REVIEW

A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data

Arthur Zimek[1]*, Erich Schubert[2] and Hans-Peter Kriegel[2]

[1]*Department of Computing Science, University of Alberta, Edmonton, AB, Canada T6G 2E8*

[2]*Institute for Informatics, Ludwig-Maximilians Universität München, Germany*

Received 30 January 2012; revised 22 June 2012; accepted 2 August 2012
DOI:10.1002/sam.11161
Published online 27 August 2012 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: High-dimensional data in Euclidean space pose special challenges to data mining algorithms. These challenges are often indiscriminately subsumed under the term 'curse of dimensionality', more concrete aspects being the so-called 'distance concentration effect', the presence of irrelevant attributes concealing relevant information, or simply efficiency issues. In about just the last few years, the task of unsupervised outlier detection has found new specialized solutions for tackling high-dimensional data in Euclidean space. These approaches fall under mainly two categories, namely considering or not considering subspaces (subsets of attributes) for the definition of outliers. The former are specifically addressing the presence of irrelevant attributes, the latter do consider the presence of irrelevant attributes implicitly at best but are more concerned with general issues of efficiency and effectiveness. Nevertheless, both types of specialized outlier detection algorithms tackle challenges specific to high-dimensional data. In this survey article, we discuss some important aspects of the 'curse of dimensionality' in detail and survey specialized algorithms for outlier detection from both categories. © 2012 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 5: 363–387, 2012

Keywords: curse of dimensionality; anomalies in high-dimensional data; outlier detection in high-dimensional data; approximate outlier detection; subspace outlier detection; correlation outlier detection

1. INTRODUCTION

High-dimensional data poses special challenges for data mining in general and outlier detection in particular. Though in recent years, several surveys on outlier detection have been published (see refs. [1–8], to name a few), the difficulties in high-dimensional data and specialized approaches in this area have not been sketched in any of those (though, notably, a recent textbook edition sketches three example approaches [9]). In fact, most approaches to this problem have been proposed just recently in the past two or three years. Since the development of unsupervised methods for outlier detection in high-dimensional data in Euclidean space appears to be an emerging topic, this survey is specialized on this topic. We hope to help researchers working in this area to become aware of other approaches, to understand the advancements and the lasting problems, and to better identify the true achievements of the various approaches. Though a lot of variants for outlier detection are around, such as supervised versus unsupervised or specialized approaches for specific data types (such as item sets, time series, sequences, categorical data, see refs. 7,10 for an overview), here we focus on unsupervised approaches for numerical data in Euclidean space.

Albeit the infamous 'curse of dimensionality' has been credited for many problems and has indiscriminately been used as a motivation for many new approaches, we should try to understand the problems occurring in high-dimensional data in more detail. For example, there is a widespread mistaken belief that every point in high-dimensional space is an outlier. This—misleading, to say the least—statement has been suggested as a motivation for the first approach specialized to outlier detection

Correspondence to: Arthur Zimek (zimek@ualberta.ca)

in subspaces of high dimensional data [11], recurring superficially to a fundamental paper on the 'curse of dimensionality' by Beyer et al. [12]. Alas, as we will discuss in the following section, it is not as simple as that.

Indeed, this often cited but less often well-understood study [12], has been reconsidered recently by several researchers independently in different research areas. We will, therefore, begin our survey by inspecting truths and myths associated with the 'curse of dimensionality', study a couple of its effects that may be relevant for outlier detection, and discuss the findings of the renewed interest in this more than 10-year-old study (Section 2). Afterwards, we will discuss different families of outlier detection approaches concerned with high-dimensional data: first, approaches that treat the issues of efficiency and effectiveness in high-dimensional data without specific interest in a definition of outliers with respect to subspaces of the data (Section 3), and second, those that search for outliers specifically in subspaces of the data space (Section 4). In Section 5, we will comment on some open-source tools providing implementations of outlier detection algorithms, and remark on the difficulties of understanding and evaluating the results of outlier detection algorithms. Finally, in Section 6, we conclude the paper.

2. THE CURSE OF DIMENSIONALITY

The 'curse of dimensionality' has motivated a lot of research in the area of databases due to its impacts on similarity search [13–21]. Yet still there are open questions, unresolved issues, and even the known results have found less attention in research than appropriate. Here we discuss some effects and influential characteristics of high-dimensional data and relate these to issues for outlier detection in high-dimensional data.

2.1. Concentration of Distances—Concentration of Outlier Scores

Let us recall the basic statements of the fundamental study of Beyer et al. [12]. Their key result states the following:

Assumption *The ratio of the variance of the length of any point vector (denoted by $\|X_d\|$) with the length of the mean point vector (denoted by $E[\|X_d\|]$) converges to zero with increasing data dimensionality.*

This assumption covers a broad range of data distributions and distance measures (generally: all L_p-norms with $p \geq 1$).

Consequence *The proportional difference between the farthest-point distance $D_{\max}$ and the closest-point distance $D_{\min}$ (the relative contrast) vanishes.*

Formally:

$$\text{If } \lim_{d\to\infty} \text{var}\left(\frac{\|X_d\|}{E[\|X_d\|]}\right) = 0, \quad \text{then } \frac{D_{\max} - D_{\min}}{D_{\min}} \to 0.$$

Intuitively, under the given assumption the relative contrast between near and far neighbors does diminish as the dimensionality increases. This *concentration effect* of the distance measure hence reduces the usability of the measure for discrimination between near and far neighbors [22].

The influence of different values for p in L_p-norms on the concentration effect has been studied in refs. 23,24. In ref. 23, the authors showed by means of an analytic argument that L_1 and L_2 are the only integer norms useful for higher dimensions. In addition, they studied the use of projections for discrimination, the effectiveness of which depended on localized dissimilarity measures that did not satisfy the symmetry and triangle inequality conditions of distance metrics. In ref. 24, fractional L_p distance measures (with $0 < p < 1$) have been studied in a similar context. The authors provide evidence supporting the contention that smaller values of p offer better results in higher dimensional settings. Their result, however, is valid only for uniformly distributed data [25]. The concentration of the cosine similarity has been studied in ref. 26.

This does, however, only tell part of the story. A change in the value range for the distances does not come unexpected. The maximum distance from the origin in the unit cube is $\sqrt[p]{d}$ for L_p-norms, the average distance converges to $\sqrt[p]{d} \cdot \frac{1}{\sqrt{3}}$ with increasing d. For a normalized maximum distance, for example, in the unit-hypercube, this comes down to $\frac{1}{\sqrt{3}} \approx 0.577$. So at first sight, it might be feasible to counter these effects by rescaling the distances appropriately.

The resulting effects can be seen in Fig. 1 (compare to ref. 25, which shows similar results without normalization). In this figure, we plot some characteristics of the distribution of vector length values for vectors uniformly distributed in the unit cube over increasing dimensionality d. The observed length of each vector is normalized by $\frac{1}{\sqrt{d}}$. Plotted are mean and standard deviation of the observed lengths as well as the actually observed minimal (maximal) length in each dimensionality. While the observed distances now have a comparable average value, there is still a loss in relative contrast. On this data set of a sample of 10^5 instances, the curse (more explicitly, the effect of distance concentration) is now clearly visible: the standard deviation disappears already at 10 dimensions, and the observed actual minimum and maximum shrink rapidly with increasing dimensionality. This is actually not very

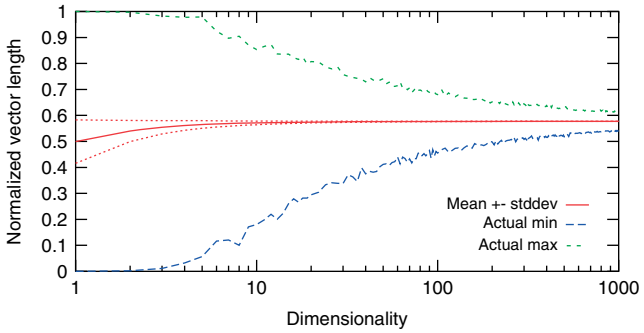

Fig. 1 Normalized Euclidean length, uniform [0,1], logarithmic scale for dimensionality d. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

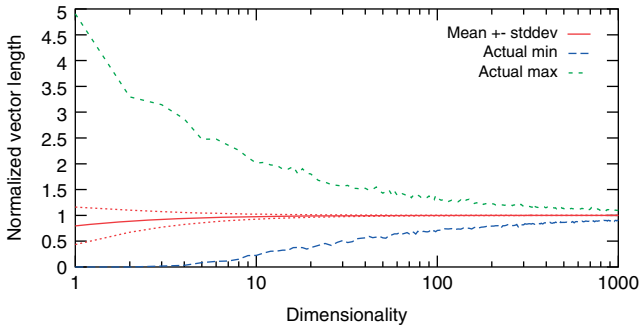

Fig. 2 Normalized Euclidean length, Gaussian [0,1], logarithmic scale for dimensionality d. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

surprising—in this i.i.d. setting, we are observing the central limit theorem with a shrinking variance. If we do not assume a uniform distribution but instead use standard normal (Gaussian) distributions, the result looks virtually the same (Fig. 2), except that the average now converges to the standard deviation of 1. The distance from the origin (the length of the vector) is a certain simplification; however, the same effects occur when looking at pairwise distances for any two points in the data set, as we can observe in Figs. 3 and 4, except with different average values.

For many distance-based data mining algorithms, this loss of numerical contrast is a major problem. To study this effect, we generate two series of very basic distributions. All points are drawn from the same distribution, so in this sense, there are no outliers in either data set series. The first series is uniformly distributed on the interval $[0 \ldots 1]$, the second series is standard normal distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$. Since the second distribution contains objects on the long tail of the distribution, some of these can be considered as outliers in the common sense of objects with a low probability density function (pdf). For example, the kNN distances as used for kNN outlier detection [27] for uniformly

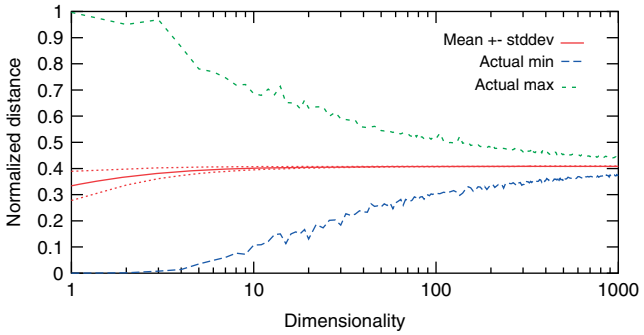

Fig. 3 Normalized Euclidean pairwise distance, uniform [0,1], logarithmic scale for dimensionality d. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

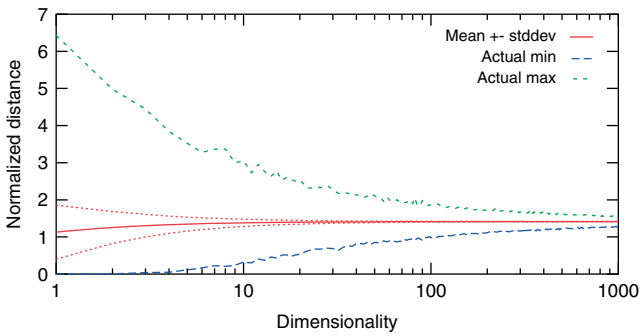

Fig. 4 Normalized Euclidean pairwise distance, Gaussian [0,1], logarithmic scale for dimensionality d. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

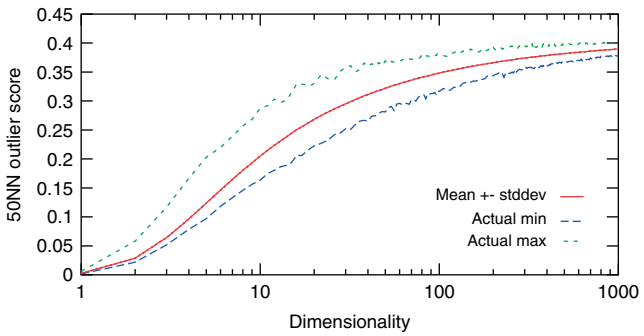

Fig. 5 Normalized Euclidean 50NN-distance, uniform [0,1], logarithmic scale for dimensionality d. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

distributed data (Fig. 5) and normally distributed data (Fig. 6) become virtually the same in high dimensionality, while in low dimensionality, the maximum distance was clearly different for the two distributions. But already at just 10 dimensions, even the uniform distribution seems to produce outliers for this method that are clearly larger

