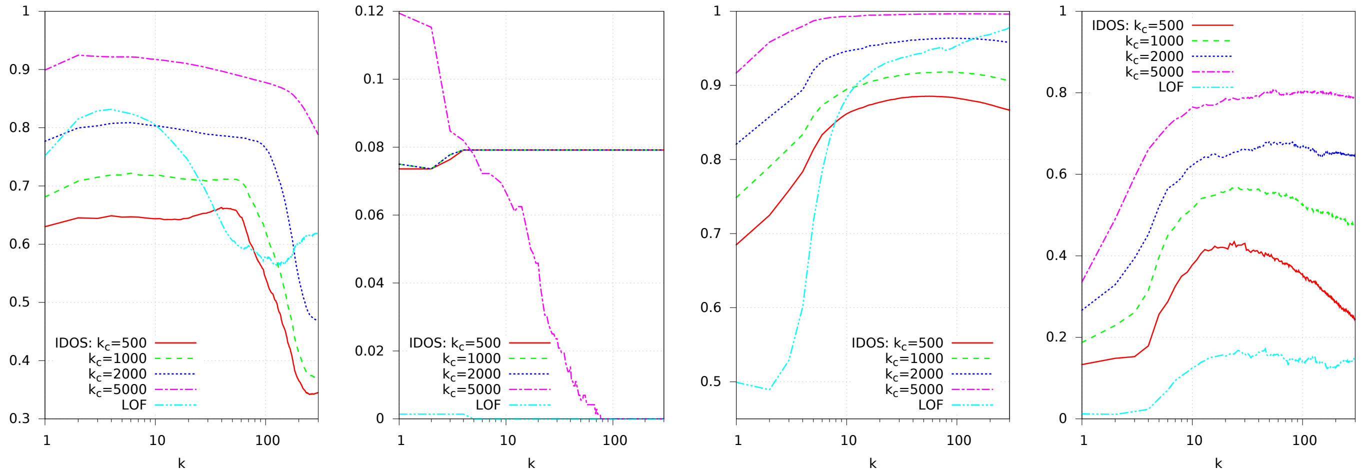
especially with arc cosine distance and a reference set size of $k \approx 20$. For most of the range of choices of k, the performance of IDOS dominates that of FastABOD, which in turn dominates that of LOF. The superiority of IDOS is even more clear from the plots of precision values (Figures 5b and 5d). However, the relatively low gain in performance between the choices k_c=200 and k_c=300 indicates that further increases in k_c will not likely result in a significant improvement in performance.

On the large and very high-dimensional Reuters dataset, FastABOD and SOD could not be tested. Both would need massive parallelism to accommodate their time complexities (quadratic in the dataset size, and linear in the dataset dimensionality). In addition, FastABOD requires the storage of a similarity matrix for efficient computation, which for the Reuters set would be at the expense of 80 GB of main memory over and above the already high memory requirements for the storage of the dataset. For LOF and IDOS, we computed the nearest neighbors required in a preprocessing step, and stored them for later use by both algorithms. In all cases the performance using the arc cosine similarity was far better than using Euclidean distance — not unexpectedly, given that the use of Euclidean distances on text data is not considered to be appropriate in practice [61]. In the experiments using arc cosine similarity, the large context set sizes required by IDOS are commensurate with the large dataset and cluster sizes. For IDOS, a context set size of 5000 (roughly 5% of the dataset) leads to near-perfect ROC AUC values of 0.9967, and a precision of 0.8069 (581 of 720 outliers found). Although LOF also achieves reasonably high ROC AUC results, its



(a) Euclidean and ROC AUC (b) Euclidean and Precision (c) Arc cosine and ROC AUC (d) Arc cosine and Precision

Fig. 5: Performance of IDOS, LOF, SOD, and FastABOD on the Amazon dataset for Euclidean and arc cosine distances, in terms of precision and ROC AUC.