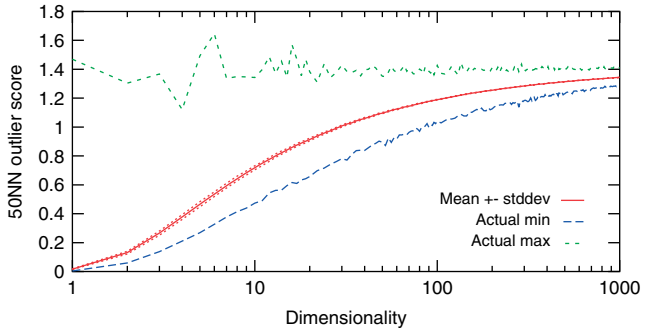

Fig. 6 Normalized Euclidean 50NN-distance, Gaussian [0,1], logarithmic scale for dimensionality d. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

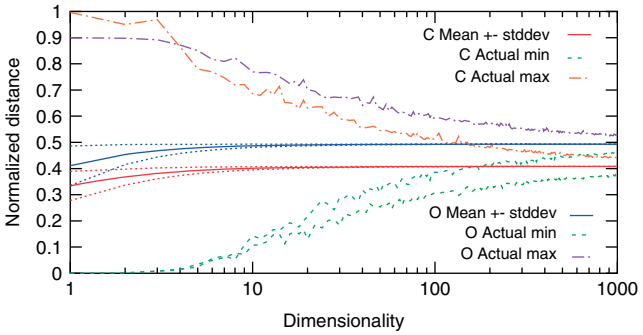

Fig. 9 Normalized Euclidean pairwise distance, uniform [0,1], 1 outlier, logarithmic scale for d. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

than the mean distance ($0.28740 \gg \mu = 0.20512 \pm \sigma = 0.00027$). The more advanced method LOF [28] is less likely to detect false outliers at this medium dimensionality, but also quickly fails to recognize objects as atypical by their distance (Figs. 7 and 8).

Up to now, the complete data set was generated by a single distribution. As such, it does not contain outliers in the sense of objects that actually are generated by a different generating mechanism. We now modify the datasets by manually placing an outlier at a constant position.

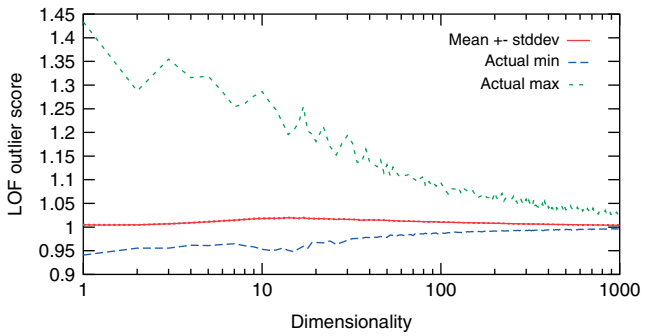

Fig. 7 LOF with $k = 50$, uniform [0,1], logarithmic scale for dimensionality d. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

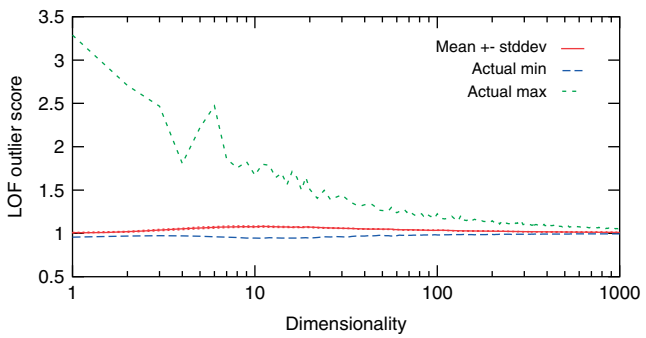

Fig. 8 LOF with $k = 50$, Gaussian [0,1], logarithmic scale for dimensionality d. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

For the uniform distribution, we position the manual outlier at 0.9 in all dimensions. While this is within the domain of the uniform distribution, the point stands out clearly in high dimensions, as it can be seen in Fig. 9. The reason for this outstanding behavior is that, for this object, all attributes are, by construction, strongly correlated. The point has an average normalized distance to the i.i.d. objects of the uniform distribution that is approximately 0.1 larger than the corresponding values of the remainder of the data. At around 700 dimensions however, the closest neighbor of this outlier is farther away from the outlier than from any other object in the uniform distribution! We see here that this kind of outlierness becomes not less but ever more prominent when increasing the dimensionality (as long as the dimensions add information).

A similar effect can be seen in the data generated by (i.i.d.) Gaussian distributions in all dimensions. For Fig. 10, we placed the manual outlier at the fixed value of 2 in every dimension, resulting in a point that is at 2σ in every dimension and thus well inside the one-dimensional distributions. At one dimension, there exist larger distances within the Gaussian distribution than to this outlier. With increasing dimensionality, however, it remains at a constant distance to the cluster mean, while the other distances concentrate. At around 100 dimensions, again the nearest neighbor of the manual outlier is farther away from the outlier than the maximum observed distance inside the Gaussian distribution.

Applying outlier detection on these two data sets produces the expected results, as we exemplify for LOF in Figs. 11 and 12. At low dimensionality, the outlier remains well hidden inside the distribution of LOF values for the clustered points. However, with increasing dimensionality the outlier achieves a score well distinguishable from the scores of the cluster objects (the standard deviations are

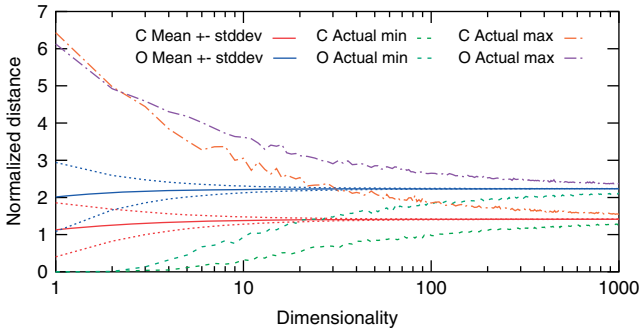

Fig. 10 Normalized Euclidean pairwise distance, Gaussian [0,1], 1 outlier, logarithmic scale for d. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

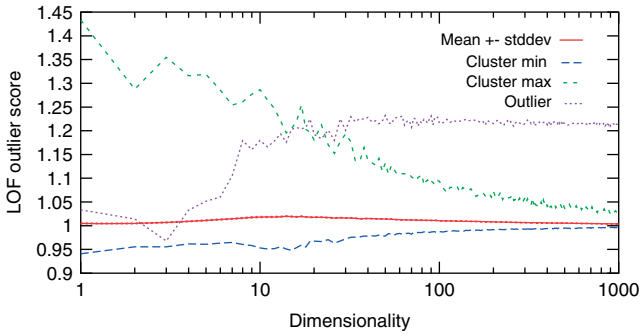

Fig. 11 LOF with $k = 50$, uniform [0,1], 1 outlier, logarithmic scale for d. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

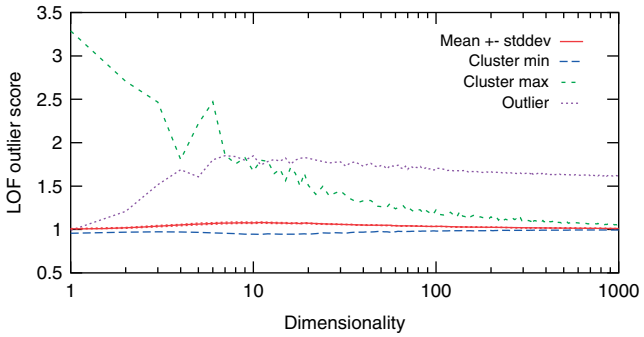

Fig. 12 LOF with $k = 50$, Gaussian [0,1], 1 outlier, logarithmic scale for d. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

barely visible, but the minimum and maximum observed LOF values from the data set are clearly bypassed already at 20–30 dimensions).

All this demonstrates that the concentration effect *per se*—although often seen this way—is *not* the main problem for outlier detection in high-dimensional data. On the contrary, for points that deviate in every attribute from the usual data distribution, the outlier characteristics just become even stronger and more pronounced with increasing dimensionality. And this observation is all but unexpected since adding another dimension where the behavior of the outlier and the remainder of usual data objects is different means adding more information that helps discriminating the different characteristics.

2.2. Relevant and Irrelevant Attributes

All the studies on the concentration effect we sketched above generally assumed that the full data set followed a single data distribution, subject to certain restrictions. In fact, when the data follows a mixture of distributions, the concentration effect is not always observed. Instead, in such cases, distances between members of different distributions may not necessarily tend to the global mean as the dimensionality increases. As briefly noted in refs. 12,29, if a data set is composed of many natural groupings or clusters, each following their own distribution, then the concentration effect will typically be less severe for queries based on points within a cluster of similar points generated according to the same mechanism, especially when the clusters are well separated.

The fundamental differences between singly distributed data and multiply distributed data are already discussed in detail in ref. 30. The authors demonstrate that nearest-neighbor queries are both theoretically and practically meaningful if the search is limited to objects from the same cluster (distribution) as the query point (i.e., any point belonging to the same cluster is considered a valid answer to a nearest-neighbor query), and other clusters are well separated from the cluster in question. The key concept is that of *pairwise stability* of clusters, which is said to hold whenever the mean distance between points of different clusters dominates the mean distance between points belonging to the same cluster. When the clusters are pairwise stable, for any point belonging to a given cluster, its nearest neighbors tend to belong to the same cluster. Here, a nearest-neighbor query of size on the order of the cluster size *can* be considered meaningful, whereas differentiation between nearest and farthest neighbors within the same cluster may still be meaningless. Note that for many common distributions these considerations may remain valid even as the dimension d tends to infinity: for example, two Gaussian distributions with widely separated means may find that their separability *improves* as the data dimension increases. However, it should also be noted that these arguments are based on the assumption that all dimensions bear information relevant to the different clusters, classes, or distributions. Depending on the ratio of relevant versus irrelevant attributes, and on the actual separation of sets of points belonging to different distributions, irrelevant attributes in a data set may impede

the separability of different distributions and thus have the potential for eventually rendering nearest neighbor query results meaningless.

The observations of ref. 30, and the important distinction between the effects of relevant and irrelevant attributes, both seem to have received little if any attention in the research literature. Despite the demonstrated deficiency of conventional L_p norms for high-dimensional data, a plethora of work based on the Euclidean distance has been dedicated to clustering strategies, which appear to be effective in practice to varying degrees for high-dimensional data [31–34]. Many heuristics have been proposed or evaluated for clustering [35–42], outlier detection [11,43–45], and indexing or similarity search [46–51] that seek to mitigate the effects of the curse of dimensionality. While some of these strategies, such as projected or subspace clustering or subspace outlier detection, do recognize implicitly the effect of relevant versus irrelevant attributes for a cluster, all these papers (as well as others) abstain from discussing these effects, let alone studying them in detail. In particular, the concept of pairwise stability of clusters as introduced in ref. 30 has not been taken into account in any of these papers. Although their underlying data models do generally assume (explicitly or otherwise) different underlying mechanisms for the formation of data groupings, they motivate their new approaches with only a passing reference to the curse of dimensionality. Also in a recent study on the 'distance compression effect' [52], although they discuss other interesting aspects, the importance of a cluster structure for the curse to apply or to not apply has been missed. Hence they fail to convincingly explain their (allegedly surprising) finding that L_p metrics seem to perform better (in their experiments) in high-dimensional data than in low-dimensional data. Considering that a cluster structure will become stronger with an increasing number of relevant dimensions (i.e., attributes actually contributing information for the cluster separation), this observation is far from surprising—see the extensive experiments in ref. 53 and the related reasoning in refs. [54–56].