(a) Euclidean and ROCAUC (b) Euclidean and Precision (c) Arc cosine and ROC AUC (d) Arc cosine and Precision

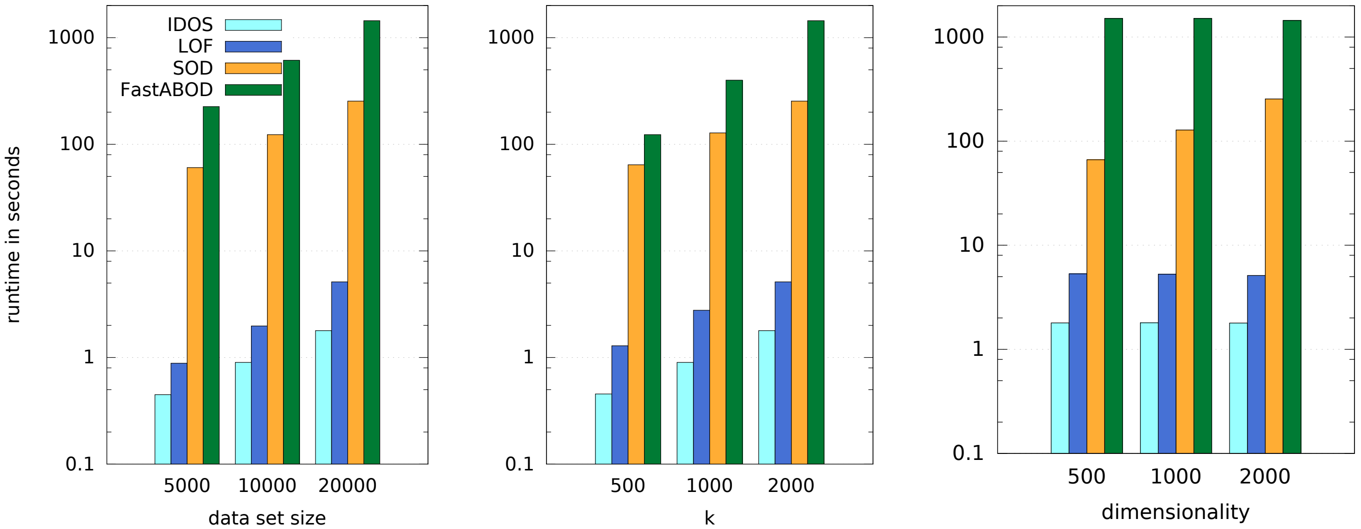
Fig. 6: Performance of IDOS and LOF on the Reuters dataset, for Euclidean and arc cosine distances, in terms of precision and ROC AUC.

performance is generally dominated by that of IDOS, particularly for the smaller values of the reference set $k = k_r$. However, the precision scores achieved by LOF are far lower than those achieved by IDOS.

4.4 Efficiency

We compare the CPU time of the algorithms. IDOS, LOF, SOD, and FastABOD, all implemented in the unified framework ELKI [56], were benchmarked at neighborhood sizes $k = 500$, 1000 and 2000 on synthetic datasets containing a single, normally distributed cluster of size $n = 5000$, 10,000 or 20,000, with representational dimension of $m = 500$, 1000 or 2000. In the case of IDOS, both parameters k_c and k_r were set to the same value, k. All tests were run on the same machine, equipped with an Intel Core i5-3570 CPU and 16GB RAM, without the use of any form of parallelism in the code. Every test was run 100 times to generate 100 scores (execution costs), from which the fastest and slowest 10% were discarded. From the remaining 80 scores, we report the arithmetic mean. We excluded the time required for the computations of shared nearest neighbors (SOD), k-nearest neighbor sets (IDOS, LOF), or the similarity matrix (FastABOD), as the efficiency of these steps depends on the employed approximation or indexing strategies, and is thus outside the scope of this paper.

As expected, when varying the dataset size, IDOS, LOF, and SOD all show a linear increase in running time. The computational cost of FastABOD implementation increases approximately quadratically (Figures 7a), which is consistent with its $O(n^2 + nk^2)$ asymptotic time complexity. Figure 7b shows the resulting increase in ex-



(a) Varying dataset size, with neighborhood size $k = k_r = 200$ and dimension $m = 2000$.

(b) Varying neighborhood size $k = k_r$, with dataset size $n = 20,000$ and dimension $m = 2000$.

(c) Varying dimension m, with dataset $n = 20,000$ and neighborhood size $k = k_r = 200$.

Fig. 7: CPU time comparison on synthetic datasets containing a single normally distributed cluster.

ecution time as the parameter k is doubled. Again, IDOS, LOF, and SOD all show a linear increase in running time, while the computational cost of FastABOD increases approximately quadratically (Figures 7b). The dimensionality has linear influence on the runtime of SOD, while for the other algorithms no significant changes are observed (Figure 7c). Overall, we find that FastABOD and SOD are considerably more time-consuming than LOF and IDOS. For lower dimensionalities, SOD usually requires less computation time than FastABOD, but still far more than LOF and IDOS. Of the two faster methods, LOF requires roughly double the execution time of IDOS.