A more general picture has been drawn by Durrant and Kabán [57]. They show that the correlation between attributes is an important effect for avoiding the concentration of distances. Correlated attributes will also result in an intrinsic dimensionality that is considerably lower than the representational dimensionality, an effect that also led to opposing the curse of dimensionality with the 'self-similarity blessing' [58]. Contrary to the conjecture in refs. 12, 25; however, Durrant and Kabán [57] show that a low intrinsic dimensionality alone does not suffice to avoid the concentration, but essentially the correlation between attributes needs to be strong enough. Let us note that, actually, different aspects like a strong cluster-structure of data [30] and low intrinsic dimensionality [58] may be merely symptoms for a strong, though possibly latent, correlation between attributes. Durrant and Kabán also identify irrelevant dimensions as the core of the problem of the curse. In essence, the ratio between noise (e.g., irrelevant attributes or additive noise masking information in relevant attributes) and (latent) correlation in the attributes of a dataset will determine whether asking for the 'nearest neighbor' actually is meaningful.

Let us modify the experimental setup used before slightly in order to study the effect of relevant versus irrelevant attributes on outlier detection: we fix the dimensionality to $d = 100$. Instead, we vary the number of attributes d' where the outlier is set to its fixed value, and draw the remaining attributes from the random distribution that is also used for the usual data. At $d' = 0$, the outlier will be distributed identically as the remainder of the data, while at $d' = d$ all attributes will be set as before. Figure 13 shows the uniformly distributed setting, while Fig. 14 shows the normally distributed setting. With increasing d' (along the x-axis), the score of the outlier becomes more and more prominent. With a low portion of relevant attributes, the outlier is well hidden in the data set, but once d' is about

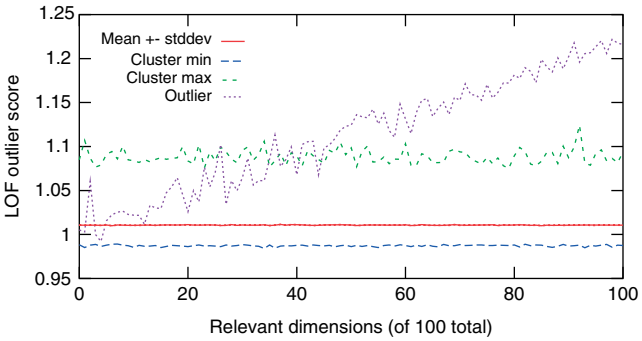

Fig. 13 LOF with $k = 50$, Uniform [0,1], 1 outlier. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

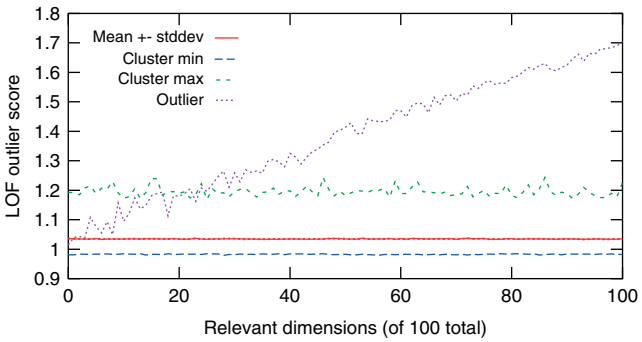

Fig. 14 LOF with $k = 50$, Gaussian [0,1], 1 outlier. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

30–40 (just in this setting, this is not a general rule of thumb), it becomes the highest scoring outlier.

This is the key motivation for subspace outlier detection: if we have too many irrelevant (noise) dimensions in the data set, the outlier can easily be masked. Once we, however, choose mostly relevant attributes (or projections), it becomes much more detectable. The challenge is to choose the right subspace.

That the relevance of attributes may be related to certain subgroups of the data has been recognized in applications like gene expression analysis [59], a problem setting that also motivated research on specialized methods for subspace clustering [31–33,60]. Similar traits can be found in areas like multiview clustering or, sometimes, alternative clustering or ensemble clustering [61]. For subspace outlier detection, a similar reasoning motivates the identification of objects that are outlying with respect to certain attributes, and this set of relevant attributes may differ from object to object. On the basis of such considerations, only few methods have been proposed so far. However, the interest in this topic seems to be increasing lately, which is a major motivation for this survey. We will discuss all methods known in the literature for subspace outlier detection in Section 4.

2.3. Discrimination Versus Ranking of Values

The distinction between relevant and irrelevant attributes in sight of increasing dimensionality has been studied recently in ref. 53 for vector data and cosine distance along with L_p norms as well as in ref. 62 for time series data and related specialized distance measures. The focus of these studies was to experimentally evaluate (and in effect to mostly confirm) the advantage of shared-neighbor distance measures in high-dimensional data, that was only assumed before. However, they can also serve as an investigation of the effects of relevant versus irrelevant attributes in data separation, based on distance *rankings* discerning between objects belonging to the same distribution versus objects belonging to some other distribution.

So, still, the differentiation between near and far points (the latter possibly being outliers) seems feasible by means of a *ranking* of distance values. Yet judging similarity or dissimilarity (as well as outlierness) by means of absolute distance values becomes more and more difficult in higher dimensional space. Let us illustrate this with another aspect of high-dimensional data exemplified for Euclidean distances (L_2-norms), the counter-intuitive behavior of the volume of a hypersphere with increasing dimensionality. This effect does not even depend on specific data distributions but on the mere dimensionality of the data space. In Fig 15, we depict the fundamental effect of this counter-intuitive behavior, the volume of a hypersphere

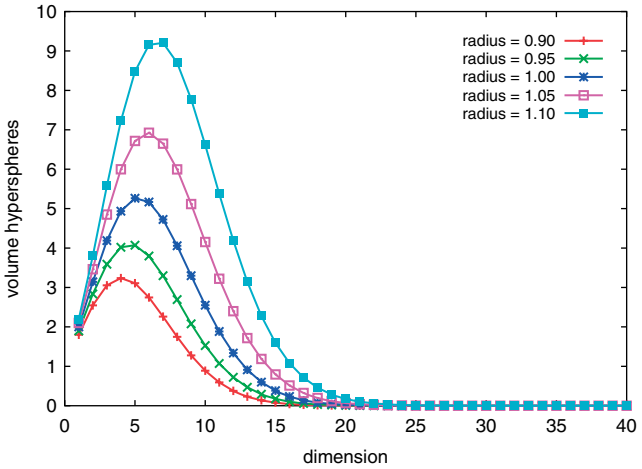

Fig. 15 Volumes of hyper-spheres. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

over a range of dimensionalities. For the unit-hypersphere (radius equals 1), we see the volume increasing from dimension 1 (a segment of length 2), dimension 2 (the volume equals π) to dimension 5. After that, on first sight surprisingly, the volume decreases again and soon approaches 0. The volume of a 20-dimensional hypersphere with a radius of 1 is almost 0. The same behavior is given with spheres of slightly varied radii. A small change of the radius moves the peak in altitude and position, but the general behavior remains the same. It would be rather naïve, though, to just interpret this as 'decreasing volume' since it is meaningless to compare a volume $unit^x$ with a volume $unit^y$, where $x \neq y$, just as we cannot say a square meter was more or less than half a cubic meter.

A meaningful interpretation of this plot is, however, easily given, if we interpret the values as the ratio of the volume of a hypersphere versus the volume of the unit hypercube (side length of 1). This way, also the behavior becomes understandable: for any radius, with increasing dimensionality the corners of the unit hypercube will eventually emerge toward the space outside the sphere since the diameter between the farthest opposite corners increases by $\sqrt{d}$, that is, unbounded, albeit slowly, while the radius of the sphere, how large it ever may be, remains constant. A more informal way of putting it is: high-dimensional spaces consist mostly of corners.

The second plot (Fig. 16) shows such ratios[1] for hyperspheres with radius ≤ 0.5 vs. the volume of the

[1] Note that, as the ratio in comparison to the unit hypercube is obtained dividing the volume by 1, the values of the y-axis in this plot are just interpreted differently, but are not changed numerically and, hence, are directly comparable to the previous plot. We can interpret the first plot as it is exactly the same way. It is just convenient to have two different plots as the shape of the curves is different for smaller radii versus larger radii.

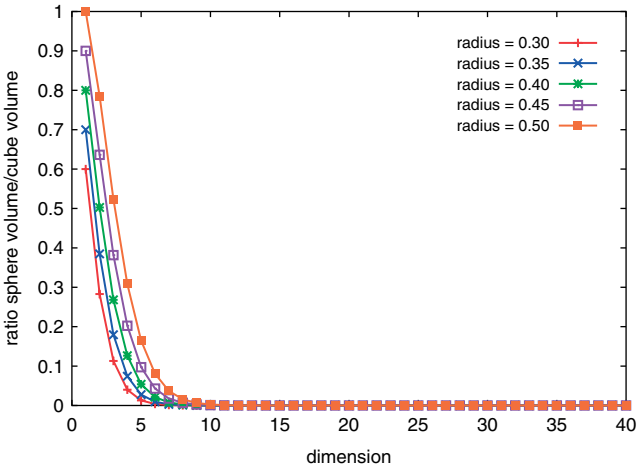

Fig. 16 Ratio: volume of a sphere vs. volume of the unit hypercube. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

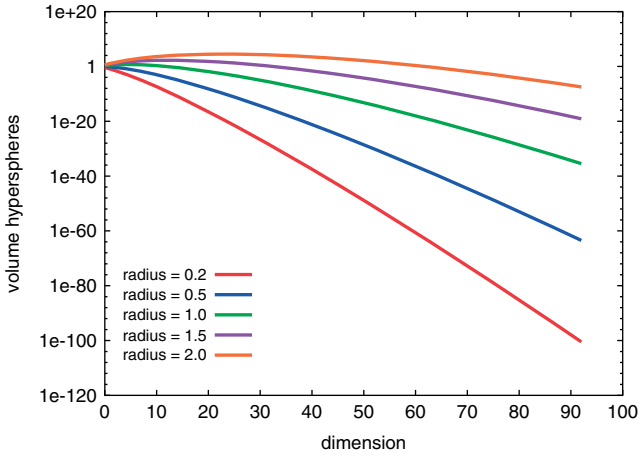

Fig. 17 Small change of radius, big change of volume. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

unit cube (side length of 1), that is, all hyperspheres here are completely inside the hypercube. The cube is in fact the minimal bounding box of the sphere with radius 0.5. Consequently, the corners of the unit hypercube are beyond the boundaries of the sphere in two dimensional space already, the maximum of the volume ratio is with dimensionality 1. In the one-dimensional case, for radius $= 0.5$ both volumes are identical. In two dimensions, the ratio is obviously $\frac{\pi}{4}$. We see the ratio rapidly decreasing again, this time allowing meaningful conclusions. In 10 dimensions already, the volume of the sphere is almost negligible compared to the volume of the bounding box. As the figure shows, the general behavior is the same for different radii as well (shown are hyperspheres with slightly smaller radii). The absolute values of the ratio are varying, though. These effects (and some others) are discussed nicely in ref. 63. In our context, this observation suggests that the choice of a certain radius ε to select a neighborhood (a fundamental step in many outlier detection methods) is notoriously rather sensitive to dimensionality. Small changes in the radius may decide whether everything or nothing is selected in a given dimensionality (e.g., a certain subspace), but the same amount of change may have no effect whatsoever if we have another data set (or another subspace) with some dimensions more. In order to describe a neighborhood containing at least some point, in high dimensions we need a radius that, if it were used in a lower dimensional subspace, would engulf the complete data space multiple times (see Fig. 17 for a visualization, note here the log-scale of the volume). It is more stable to use k-nearest neighbors instead of some ε-neighborhood since this guarantees to always have k objects for further computations. However, the effect on a refined modeling of data characteristics (as, e.g., deriving a density model) remains, as a certain density level (that may be a threshold discerning between inlierness and outlierness) is not comparable over subspaces of different dimensionalities.

There are two obvious consequences for outlier detection.

First, while outlier rankings may be still good, the underlying outlier scores do not allow to separate between outliers and inliers, that is, considering the outlier scores, there is no obvious gap between outliers and inliers. Aside from a viable ranking, however, such a gap would be highly desirable, as stated already by Hawkins [64]: *'a sample containing outliers would show up such characteristics as large gaps between 'outlying' and 'inlying' observations and the deviation between outliers and the group of inliers, as measured on some suitably standardized scale'*.

Second, if outlier scores are influenced by distance values, and distance values vary substantially over different dimensionality, how can outlier scores derived from subspaces of different dimensionality be compared? Outlier scores based on more attributes may be more prominent numerically while the outlier characteristic may actually be less prominent.

2.4. Combinatorial Issues and Subspace Selection

Another aspect also known as the 'curse of dimensionality' in other domains is the combinatorial explosion. For example, combinatorial effects strike at the level of distance functions. For a normal distribution, an object is farther than three standard deviations away from the mean with a probability of approximately 0.27%—*nota bene*, in a single dimension. Assuming that each dimension is independently normally distributed, an object appears to be normal with a likelihood of approximately 0.9973^d; at $d = 10$ dimensions,

an object is within the 3σ range in *every* single dimension with 97.33%, at $d = 100$ with 76.31%, and at $d = 1000$ at just 6.696%. So looking at a single attribute of high-dimensional distributions, virtually every object is extreme in at least one dimension. This can to some extend be controlled with extreme significance thresholds—in order to get 99.73% correct at 1000 dimensions, we could use a significance threshold of 99.99973% in every single dimension, or about 5σ—but at the cost of missing outliers that are less significant. In addition, we might need to increase these thresholds additionally to account for data set sizes: in a normally distributed, single dimensional data set of just 10000 objects, around 27 objects will appear to be outliers at 3σ *just by chance*. Therefore, it may be necessary to increase the threshold even further.

The statistical bias introduced by testing such a large quantity of models is also known as data-snooping bias and is closely related to model overfitting. The proper statistical practice of developing a hypothesis and then validating it on independent data is violated by an exhaustive search of subspaces reusing the same data that is to be evaluated. In order to obtain statistically valid results, we must therefore not base the choice of the evaluation subspace on the object we want to test. But even at choosing the subspace, we already see a combinatorial explosion once we go beyond looking at single attributes as in the example above.

A method for generating models that does not incur a data-snooping bias is to partition the data set with a grid. This works quite well in low-dimensional data, but while a grid with 10 bins in each dimension has 100 cells in two dimensions it already has 10^{100} (i.e., one googol) cells in 100 dimensions. In order for these cells not to be already empty on average, we need at least as many objects as cells. In order to draw significant conclusions from the number of objects in a cell, we need a data set at least an order of magnitude larger than the number of cells. Therefore, a naïve grid-based approach is also not feasible in high dimensions.

When selecting a subspace out of d dimensions, there are already $2^d - 1$ unordered possible subspaces to choose from. Looking at all projected subspaces is therefore also not feasible in high-dimensional data. Instead, these decisions will need to be taken on the basis of single dimensions and aggregated to a higher order subspace. The number of combinations further increases, when we instead of selecting (or not selecting) dimensions select an interval in each dimension (not selecting a subspace can be seen as restricting it to its full value range). If we had just 10 interval borders to choose from, we then have 45^d affine, axis-parallel subspaces to choose from. Allowing arbitrarily oriented linear manifolds removes any bounds from the number of models to choose from.

In this more general setting, a subspace outlier can be thought of as an object that is outlying in some subspace or some combination of attributes, whereas it could be perfectly normal in other attributes and even in the majority of attributes. The normality in many attributes might also outweigh the abnormality in some subspace. This could render the subspace outlier undetectable (masked) in the full dimensional space. A simple model (in fact, the first model proposed for subspace outliers [11]) would partition the data space into a grid of cells. For a single cell, an expected number of contained points can be computed assuming a uniform distribution. As (subspace) clusters can be seen as (groups of) dense grid cells [31], all points contained in unexpectedly sparse grid cells could be seen as being outliers.

As we have seen, combinatorial issues of high-dimensional data can be problematic on various facets: on one hand, the model search space can explode, rendering many search methods unusable. On the other hand, evaluating an object against many possible subspaces—or even the broad search for a subspace in which the object is unusual—may introduce a statistical bias called the data-snooping bias and may cost us statistical validity and significance.

2.5. Hubness

Recently, a couple of papers studied the effect of so called 'hubness' in high-dimensional data on different data mining and machine learning tasks [26,65–68]. Hubness is a phenomenon well known in graph data analysis [69]. In vector spaces, 'hubs' are points relatively close to many other points, that is, they occur very often in k-neighborhoods of other points while other points may occur rarely if ever in k-neighborhoods. 'Hubness' (or, more specifically, k-hubness of an object o: $N_k(o)$) is thus the number of times a point o is counted as one of the k nearest neighbors of any other point in a data set. It turns out that, with increasing dimensionality, many points show a small or intermediate hubness while some points exhibit a very high hubness. As they put it in ref. 66, intuitively, we could interpret these 'hubs' as very popular neighbors. More formally, this behavior of k-neighborhoods can also be seen as a distribution of the distances to the kth nearest neighbor which is increasingly skewed with increasing (intrinsic) dimensionality, as studied in refs. 65,66. As they also notice, setting a fixed ε-neighborhood, hubs do not emerge. The phenomenon seems, thus, closely related to the diminishing contrast and discriminability of distances: while the absolute distance values are uninformative, the rankings differ considerably and result in high hubness of some points, usually those that are more central with respect to the corresponding distribution.

Why this special aspect of the concentration effect is worth making a separate point here was pointed out by Radovanović et al. [66]: the other side of the coin of hubness is probably rather relevant for outlier detection in high-dimensional data. There will also be antihubs that are unusually far away from most other points and that, specifically, exhibit a large distance from their kth nearest neighbor which would qualify them as outliers according to an outlier model based on the kth nearest neighbor distances [27], despite the fact that they are generated by the usual distribution. Probabilistically, however, as argued in ref. 66, hubs are also outlierish since they are rare. However, Radovanović et al. only touched upon this question. Overall, the relation of hubness and outlier degree appears to be remaining an open issue for high-dimensional outlier detection.

2.6. Consequences

What did we learn about the impact of the curse of dimensionality on outlier detection in high-dimensional data, what are open issues and questions? Let us recollect the important points as a couple of problems or challenges for outlier detection in high-dimensional data:

Problem 1 (Concentration of Scores)
Due to the central limit theorem, the distances of attribute-wise i.i.d. distributed objects converge to an approximately normal distribution with low variance, giving way to numerical and parametrization issues.

Problem 2 (Noise attributes)
A high portion of irrelevant (approximately i.i.d. distributed) attributes can mask the relevant distances.

Problem 3 (Definition of Reference-Sets)
Common notions of locality (for local outlier detection) rely on distance-based neighborhoods, which often leads to the vicious circle of needing to know the neighbors to choose the right subspace, and needing to know the right subspace to find appropriate neighbors.

Problem 4 (Bias of Scores)
Scores based on L_p norms are biased toward high-dimensional subspaces, if they are not normalized appropriately. In particular, distances in different dimensionality (and thus distances measured in different subspaces) are not directly comparable.

Problem 5 (Interpretation & Contrast of Scores)
Distances and distance-derived scores may still provide a reasonable ranking, while (due to concentration) the scores appear to be virtually identical. Choosing a threshold boundary between inliers and outliers based on the distance or score may be virtually impossible.

Problem 6 (Exponential Search Space)
The number of potential subspaces grows exponentially with the dimensionality, making it increasingly hard to systematically scan through the search space.

Problem 7 (Data-Snooping Bias)
This is the correct version of the misconception that every point in high-dimensional data is an outlier: Given enough subspaces, we can find at least one subspace such that the point appears to be an outlier. Statistical principles of testing the hypothesis on a different set of objects need be employed. This problem is probably the most subtle and persistent one. We could see it also as problem of overfitting, since the outlier model is overly adapted to the data point given. This point of view makes it obvious that this problem is the most fundamental one to avoid in any learning procedure.

Problem 8 (Hubness)
What is the relationship of hubness and outlier degree? While antihubs may exhibit a certain affinity to also being recognized as distance-based outliers, hubs are also rare and unusual and, thus, possibly are outliers in a probabilistic sense.

Aside from these problems and challenges for correctness and statistical validity, high-dimensional data pose special problems for the feasibility, for example, for efficient neighborhood search which is usually a fundamental step to derive a reference-set for each point to judge on its deviation in characteristics from the reference-set.

3. EFFICIENCY AND EFFECTIVENESS FOR OUTLIER DETECTION IN HIGH-DIMENSIONAL DATA

The outlier detection methods we are discussing in this survey (unsupervised methods for high-dimensional numeric data in Euclidean space), usually rely on the assessment of neighborhoods (e.g., counting objects in ε-neighborhoods [70,71], using kth-nearest-neighbor-distances or aggregates thereof as outlier score [27,72], or even comparing the object's neighbors with the neighbors of the object's neighbors [28]).

Various efficient variants have been proposed for these basic approaches [27,72–79] usually based on pruning, on sampling, or on ranking strategies (see ref. 80 for an overview and unifying discussion). Using efficient indexing-structures, such as R-trees [81], R*-trees [82], or X-trees [15], can also yield $O(n \log n)$ performance for many approaches. However, these techniques usually deteriorate in efficiency with increasing dimensionality [20] due to the same issues described in Section 2: rectangular

pages offer an increasingly bad approximation for the data, the options for splitting increase, and the distances between closest and nearest objects concentrate.

Hence, some recent outlier detection methods tuned for efficiency in high-dimensional data use approximate neighborhood computation (see Section 3.1).

Regardless of the efficiency, the neighborhood in high-dimensional data is also less expressive (however depending on the given data distribution as discussed in Section 2). Hence, other adaptations to high-dimensional data address the issue of effectiveness and stability (see Section 3.2).

3.1. Efficiency

3.1.1. Background: approximate neighborhoods

Locality sensitive hashing (LSH) was first proposed in ref. 19 as a cure to the curse of dimensionality for closest-pairs search and later refined, for example, in refs. [83–85]. The key ingredient for this is the Johnson-Lindenstrauss Lemma [86], which proved the existence and bounds of projecting n objects into a lower dimensional space of dimensionality $\mathcal{O}(\log n/\epsilon^2)$, such that the distances are preserved within a factor of $1+\epsilon$. Matoušek [87] further improves these error bounds. The most interesting and surprising property is that the reduced dimensionality depends only logarithmically on the number of objects and on the error bound, but *not* on the original dimensionality d. Different ways of obtaining such a projection have been proposed for common norms such as Manhattan and Euclidean distance. A popular choice are the 'database-friendly' random projections [88, 89], where $\frac{2}{3}$ of the factors are 0 and the others ± 1, which can be computed more efficiently than the previously used matrices. See ref. 90 for an overview and empirical study on different variations of the Johnson-Lindenstrauss transform, indicating that a reduced dimensionality of $k = 2 \cdot \log n/\epsilon^2$ will usually maintain the pairwise distances within the expected quality. The known error bounds give a controlled way of reducing the dimensionality of the data set, conveniently based on random projections that are independent of the actual data and often rather cheap to compute as opposed to using, for example, principal component analysis. The error bounds allow for choosing a controlled trade-off between efficiency and precision. It should however be noted, as pointed out in ref. 91, that random projection methods are not suitable to defy the concentration of distances (see Section 2.1). They do, however, preserve strong structure that may be present in the data (see also discussion in Section 2.2).

Space-filling curves, also known as Peano curves [92], are another method of reducing the dimensionality while approximately preserving neighborhoods. In contrast to random projections, the projection to a space-filling curve does not directly preserve distances, but neighborhoods (to a certain extend). However they can still be used to compute neighbor candidates efficiently. Then only for those candidates the full-dimensional distance needs to be computed.