To conclude, we note that while computing the results for the Reuters dataset shown in Figure 6, the gap between the running times of LOF and IDOS widened substantially as k increased. This is likely due to the need for LOF to compute $k + 1$ lookups of object k-distances for each lookup of an object k-distance by IDOS, during the construction of the outlier model.

5 Conclusion

We have presented a novel outlier scoring method, IDOS, that takes advantage of variations in local intrinsic dimensionality (ID) to distinguish between inliers and outliers. ID allows for inlier members of a subspace cluster to come to a better agreement with other members of the same cluster, and for better distinguishability from nonmembers. Local outliers tend to experience proximity to a cluster as an increase in the estimated value of their continuous ID.

As a comparison, we argued that the well-known LOF outlier score can be reinterpreted in light of the model of continuous intrinsic dimensionality introduced in [51]. In comparison with IDOS, LOF is revealed to have the potential for more variability in its assessment of local density within clusters (groups of inliers) than IDOS has in its assessment of local intrinsic dimensionality, which would make it harder to distinguish outliers in the vicinity of such clusters. These claims are borne out by experimental results presented for high- to very high-dimensional datasets (up to 107,756 dimensions), that show the superiority of our method over the state-of-the-art competitors LOF, ABOD and SOD, in terms of both effectiveness and efficiency.

Acknowledgments

Michael E. Houle acknowledges the financial support of JSPS Kakenhi Kiban (A) Research Grant 25240036, and the JST ERATO Kawarabayashi Large Graph Project.

References

[1] E. M. Knorr and R. T. Ng, “A unified notion of outliers: Properties and computation,” in *Proceedings of the 3rd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Newport Beach, CA*, 1997, pp. 219–222.

[2] S. Ramaswamy, R. Rastogi, and K. Shim, “Efficient algorithms for mining outliers from large data sets,” in *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX*, 2000, pp. 427–438.

[3] F. Angiulli and C. Pizzuti, “Fast outlier detection in high dimensional spaces,” in *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD), Helsinki, Finland*, 2002, pp. 15–26.

[4] M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, “LOF: Identifying density-based local outliers,” in *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX*, 2000, pp. 93–104.

[5] E. Schubert, A. Zimek, and H.-P. Kriegel, “Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection,” *Data Mining and Knowledge Discovery*, vol. 28, no. 1, pp. 190–237, 2014.

[6] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statistical Analysis and Data Mining*, vol. 5, no. 5, pp. 363–387, 2012.

[7] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *Proceedings of the 7th International Conference on Database Theory (ICDT), Jerusalem, Israel*, 1999, pp. 217–235.

[8] V. Pestov, "On the geometry of similarity search: Dimensionality curse and concentration of measure," *Information Processing Letters*, vol. 73, no. 1–2, pp. 47–51, 2000.

[9] D. François, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 7, pp. 873–886, 2007.

[10] D. R. Karger and M. Ruhl, "Finding nearest neighbors in growth-restricted metrics," in *Proceedings of the 34th annual ACM symposium on Theory of computing (STOC), Montreal, QC, Canada*, 2002, pp. 741–750.

[11] M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Can shared-neighbor distances defeat the curse of dimensionality?" in *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management (SSDBM), Heidelberg, Germany*, 2010, pp. 482–500.

[12] M. E. Houle, H. Kashima, and M. Nett, "Generalized expansion dimension," in *ICDM Workshop Practical Theories for Exploratory Data Mining (PTDM)*, 2012, pp. 587–594.

[13] W. Jin, A. Tung, and J. Han, "Mining top-n local outliers in large databases," in *Proceedings of the 7th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Francisco, CA*, 2001, pp. 293–298.

[14] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Taipei, Taiwan*, 2002, pp. 535–548.

[15] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," in *Proceedings of the 19th International Conference on Data Engineering (ICDE), Bangalore, India*, 2003, pp. 315–326.

[16] W. Jin, A. K. H. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore*, 2006, pp. 577–593.

[17] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "LoOP: local outlier probabilities," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, China*, 2009, pp. 1649–1652.

[18] E. Schubert, A. Zimek, and H.-P. Kriegel, "Generalized outlier detection with flexible kernel density estimates," in *Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia, PA*, 2014, pp. 542–550.