The key idea of a space-filling curve is to draw a one-dimensional curve through the data space that—by being recursively defined, which results in a fractal curve—gets arbitrary close to every point without intersecting itself. The most popular variants are the Z-curve [93] and the Hilbert curve [94] that recursively split the space into halves. The original Peano curve [92], which divides into three parts, is less often used. Originally, these curves were developed for two-dimensional spaces. They can be generalized to higher dimensional spaces, though. Morton [93] was the first to apply them in database systems for indexing geographic data. Since then, they have been used for example in Hilbert R-Trees [95].

As the original motivation of space-filling curves was to provide a complete order of (two-dimensional) vectors, space-filling curves are an extreme dimensionality reduction technique: they always reduce to a one-dimensional space, a linear order of all points. Intuitively, they can be interpreted as repeatedly cutting and opening the data space along some edge in the process of linearization. With an increasing number of dimensions, the number of such cuts—where neighborhoods are not preserved—increases. Figure 18 is a visualization of the 3rd order Hilbert curve, with the actual curve in black while the red lines indicate neighborhoods that are not preserved well. When the curve is unfolded, the red edges are no longer adjacent along the curve. It can be seen that space filling curves rely on an intricate (and just as infinite and fractal) pattern of cutting the data space into fragments that are then ordered by the curve.

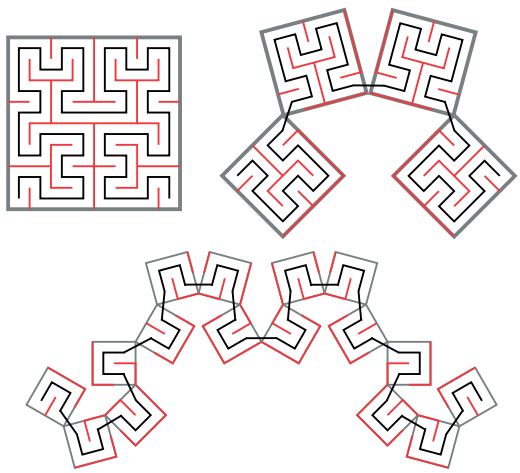

Fig. 18 Visualization of Hilbert curve unfolding. Red edges are cut at curve iteration 3. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

3.1.2. Outlier detection using approximate neighborhoods

The concept of pruning based on approximate nearest neighbor search has been introduced to outlier detection by ref. 27. This concept is sometimes supported by data partitioning techniques as, for example, in refs. 27,96.

For high-dimensional data, a variant has been proposed in ref. 97. Their approach RBRP (recursive binning and re-projection) is a combination of binning and projecting the data. In a first phase, they bin the data, recursively, into k clusters (using a k-means like clustering [98] at each level). This results in k bins, and again, in each bin, k bins and so forth, unless a bin does not contain a sufficient number of points. For each resulting bin, in a second phase, approximate neighbors are listed following their linear order as projected onto the principal component of the bin. Finally, within each bin (as long as necessary), they use a variant of the nested loop algorithm [74], to derive the top-n outliers. The resulting outliers, though based on approximate neighborhood, are reported to be the same as delivered by ORCA [74], yet they are retrieved by RBRP far more efficiently in high-dimensional data than they are by ORCA. RBRP could thus be seen as an adaptation of ORCA to high-dimensional data.

An outlier detection method, LSOD (locality sensitive outlier detection), based on LSH has been proposed in ref. 99. In LSOD, also, the approximate neighborhood search (here based on LSH) is combined with a data partitioning step using a k-means type clustering. The idea of outlierness is intuitively that points in sparse buckets will probably have fewer neighbors and are therefore more likely to be (distance-based) outliers. The pruning is based on a ranking of this outlier *likelihood*, based on statistics on the partitions. As this intuition is closely connected to the distance-based notion of outliers, their conjecture that their approach *'can be used in conjunction with any outlier detection algorithm'* may be overly optimistic. Nevertheless, as a ranking approach for distance-based outliers LSOD constitutes an interesting contribution to the research area.

PINN [100] (projection-indexed nearest-neighbors), a variant of LOF [28], uses random projections to perform outlier detection in high-dimensional data. PINN facilitates an approximate neighborhood search that is used as a replacement of the neighborhood and distances computation of LOF. The distances required by LOF are estimated based on the random projections. They prove that the projection also preserves the outlier scores within the known error bound of random projection which is a theoretically very interesting result. In order to further improve the result quality, they retrieve the $c \cdot k$ nearest neighbors in the projected space, and then refine these candidates to the intended number k, which for low values of $c = 2$ or 3 offered considerable benefits. We conjecture that the combination of PINN with outlier detection methods other than LOF will prove a very interesting topic for further research.[2]

Let us also note that there is an approximate method for the well-known LOCI (local correlation integral) algorithm [102]. The basic version is estimating a multi-granularity deviation factor (MDEF) for each point and its local neighborhood. The approximate method aLOCI approximates the neighborhood by means of a space partitioning grid, leading to practically linear performance. Both variants, LOCI and aLOCI, are however not especially suitable for high-dimensional data. Since the approximation in aLOCI is grid-based, its quality and efficiency deteriorates with increasing dimensionality.

In refs. 72,103, space-filling curves are used for an approximation step. The main achievements of this work [72,103] are (i) the proposal of a new variant of the distance-based notion of outlierness, using the aggregated distances to the k nearest neighbors instead of using solely the distance to the kth nearest neighbor; and (ii) an efficient approximation of the nearest neighbors, based on Hilbert space filling curves [94]. Using the aggregated distances (i) has a certain smoothing effect and is usually more stable than using solely the distance to the kth nearest neighbor. The effectiveness of Hilbert curves (ii) has been widely studied [104]. The problem with space filling curves for approximations is that, though neighbors along the space filling curve are also close in the data space, the opposite relationship does not necessarily hold. In order to limit this loss for approximate nearest neighbor search, in [105], the authors proposed shifting the objects d times (where d is the data space dimensionality), resulting in $d + 1$ copies of the data. They hereby yield a guaranteed approximation quality. This idea is also used in [72,103] as a filter step before the exact computation of the top-n outliers.

3.2. Effectiveness and Stability

Some statistical methods for outlier detection meet the requirements of high-dimensional data by feature selection or dimensionality reduction methods (see ref. 106 for a discussion of dimensionality reduction and e.g., refs. [107–109] for the use of PCA for high-dimensional outlier detection; a methodology for quality assessment of dimensionality reduction is discussed in ref. 110). An example for a more data mining-oriented method is ref. 111, where a feature extraction method is proposed specifically to enhance outlier detection. Actually, this method retrieves not a subset of the features but an optimized combination of features, based on separation of outliers versus inliers

[2] Lately [101], random projections have been used to speed up ABOD [44], which we discuss in Section 3.2.2.

in training data. In these methods, the dimensionality is reduced and all the outliers are sought in the remaining or transformed feature space where they are expected to be more prominent.

In such approaches to outlier detection based on dimensionality reduction, essentially a single subspace for the complete data set is assumed to be sufficient to derive all outliers. As a side effect of the evaluation of their method, it was demonstrated in ref. 112 that a global dimensionality reduction by PCA and outlier detection as a second step in the reduced feature space is likely to fail in the typical subspace setting. As opposed to global feature reduction methods, in this survey, we are interested in methods that still define outliers in the full space (present section) or identify potentially different subspaces for different outliers (Section 4).

3.2.1. *Combination of feature subsets*

Combining different feature subsets (i.e., subspaces) for outlier detection has been proposed first in the 'feature bagging' approach [113]. There, using different feature subsets was motivated by improving the quality of the overall prediction of an ensemble outlier detector, built from the single outlier detectors on the different (randomly selected) feature subsets. Though not aiming at high-dimensional data, this approach could be seen as an approach to deal with high dimensionality. This has been made more explicit in ref. 114, combining randomly selected subspaces for an (unsupervised) ensemble. Motivated by the 'curse of dimensionality', computing distances in randomly selected subspaces and combining the findings is expected to be more stable and effective than computing the distances (as required by most outlier detection methods) in the high-dimensional complete feature space.

Let us note that, aside from the mentioned approaches, there exist only a few outlier detection ensemble methods, namely [115–117]. These ensemble approaches are not specialized on feature subsets although they straightforwardly could use a similar setup as well.

Common to all these approaches is that, though they are using subspaces in the process of identification of outliers, they do not actually define outlierness with respect to specific subspaces but, as a result of the combination of different subspaces, they define outlierness with respect to the full, potentially high-dimensional, space. Using subspaces explicitly for modeling outlierness will be discussed in Section 4.

3.2.2. *Angle-based method*

Since the cosine distance is often successfully used in high-dimensional data, it is not surprising that an outlier detection method tailored for high-dimensional data was based on the use of angles instead of plain distances: The outlier detection method ABOD (angle-based outlier detection) [44] assesses the variance in angles between an outlier candidate and all other pairs of points. For outliers, most other data objects will be concentrated in some directions, while for inliers, data objects are distributed in all directions. Hence the variance of angles is lower for outliers than for inliers. The lower variance therefore signals the higher outlierness (as opposed to many other approaches where the higher score signals stronger outlierness).

ABOD has been shown to remain stable, where plainly distance-based methods deteriorate. However, the method scales cubic with respect to the number of data objects, since for each object, all pairs of other objects are checked. A more efficient variant is based on samples but still relatively stable [44].

A near-linear time efficient approximation variant of ABOD is proposed in ref. 101, based on random projections.

4. SUBSPACE OUTLIER DETECTION

The idea of defining outliers with respect to subspaces of the original feature space can be traced back to ref. 118. There, the explanation of outlierness is based on the features of major deviation from the majority of the usual data. The nature of anomaly of an already identified outlier is explained by referring to those attributes where this outlier shows a high deviation from the remainder of the data set, that is, the 'intensional knowledge of distance-based outliers' takes the form of explanations using a subset of attributes.

Yet, in the approach of Knorr and Ng, the outliers were defined in the full dimensional space and only in retrospect explained by subspace properties. The question arises, if outliers could also be defined in terms of subspace properties in the first place. Methods to do so came up mostly very recently and focus mainly on two aspects of the problem: (i) how to identify the subspace where an object is an outlier (tackling, e.g., Problems 2, 3, 6), and (ii) how to make scores based on different characteristics (such as the differing dimensionality of some subspaces) actually comparable in a meaningful way (tackling, e.g., Problems 4, 5).

We discuss the existing methods with respect to these problems (i) and (ii) in the following Sections 4.1 and 4.2, respectively. Finally, we will have a short comment on the possible types of subspaces identified by the methods (such as: axis-parallel, or arbitrarily-oriented, that is, based on correlations among attributes), in Section 4.3.

4.1. Identification of Subspaces

One possible fundamental approach to outlier detection is to identify clusters, for example, using the density-based paradigm [119], where objects need not belong to any cluster but could remain as 'noise'. Those noise objects could be interpreted as being outliers. For low-dimensional outlier detection, this is usually seen as a rather crude approach. Many subspace outlier detection approaches, however, can be seen that way. They define clusters (implicitly or explicitly), and identify those objects not belonging to any cluster as outliers.

What makes this approach suitable for subspace outlier detection is the fact that many subspace clustering approaches are density-based or grid-based (i.e., the data space is partitioned by some grid, and those grid cells containing more than some expected number of objects constitute cluster elements [31]). To avoid searching a number of grid-cells that explodes exponentially with the number of dimensions, these approaches usually start with partitioning each dimension individually and combining only those dimensions for two-dimensional (three-dimensional, and so on) grids that at least contain one cluster (usually adopting an Apriori-like [120] search heuristic).

What makes this approach questionable for subspace outlier detection, however, is, first, that an Apriori-like strategy is not possible if we are searching for sparse regions, instead of dense regions, since the sparsity is not anti-monotonic over increasing dimensionality. Besides, even in subspace clustering the monotonicity is based on restrictions that render the usage of results limited [33]. Second, it should be kept in mind that the subspace clustering problem is still not sufficiently well defined and does not have satisfying solutions. It is not even clear what a satisfying solution in terms of retrieved subspace clusters would be as the evaluation of these approaches is an open and challenging problem (some aspects are discussed in refs. 41,60,61,121,122).

Accordingly, the first approach claimed to be suitable for high-dimensional data [11] (with the same content also published in ref. 123) resembles a grid-based subspace clustering approach where not dense but sparse grid cells are sought to report objects that are contained within these sparse grid cells as outliers. The grid is defined to partition the data space dimension-wise using equidepth ranges, that is, each attribute is divided into ϕ ranges such that each range contains a fraction of

$$f = \frac{1}{\phi}$$

of the data objects. Note that, although this way each one-dimensional range contains the same number of objects, the intersection of two one-dimensional partitions can contain more or less than the fraction f. The algorithm now shall assess whether an intersection of k one-dimensional partitions (i.e., a k-dimensional hyper-cuboid in the d dimensional space, $k \leq d$) contains unexpectedly few data points. Assuming the data consisting of N points being uniformly distributed in d dimensions, the expected value is given by

$$N \cdot f^k,$$

and the standard deviation is given by

$$\sqrt{N \cdot f^k \cdot (1 - f^k)}.$$

The outlier model therefore designates any point an outlier that is contained in a cuboid containing (significantly) less than expected data points.

This approach has various problems. As discussed in Problem 6, with increasing dimensionality, the expected value of a grid cell quickly becomes too low to find significantly sparse grid cells. The authors are aware of this problem [123, p. 217], and argue that k must be chosen small enough. Secondly, the parameter k must be fixed, as the scores are not comparable across different values of k (Problem 4). The search space is too large even for a fixed k only (Problem 6), which is why the authors propose a genetic search that ensures the value of k is preserved across mutations. With a restricted computation time, their method will inspect just a tiny subset of the $\binom{n}{k}$ projections (not yet to speak of individual subspaces), and their randomized search strategy does neither encourage fast enough convergence nor diversity. The method can thus not give any guarantees about the outliers detected or missed.

Using this randomized model optimization without a statistical control also raises the question of the statistical bias (Problem 7) and how meaningful the detected outliers actually are. By assuming that the quantile grid will have approximately equally filled bins, the statistical test employed essentially assumes that the joint probability always equals the product of the marginal probabilities. Hence the presence of clusters in the data set will skew the results considerably. Additionally, the proposed equidepth binning is likely to include outliers in the grid cell of a nearby cluster and thus hide them from detection entirely; in fact a non-equidepth grid approach is much more likely to have outliers within sparse grid cells.

In general, when using a grid-based search, the identified dense areas also need to be refined to detect outliers that happen to fall into a cluster bin. In the presence of correlations, deviations from such a trend can otherwise not be recognized. Nevertheless, the authors of ref. 11 are credited for bringing the problem of subspace outliers to the attention of the data mining community.

Historically, the second approach dedicated to this problem was HOS-Miner [124]. They propose not to identify

subspace outliers but rather to identify the subspaces in which a given point is an outlier. However, for that purpose, they also define the outlying degree of a point with respect to a certain space (or possibly a subspace) s in terms of the sum of distances to the k nearest neighbors in this (sub-)space s. For a fixed subspace s, this is the outlier model of ref. 72.

Defining subspace outlierness this way has the advantageous property of a monotonic behavior over subspaces and superspaces of s, since the outlying degree OD is directly related to the distance-values. At least for L_p-norms (this restriction is not discussed in ref. 124), the following property holds for any object o and subspaces s_1, s_2:

$$OD_{s_1}(o) \geq OD_{s_2}(o) \iff s_1 \supseteq s_2. \quad (1)$$

This is trivially true and is used to facilitate an Apriori-like search strategy (down-ward as well as up-ward) for outlying subspaces for any query point by setting a threshold T discriminating outliers (where $OD_s(o) \geq T$) from inliers (where $OD_s(o) < T$) in any subspace s. Alas, setting the same fixed threshold to discern outliers with respect to their score OD in subspaces of different dimensionality ignores the problem that these scores are rather incomparable (see Problem 4). Later approaches specifically addressed this comparability-problem, see Section 4.2. The authors of ref. 125 (see below) even assert that the monotonicity of Eq. (1) must not be fulfilled for true subspace outliers (since it implies the outlier can be found trivially in the full-dimensional space then). Furthermore, the systematic search for the subspace with the highest score raises the question of a data-snooping bias (see Problem 7).

OutRank [45] (or the variant described shortly in ref. 126) extends a grid-based clustering approach with an analysis of the non-clustered objects that is much more feasible for high dimensionality. Clusters as opposed to outliers are not rare objects, and are likely to be recognizable in both lower and higher dimensionality (although a meaningful definition of subspace clusters is still challenging, as noted above). Since OutRank is based on finding a cluster model first (with some grid-based subspace clustering such as DUSC (dimensionality unbiased subspace clustering) [127] or some suitable density-based subspace clustering such as EDSC (efficient density-based subspace clustering) [128]), it avoids the aforementioned statistical bias (Problem 7) that comes along with approaches searching some subspace for each object where the object is an outlier. Instead, the outlierness is essentially asserted based on how often the object is recognized as part of a cluster and on the dimensionality and size of the subspace clusters it is a part of. In order for this to work, a strong redundancy in the clustering is implicitly assumed. The result then may be biased towards hubs (Problem 8). It is, however, not totally satisfying to see outliers as just a side-product of density-based clustering. Depending on the parametrization of the density-based clustering algorithm, this can result in a large set of outliers: On a data set where no clusters are recognized, even all objects become outliers. Furthermore, outlier detection based on subspace clustering relies on the subspace clusters being well separated which links back to the issues we discussed in Section 2.2 on distributions being well-separated or not.

A method proposed for tackling the specialized problem of finding outliers in subspaces without a previous *explicit* clustering-step is SOD (subspace outlier detection) [129]. Based on a reference set for a point, the outlierness of the point is assessed. The reference set is seen as possibly defining (implicitly) a subspace cluster (or a part of such a cluster), that is, the points may scatter along some dimensions while they may concentrate within others. If the query point deviates considerably from the reference set in the latter attributes, it is a subspace outlier with respect to the corresponding subspace. SOD provides actually not a decision (outlier vs. inlier) but an outlier score which could be interpreted as a probability-score of being an outlier according to the deviation from the reference set in the corresponding subspace by assuming a distribution of the distances in the relevant attributes, as visualized in Fig. 19.

For a single object, there is no comparison of multiple subspaces in SOD, but the subspace decision is based on the whole reference set, this way avoiding the statistical bias (Problem 7). It is crucial, though, to select

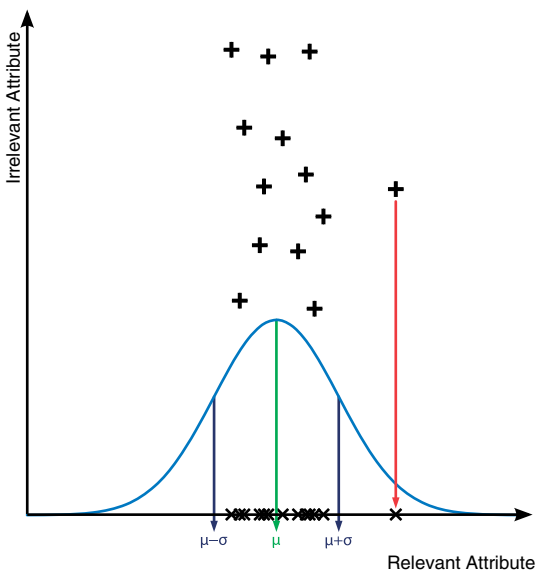

Fig. 19 SOD projects to locally relevant attributes, then assumes a normal distribution to judge outlierness. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

a meaningful reference set for each query point. In ref. 129, the authors proposed to use shared nearest-neighbors in order to stabilize the assessment of neighborhoods in high-dimensional spaces to attenuate Problem 3. Shared nearest-neighborhood was indeed demonstrated to enhance a meaningful selection of neighbors in high-dimensional space [53]. The normalization of scores and the interpretation as 'probability estimates' aims at improving with respect to Problem 5, which will be discussed further in Section 4.2.

OUTRES [130] specifically addresses the problem of bias of outlier scores due to differing value range of distance values in subspaces of different dimensionality (Problem 4). For their approach, this is actually even more important since they assess deviations of each object in several subspaces simultaneously and are interested in a general (combined) ranking of the objects according to their outlier scores in all 'relevant subspaces'. They restrict the assessment to this set of 'relevant subspaces', excluding subspaces with highly scattered objects and, thus, low contrast between outliers and inliers. OUTRES needs comparable neighborhoods (see Problem 3) for each point to estimate densities. In order to adjust for different subspace dimensionalities (see Problem 4), they adjust the ε radius for the given subspaces dimensionality. The score of a single subspace is then obtained by comparing the object's density to the average density of its neighborhood. The total score of an object is the product of all its scores in all relevant subspaces. Assuming a score in [0, 1], where the smaller score denotes the stronger outlier, this should provide a good contrast for those outliers with very small scores in many relevant subspaces. Details of the selection of relevant subspaces for OUTRES are revealed in the follow-up paper [131]. Any attribute that exhibits uniformly distributed values in the neighborhood of the currently considered point o is rejected, based on a statistical significance test. This excludes, for this o, also any superspaces of uniformly distributed attributes (and, hence, an Apriori-like search strategy can be applied). The assumption is, that subspaces where the objects are not uniformly distributed (in the neighborhood of o), exhibit a clustering structure where outliers can possibly meaningfully deviate. The remaining attributes (and their combinations) are seen as *relevant*. Note that the relevance is assessed for each data object individually, that is, a subspace can be relevant for one object and irrelevant for another object which renders this approach relatively expensive. This approach alleviates the problem of noise attributes (Problem 2) while it is based on a statistic on the neighborhood of the point and thus not likely susceptible to a statistical bias (Problem 7).

For HighDOD [125], the authors motivate their method with an interesting observation on the sum of distances to the k nearest neighbors as the outlier score (as defined by Angiulli and Pizzuti [103]) and the monotonicity property of Equation 1: while this enables an Apriori-like subspace search, it also implies that the outlier score must be maximal in the full-dimensional space, and the method will thus not find true subspace outliers (and a full-dimensional algorithm could have been used in the first place). Instead, they modify the outlier score of ref. 103 to use the L_p norm normalization we also used in Section 2. They conclude that the pruning of subspaces is impossible, and therefore examine all subspaces up to a user-defined maximum dimensionality m, similar to ref. 11. This however suggests that they possibly also fall prey to the data-snooping bias. To lessen the exponential number of subspaces, they only explore them up to a maximum dimensionality of $m = \lfloor \log_{10} n \rfloor$, and use a linear-time ($\mathcal{O}(n \cdot m)$) density estimation to generate outlier candidates they compute the nearest neighbors for.

At first sight, the observation that neither an upward- nor a downward-pruning is possible may appear to contrast to most of the previous efforts. However, this only holds when applied to the actual outlier score. Methods that use a pruning strategy for example for clustering will still work. It does however indicate that Problem 4 is fatal for HOS-Miner [124], and the maximum outlier score of HOS-Miner will always be in the full-dimensional space.

The latest paper on the topic of subspace outlier detection is dedicated to identifying high contrast subspaces (HiCS) [112], using Monte Carlo sampling to obtain a notion of contrast of a subspace. In these subspaces, they use LOF and aggregate the LOF scores for a single object over all interesting subspaces (although not using score normalization, as we discuss below in Section 4.2). Their philosophy of decoupling subspace search and outlier ranking may raise some questions since a certain measure of contrast to identify interesting subspaces will relate quite differently to different outlier ranking measures. It is doubtful that, as they state, instead of LOF, any other outlier score could be used interchangeably without a considerable impact on the results. As their measure of interestingness is based on an implicit notion of density, it may only be appropriate for density-based outlier scores. Nevertheless, this decoupling allows them to discuss the issue of subspace selection with great diligence as this is the focus of their study. The core concept for these subspaces with high contrast is a measure of correlation among the attributes of a subspace, based on the deviation of the observed probability density function from the expected probability density function, assuming independence of the attributes. These deviations are aggregated over several Monte Carlo samples to assign a value of contrast to a subspace. This measure is closely related to the notion of mutual information. Actually, it can be seen as a Monte Carlo estimation of the mutual information. The intuition is that,

in these subspaces, outliers are not trivial (e.g., identifiable already in one-dimensional subspaces) but deviate from the (although probably nonlinear and complex) correlation trend exhibited by the majority of data in this subspace.