[19] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Outlier detection in axis-parallel subspaces of high dimensional data," in *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand*, 2009, pp. 831–838.

[20] E. Müller, M. Schiffer, and T. Seidl, "Adaptive outlierness for subspace outlier ranking," in *Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM), Toronto, ON, Canada*, 2010, pp. 1629–1632.

[21] ——, "Statistical selection of relevant subspace projections for outlier ranking," in *Proceedings of the 27th International Conference on Data Engineering (ICDE), Hannover, Germany*, 2011, pp. 434–445.

[22] H. V. Nguyen, V. Gopalkrishnan, and I. Assent, "An unbiased distance-based outlier detection approach for high-dimensional data," in *Proceedings of the 16th International Conference on Database Systems for Advanced Applications (DASFAA), Hong Kong, China*, 2011, pp. 138–152.

[23] F. Keller, E. Müller, and K. Böhm, "HiCS: high contrast subspaces for density-based outlier ranking," in *Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, DC*, 2012.

[24] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Outlier detection in arbitrarily oriented subspaces," in *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM), Brussels, Belgium*, 2012, pp. 379–388.

[25] X. H. Dang, I. Assent, R. T. Ng, A. Zimek, and E. Schubert, "Discriminative features for identifying and interpreting outliers," in *Proceedings of the 30th International Conference on Data Engineering (ICDE), Chicago, IL*, 2014, pp. 88–99.

[26] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV*, 2008, pp. 444–452.

[27] A. Zimek, R. J. G. B. Campello, and J. Sander, "Ensembles for unsupervised outlier detection: Challenges and research questions," *ACM SIGKDD Explorations*, vol. 15, no. 1, pp. 11–22, 2013.

[28] T. de Vries, S. Chawla, and M. E. Houle, "Density-preserving projections for large-scale local anomaly detection," *Knowledge and Information Systems (KAIS)*, vol. 32, no. 1, pp. 25–52, 2012.

[29] N. Pham and R. Pagh, "A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data," in *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Beijing, China*, 2012.

[30] Y. Wang, S. Parthasarathy, and S. Tatikonda, "Locality sensitive outlier detection: A ranking driven approach," in *Proceedings of the 27th International Conference on Data Engineering (ICDE), Hannover, Germany*, 2011, pp. 410–421.

[31] J. Bruske and G. Sommer, "Intrinsic dimensionality estimation with optimally topology preserving maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 5, pp. 572–575, 1998.

[32] K. Fukunaga and D. R. Olsen, "An algorithm for finding intrinsic dimensionality of data," *IEEE Transactions on Computers*, vol. C-20, no. 2, pp. 176–183, 1971.

[33] E. Pettis, T. Bailey, A. Jain, and R. Dubes, "An intrinsic dimensionality estimator from nearest-neighbor information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 25–37, 1979.

[34] P. Verveer and R. Duin, "An evaluation of intrinsic dimensionality estimators," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 1, pp. 81–86, 1995.

[35] C. Bouveyron, G. Celeux, and S. Girard, "Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA," *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1706–1713, Oct. 2011.

[36] I. Jolliffe, *Principal Component Analysis*. Springer-Verlag, 1986.

[37] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[38] B. Schölkopf, A. J. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[39] J. Venna and S. Kaski, "Local multidimensional scaling," *Neural Networks*, vol. 19, no. 6–7, pp. 889–899, 2006.

[40] J. Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning," *Neural Networks*, vol. 7, no. 1, pp. 113–127, 1994.

[41] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[42] F. Camastra and A. Vinciarelli, "Estimating the intrinsic dimension of data with a fractal-based method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 10, pp. 1404–1407, 2002.

[43] C. Faloutsos and I. Kamel, "Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension," in *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Minneapolis, MN*, 1994.

[44] A. Gupta, R. Krauthgamer, and J. R. Lee, "Bounded geometries, fractals, and low-distortion embeddings," in *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS), Cambridge, MA*, 2003, pp. 534–543.

[45] V. Vapnik, *The nature of statistical learning theory*. Springer, 2000.

[46] P. Larrañaga and J. A. Lozano, *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Springer, 2002, vol. 2.

[47] E. Levina and P. J. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), Vancouver, BC*, 2004.

[48] A. Beygelzimer, S. Kakade, and J. Langford, "Cover trees for nearest neighbors," in *Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA*, 2006, pp. 97–104.