The idea of comparing the joint probability with the marginal probabilities can be traced back to implicit assumptions in ref. 11. Their approach considered objects in areas less dense than expected immediately and indiscriminately as outliers, while HiCS refines them using LOF in the corresponding subspace.

Let us note that, while the focus of research was the identification of meaningful subspaces for outlier detection, an open and interesting future issue for finding outliers in different subspaces *efficiently* may be the appropriate use of suitable index structures for subspace similarity search, which is, however, a rather immature research area itself. To the best of our knowledge, all approaches to subspace similarity search potentially suitable are refs. [132–137]. However, unless the effectiveness of subspace selection for outlier detection is sufficiently resolved (and it is our impression that there is potential for improvement), it is not yet the turn of tackling efficiency issues.

4.2. Comparability of Outlier Scores

An outlier score provided by some outlier model should help the user to decide whether an object actually is an outlier. For many approaches even in low-dimensional data the outlier score is not readily interpretable. The scores provided by varying methods differ widely in their scale, their range, and their meaning. For many methods, the scaling of occurring values of the outlier score even differs within the same method from data set to data set, that is, outlier score o in one data set means, we have an outlier, while in another data set it may not be really exceptional. Even within one data set, the identical outlier score o for two different database objects can denote actually substantially different degrees of outlierness, depending on different local data distributions. Obviously, this makes the interpretation and comparison of different outlier detection models a very difficult task.

For low dimensional data, this problem has been recognized. LOF [28], for example, intends to level out different density values in different regions of the data, as it assesses the *local* outlier factor. Yet still, the same LOF value can designate very different degrees of outlierness in particular across different data sets and dimensionalities. LoOP [138] is a recent LOF variant that addresses this problem more specifically, providing a statistical interpretation of the outlier score by translating it into a probability estimate. This includes a normalization to become independent from the specific data distribution in a given data set. In ref. 116, generalized scaling methods for a range of different outlier models have been proposed. The unified scaling and better comparability of different methods could also facilitate a combined use to get the best of different worlds, for example, by means of setting up an ensemble of outlier detection methods. This is experimentally also shown for different projections, a common situation in subspace outlier detection, when scores of different subspaces need to be combined.

Considering subspaces for outlier detection sets up additional problems of comparability and bias (cf. Problems 4 and 5). Most outlier scorings are based on assessment of distances, usually L_p distances, which can be expected to grow with additional dimensions, while the relative variance decreases. Hence a numerically higher outlier score, based on a subspace of more dimensions, does not necessarily mean the corresponding object is a stronger outlier than an object with a numerically lower outlier score, based, however, on a subspace with less dimensions. Many methods that combine multiple scores into a single score neglect to normalize the scores before the combination. Another effect considerably affecting L_p distances is the normalization and scaling of the data set.

This specific problem has been accounted for by some of the subspace outlier detection methods. The model of refs. 11,123 circumvents the problem since they restrict the search for outliers to subspaces of a fixed dimensionality (given by the user as input parameter). OutRank [45] weights the outlier scores by size and dimensionality of the corresponding reference cluster. SOD [129] may be of special interest here since it also provides a scoring interpretable as probability, including a normalization over the dimensionality (cf. Fig. 19).

For OUTRES [130,131], this problem of bias is the core motivation. The score definition of OUTRES uses density estimates that are based on the number of objects within an ε-range in a given subspace (or, enhanced: on the sum of their weighted distances, using an Epanechnikov Kernel [139] with a bandwidth adaptive to the dimensionality). To account for the bias of distance-based scores towards high-dimensional subspaces, an adaptive neighborhood (ε is increasing with dimensionality) is proposed as well as an adaptive density by scaling the distance values accordingly. The score is also adapted to locally varying densities as the score of a point o is based on a comparison of the density around o vs. the average density among the neighbors of o (i.e., they seek local outliers in the sense of LOF [28]). This renders, however, the time complexity $O(n^3)$ for a database of n objects unless suitable data structures (such as, e.g., precomputed neighborhoods) are used. Because of the adaptation to different dimensionality of subspaces, where different neighborhoods are relevant for any point in different subspaces, this is not trivial here, yet it is not discussed in the paper.

The bias of distance-based outlier scores toward higher dimensions is also the main motivation for HighDOD [125]. Their approach is based on the outlier notion of ref. 72 (sum of distances to the k nearest neighbors), adapting the distances to the dimensionality d of the corresponding subspace by scaling the sum with $\frac{1}{\sqrt{d}}$ (or, if using another L_p distance than Euclidean, with $\frac{1}{\sqrt[p]{d}}$). Furthermore assuming normalized attributes (with a value range in $[0, 1]$), this results in restricting each summand to ≤ 1 and the sum therefore to $\leq k$, irrespective of the considered dimensionality. A normalization to adjust variances is, however, not done, and as such the curse of dimensionality is not completely handled.

For HiCS [112], it seems questionable whether the scores are comparable in a meaningful way since LOF scores retrieved in subspaces of different dimensionality are aggregated for a single object without normalization. As we have seen on the basis of Figs. 7 and 8, the LOF scores differ widely with varying dimensionality. An improvement could be based on LoOP [138] or the general scaling and normalization methods discussed in ref. 116. In the experiments carried out in ref. 112, however, this is not a major issue since the relevant subspaces vary only between 2 and 5 dimensions. Yet in general, while much diligence has been spent in ref. 112 on the selection of subspaces, the issue of comparability of scores has been ignored.

4.3. Types of Subspaces: Axis-parallel or Arbitrarily Oriented

Most of the algorithms for subspace outlier detection so far are restricted to outliers in axis-parallel subspaces. This restriction is, for example, due to grid-based approaches or to the required first step of subspace or projected clustering. The type of subspace sought by HiCS [112] is not restricted in this sense or even biased to find not-axis-aligned subspaces due to the analysis of correlation between attributes. Another example of definition of outliers in arbitrarily-oriented subspaces due to correlation among some attributes and the deviation of outliers from this correlation pattern has been proposed earlier as COP (correlation outlier probability) in ref. 140, ch. 18 (see also ref. 141), as an application of the correlation clustering concepts discussed in ref. 142. According to this model, points have a high probability of being a 'correlation outlier' if their neighbors show strong linear dependencies among attributes and the points in question deviate substantially from the corresponding linear model. Identifying so called 'correlation outliers' may, thus, be another interesting way to define outliers in high-dimensional data. But for both, axis-parallel and arbitrarily oriented outlier models, many (and to a good part, identical) issues remain open and wait for future research.

5. DISCUSSION

5.1. Tools and Implementations

We would also like to point to some open-source implementation frameworks related to subspace outlier detection.

Subspace Outlier Ranking Exploration Toolkit (SOREX)[3] [143] is a collection of some outlier detection algorithms for high-dimensional data, mainly providing variants of OutRank that are based on using different subspace clustering algorithms. SOREX is part of the OpenSubspace project [144], which again is an extension of WEKA [145] for subspace clustering. WEKA does however not include index structures, which can provide substantial speedups for many data mining applications. Let us remark, though, that SOREX does unfortunately not come along with the source code of the available methods.

ELKI[4] is an actively developed and maintained '**E**nvironment for deve**L**oping **K**DD-applications supported by **I**ndex-structures'. Two recent releases [146,147] were especially dedicated to outlier detection, release 0.3 [146] focusing on visual evaluation of outlier scores, release 0.4 [147] specialized on geographical or spatial outlier detection, comprising—additionally to a selection of fundamental outlier detection methods such as those in refs. 27,28,71,72,75,78,138,148 and LOCI including the approximate aLOCI [102]—also some spatial outlier detection methods such as those in refs. [149–151], among others. With the latest release 0.5 [152], several implementations of algorithms for outlier detection in high-dimensional data are provided, such as refs. 44,45,112,113,129,130,153 and the implementation of refs. 72,103 based on Hilbert curves, more are going to be included. Some of the figures in this article were generated using ELKI.

5.2. Data Preparation, Normalization, and Bias

Data preprocessing affects the outcome considerably. Some methods require the data to be normalized to the interval $[0 : 1]$, others expect standardized attributes ($\mu = 0, \sigma = 1$) or work on the ranks. Any such transformation implies a slight information loss, but often improves the results considerably. Some algorithms can only be used with normalized attributes, while others are just biased by non-standardized attributes. Normalization can be done as a preprocessing step, or by adjusting the distance

[3] http://dme.rwth-aachen.de/de/OpenSubspace/SOREX
[4] http://elki.dbs.ifi.lmu.de/

function accordingly. Last but not least, the distance function itself is an implicit parameter to many algorithms. Frameworks such as WEKA and ELKI offer a wide choice of such normalizations, filters, preprocessors, and distance functions beyond the classic L_p norms.

To discern between high and low variance in order to distinguish irrelevant from relevant attributes, perhaps even in some local neighborhood, is also far from trivial. In presence of a single highly correlated attribute pair (for example a duplicate attribute or the sensor readings of two redundant sensors), all other correlations in the data set may appear insignificant in contrast to this. Yet, few of the established methods consider this situation of extreme (but trivial) correlations masking weaker but more interesting trends in the data set. HiCS uses a variant of mutual information to skip attribute combinations. Only the candidates with the least mutual information are kept, but there is no treshold on what amount of mutual information is significant. COP [140,141] uses PCA and a local subspace filter to compute arbitrarily oriented subspaces, and needs to decide which axes are of high or low importance. In ref. 40, improvements for PCA in presence of outliers and the interaction with such filters are discussed. The ELKI framework includes a large collection of those filters to discern relevant and irrelevant subspaces.

5.3. Evaluation of Outlier Detection Methods

Unsupervised data mining on high-dimensional data is in general hard to evaluate. Two studies on subspace clustering, as a related field, already demonstrate the difficulties associated with evaluation [144,154]. Qualitatively evaluating outlier detection is even harder. If you evaluate a new algorithm, one of the most delicate tasks is probably to choose a fair setup for the competitors. It is our hope that this survey will also help to understand the different aspects that influence the results of different algorithms. For example, to parametrize SOD [129] requires to understand the impact and appropriate choice of the neighborhood for computation of shared neighbor distances [53].

Not only for high-dimensional data, but generally for outlier detection, how to properly evaluate in an informative way the derived outlier rankings and outlier scores is an open and notorious problem.

Existing evaluation procedures exhibit, though often used, a couple of problems that should also be acknowledged when using such evaluation methods. For example, *precision*@k, which is the true positive rate for the top k results in a data set that contains k outliers, is a rather naïve way of evaluating the result. What renders this approach problematic is the imbalanced nature of the outlier problem. Consider a data set containing 100 objects. The typical expectation would be that 2 of those are outliers. A method that ranks the true outliers on rank #3 and #4 will have a *precision*@2 of 0. Thus, occasionally, the precision values are shown for a larger range of k.

For an area like information retrieval, for example, in search engines, this evaluation method is sensible. There, usually only the top, say, 10 results of a search query are displayed, and the classification of the data set provides complete information on the relevancy.

Contrariwise, for the task of evaluating unsupervised outlier detection, using some labeled data set, it is important to acknowledge that the 'ground truth' may be incomplete and that real world data may include sensible outliers that are just not yet known or were considered uninteresting during labeling. This could happen in particular when algorithms are evaluated using classification data sets, assuming that some rare (or downsampled) class contains the outliers. Such a setup is neglecting the possibility that the rare class may be clustered, whereas there are probably additional true outliers within the frequent classes. If a method is detecting such outliers, that should actually be rated as a good performance of the method. Instead, in this setup, detecting such outliers is overly punished due to class imbalance.