[49] M. E. Houle, X. Ma, M. Nett, and V. Oria, "Dimensional testing for multi-step similarity search," in *Proceedings of the 12th IEEE*

International Conference on Data Mining (ICDM), Brussels, Belgium, 2012, pp. 299–308.

[50] M. E. Houle and M. Nett, "Rank cover trees for nearest neighbor search," in *Proceedings of the 6th International Conference on Similarity Search and Applications (SISAP), A Coruña, Spain*, 2013, pp. 16–29.

[51] M. E. Houle, "Dimensionality, Discriminability, Density & Distance Distributions," in *International Conference on Data Mining Workshops (ICDMW)*. IEEE Computer Society, 2013, pp. 468–473.

[52] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, and M. Nett, "Estimating continuous intrinsic dimensionality," NII, Japan, TR 2014-001E, Mar. 2014.

[53] B. M. Hill, "A simple general approach to inference about the tail of a distribution," *Annals of Statistics*, vol. 3, no. 5, pp. 1163–1174, 1975.

[54] R. Huisman, K. G. Koedijk, C. J. M. Kool, and F. Palm, "Tail-index estimates in small samples," *J. Bus. Econ. Stat.*, vol. 19, no. 2, pp. 208–216, 2001.

[55] P. Hall, "Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems," *J. Multivariate Anal.*, vol. 32, no. 2, pp. 177–203, 1990.

[56] E. Achtert, H.-P. Kriegel, E. Schubert, and A. Zimek, "Interactive data mining with 3D-Parallel-Coordinate-Trees," in *Proceedings of the ACM International Conference on Management of Data (SIGMOD), New York City, NY*, 2013, pp. 1009–1012.

[57] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "The Amsterdam Library of Object Images," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112, 2005.

[58] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

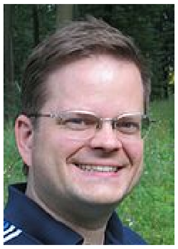
[59] E. Pampalk, A. Rauber, and D. Merkl, "Content-based organization and visualization of music archives," in *Proceedings of the 10th ACM International Conference on Multimedia (ACM MM), Juan les Pins, France*, 2002, pp. 570–579.

[60] M. R. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views-an application to multilingual text categorization," in *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS), Vancouver, BC*, vol. 1, 2009, pp. 28–36.

[61] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining," in *LDV Forum*, vol. 20, no. 1, 2005, pp. 19–62.

Jonathan von Brünken studied Computer Science at the Ludwig-Maximilians-Universität München, Germany and completed an internship at the National Institute of Informatics (NII) in Tokyo, Japan. After obtaining his Diploma degree in March 2014 he started working for the PAYBACK GmbH, part of the American Express Group, as a data analyst for marketing and market research at the customer insights division. Current personal research interests are in the fields of outlier detection, clustering and classification, especially for high-dimensional and multimedia data.

Michael E. Houle obtained his PhD degree from McGill University in 1989, in the area of computational geometry. Since then, he developed research interests in algorithmics, data structures, and relational visualization, first as a research associate at Kyushu University and the University of Tokyo in Japan, and from 1992 at the University of Newcastle and the University of Sydney in Australia. From 2001 to 2004, he was a Visiting Scientist at IBM Japan's Tokyo Research Laboratory, where he first began working on approximate similarity search and shared-neighbor clustering methods for data mining applications. Since then, his research interests have expanded to include dimensionality and scalability in the context of fundamental data mining tasks such as search, clustering, classification, and outlier detection. He has co-authored award-winning conference papers on outlier detection (Best Research Paper Award at IEEE ICDM 2010) and similarity search (Best Paper Award at SISAP 2014). Currently, he is a Visiting Professor at the National Institute of Informatics (NII), Japan.



Arthur Zimek is a Privatdozent in the database systems and data mining group at the Ludwig-Maximilians-Universität (LMU) München, Germany. 2012–2013 he was a postdoctoral fellow in the department for Computing Science at the University of Alberta, Edmonton, Canada. He finished his Ph.D. thesis at LMU in informatics on "Correlation Clustering" in summer 2008, and received for this work the "SIGKDD Doctoral Dissertation Award (runner-up)" in 2009. Together with his co-authors, he received the "Best Paper Honorable Mention Award" at SDM 2008 and the "Best Demonstration Paper Award" at SSTD 2011. His current research interests are in data mining with a focus on clustering and outlier detection, methods and evaluation, ensemble methods, and high dimensional data.