A more advanced method of evaluation, widely used in evaluation of outlier detection methods, is based on receiver operating characteristic (ROC) curves. ROC curves are a plot of the true positive rate against the false positive rate. Since a plot is hard to evaluate, this is turned into a measure by computing the area under this curve (AUC). For a random result, both rates will grow simultaneously, resulting in an area approximately filling half of the space. For a perfect result returning all outliers first and only then returning the inliers (i.e., we have 100% true positives before we even get the first false positive), the area under the corresponding curve will cover all available space. ROC curves and ROC AUC analysis inherently treat the class imbalance problem by using the relative frequencies which makes them popular for evaluation of outlier detection. However, the best thing we have is not necessarily good. A problem associated with ROC curves is that they lose the score information. The ROC AUC analysis does not assess whether the outlier score offers a reasonable contrast between outliers and inliers, though yielding such high contrast would be a plus for any outlier detection algorithm (see Problem 5).

Motivated by these findings, lately a correlation measure to compare rankings has been developed [117] that is taking the scores into account. This correlation not primarily to a given ground truth but between different outlier score rankings provided by different outlier detection algorithms, with different parameterization, and on different data (e.g., using feature subsets), allows for a finer judgement of similarities and disagreements between different outlier

detectors. Where all methods would have a similar performance as judged by ROC AUC, supplementing a ROC analysis with this ranking correlation measure allows for deeper insights in similarities and differences between different methods, parametrization, distance measures, and data.

Along with a score normalization (as, e.g., in ref. 116) and randomization methods such as feature bagging [113], these advances in evaluation methodology are important ingredients for improving ensemble approaches to outlier detection, a field of research that is rather unexplored but is promising to have also big potential for advancements of outlier detection in high-dimensional data. For example, the combination of rankings in different subspaces to a somehow aggregated ranking (the approach pursued in HiCS [112]), could be straightforwardly interpreted as an ensemble approach and, hence, could be possibly explored and understood more detailed on the basis of the theoretical background of ensemble techniques. The normalization [116] and the diversity analysis demonstrated in ref. 117 are obvious extensions that should further improve the results of methods such as HiCS.

6. CONCLUSION

In this survey, we have inspected some typical problems associated with high-dimensional data and discussed the challenges that outlier detection is presented with by the so-called 'curse of dimensionality' and related issues. On the basis of this discussion, we named some problems that should be acknowledged by research on new methods for outlier detection in high-dimensional data.

Furthermore, we discussed existing work, where we distinguish between approaches, on the one hand, not especially interested in subspace-definitions of outliers but just in efficiency and effectiveness issues for high-dimensional data. On the other hand, we discussed specialized methods for subspace outlier detection.

Finally, we gave some pointers to tools and implementations, highlighted the importance of understanding the impact of data preparation, and remarked on an important meta-problem for any unsupervised outlier detection approach, the issue of a proper evaluation procedure.

Overall, we would like to conclude this survey not without emphasizing again that the area of outlier detection specialized for high-dimensional data offers lots of opportunities for improvement. There are just a few approaches around in the literature so far, yet there are many directions to go and problems still to solve. The researcher should, though, be aware of the existing attempts of solution and the associated pitfalls. To support further research on specialized high-dimensional outlier detection methods was our motivation for writing this survey and we are looking forward to seeing new ideas emerging in the research community.

ACKNOWLEDGMENT

A.Z. is partially supported by NSERC.

REFERENCES

[1] M. Markou and S. Singh, Novelty detection: a review-part 1: statistical approaches, Signal Process 83 (2003), 2481–2497.
[2] M. Markou and S. Singh, Novelty detection: A review-part 2: neural network based approaches, Signal Process 83 (2003), 2499–2521.
[3] V. J. Hodge and J. Austin, A survey of outlier detection methodologies, Artif Intell Rev 22 (2004), 85–126.
[4] M. Agyemang, K. Barker, and R. Alhajj, A comprehensive survey of numeric and symbolic outlier mining techniques, Intell Data Anal 10 (2006), 521–538.
[5] A. Patcha and J.-M. Park, An overview of anomaly detection techniques: Existing solutions and latest technological trends, Comput Netw 51 (2007), 3448–3470.
[6] A. S. Hadi, A. H. M. Rahmatullah Imon, and M. Werner, Detection of outliers, Wiley Interdiscip Rev: Comput Stat 1(1) (2009), 57–70.
[7] V. Chandola, A. Banerjee, and V. Kumar, Anomaly detection: a survey, ACM Comput Surv 41 (3) (2009), Article 15, 1–58.
[8] X. Su and C.-L. Tsai, Outlier detection, Wiley Interdiscip Rev: Data Mining Knowledge Disc 1(3) (2011), 261–268.
[9] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, (3rd ed.), Morgan Kaufmann, Amsterdam, 2011.
[10] V. Chandola, A. Banerjee, and V. Kumar, Anomaly detection for discrete sequences: A survey, IEEE Trans Knowledge Data Eng 24(5) (2012), 823–839.
[11] C. C. Aggarwal and P. S. Yu, Outlier detection for high dimensional data, In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Santa Barbara, CA, 2001, 37–46.
[12] 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, 217–235.
[13] K. Lin, H. V. Jagadish, and C. Faloutsos, The TV-Tree: an index structure for high-dimensional data, The VLDB J 3 (1995), 517–542.
[14] S. Brin, Near neighbor search in large metric spaces, In Proceedings of the 21st International Conference on Very Large Data Bases (VLDB), Zurich, Switzerland, 1995, 574–584.
[15] S. Berchtold, D. A. Keim, and H.-P. Kriegel, The X-Tree: an index structure for high-dimensional data, In Proceedings of the 22nd International Conference on Very Large Data Bases (VLDB), Bombay, India, 1996, 28–39.
[16] S. Berchtold, C. Böhm, B. Braunmüller, D. A. Keim, and H.-P. Kriegel, Fast parallel similarity search in multimedia

databases, In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Tucson, AZ, 1997, 1–12.
[17] N. Katayama and S. Satoh, The SR-tree: an index structure for high-dimensional nearest neighbor queries, In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Tucson, AZ, 1997, 369–380.
[18] S. Berchtold, C. Böhm, D. A. Keim, and H.-P. Kriegel, A cost model for nearest neighbor search in high-dimensional data space, In Proceedings of the 16th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Tucson, AZ, 1997, 78–86.
[19] P. Indyk and R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality, In Proceedings of the thirtieth annual ACM symposium on Theory of computing (STOC), Dallas, TX, 1998, 604–613.
[20] R. Weber, H.-J. Schek, and S. Blott, A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces, In Proceedings of the 24th International Conference on Very Large Data Bases (VLDB), New York City, NY, 1998, 194–205.
[21] C. Böhm, S. Berchtold, and D. A. Keim, Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases, ACM Comput Surv 33(3) (2001), 322–373.
[22] V. Pestov, On the geometry of similarity search: dimensionality curse and concentration of measure, Inform Process Lett 73(1–2) (2000), 47–51.
[23] A. Hinneburg, C. C. Aggarwal, and D. A. Keim, What is the nearest neighbor in high dimensional spaces? In Proceedings of the 26th International Conference on Very Large Data Bases (VLDB), Cairo, Egypt, 2000, 506–515.
[24] C. C. Aggarwal, A. Hinneburg, and D. Keim, On the surprising behavior of distance metrics in high dimensional space, In Proceedings of the 8th International Conference on Database Theory (ICDT), London, UK, 2001, 420–434.
[25] D. Francois, V. Wertz, and M. Verleysen, The concentration of fractional distances, IEEE Trans Knowledge Data Eng 19(7) (2007), 873–886.
[26] M. Radovanović, A. Nanopoulos, and M. Ivanović, On the existence of obstinate results in vector space models, In Proceedings of the 33rd International Conference on Research and Development in Information Retrieval (SIGIR), Geneva, Switzerland, 2010, 186–193.
[27] 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, 427–438.
[28] 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, 93–104.
[29] U. Shaft and R. Ramakrishnan, Theory of nearest neighbors indexability, ACM Trans Database Syst (TODS) 31(3) (2006), 814–838.
[30] K. P. Bennett, U. Fayyad, and D. Geiger, Density-based indexing for approximate nearest-neighbor queries, In Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Diego, CA, 1999, 233–243.
[31] H.-P. Kriegel, P. Kröger, and A. Zimek, Clustering high dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering, ACM Trans Knowledge Disc Data (TKDD) 3(1) (2009), 1–58.
[32] P. Kröger and A. Zimek, Subspace clustering techniques. In Encyclopedia of Database Systems, L. Liu and M. T. Özsu, eds. Springer, New York, 2009, 2873–2875.
[33] H.-P. Kriegel, P. Kröger, and A. Zimek, Subspace clustering, Wiley Interdiscip Rev: Data Mining Knowledge Disc 2(4) (2012), 351–364.
[34] I. Assent, Clustering high dimensional data, Wiley Interdiscip Rev: Data Mining Knowledge Disc 2(4) (2012), 340–350.
[35] C. Domeniconi, D. Papadopoulos, D. Gunopulos, and S. Ma, Subspace clustering of high dimensional data, In Proceedings of the 4th SIAM International Conference on Data Mining (SDM), Lake Buena Vista, FL, 2004.
[36] K.-G. Woo, J.-H. Lee, M.-H. Kim, and Y.-J. Lee, FINDIT: a fast and intelligent subspace clustering algorithm using dimension voting, Inform Softw Technol 46(4) (2004), 255–271.
[37] L. Parsons, E. Haque, and H. Liu, Subspace clustering for high dimensional data: a review, ACM SIGKDD Explor 6(1) (2004), 90–105.
[38] M. L. Yiu and N. Mamoulis, Iterative projected clustering by subspace mining, IEEE Trans Knowledge Data Eng 17(2) (2005), 176–189.
[39] G. Liu, J. Li, K. Sim, and L. Wong, Distance based subspace clustering with flexible dimension partitioning, In Proceedings of the 23rd International Conference on Data Engineering (ICDE), Istanbul, Turkey, 2007, 1250–1254.
[40] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, A general framework for increasing the robustness of PCA-based correlation clustering algorithms, In Proceedings of the 20th International Conference on Scientific and Statistical Database Management (SSDBM), Hong Kong, China, 2008, 418–435.
[41] G. Moise and J. Sander, Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering, In Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV, 2008, 533–541.
[42] E. Achtert, C. Böhm, J. David, P. Kröger, and A. Zimek, Global correlation clustering based on the Hough transform, Stat Anal Data Mining 1(3) (2008), 111–127.
[43] C. Zhu, H. Kitagawa, and C. Faloutsos, Example-based robust outlier detection in high dimensional datasets, In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM), Houston, TX, 2005, 829–832.
[44] 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, 444–452.
[45] E. Müller, I. Assent, U. Steinhausen, and T. Seidl, OutRank: ranking outliers in high dimensional data, In Proceedings of the 24th International Conference on Data Engineering (ICDE) Workshop on Ranking in Databases (DBRank), Cancun, Mexico, 2008, 600–603.
[46] C. C. Aggarwal, Re-designing distance functions and distance-based applications for high dimensional data, ACM SIGMOD Record 30(1) (2001), 13–18.

[47] N. Katayama and S. Satoh, Distinctiveness-sensitive nearest-neighbor search for efficient similarity retrieval of multimedia information, In Proceedings of the 17th International Conference on Data Engineering (ICDE), Heidelberg, Germany, 2001, 493–502.

[48] C. C. Aggarwal and P. S. Yu, Finding generalized projected clusters in high dimensional space, In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX, 2000, 70–81.

[49] S. Berchtold, C. Böhm, H. V. Jagadish, H.-P. Kriegel, and J. Sander, Independent Quantization: an index compression technique for high-dimensional data spaces, In Proceedings of the 16th International Conference on Data Engineering (ICDE), San Diego, CA, 2000, 577–588.

[50] H. Jin, B. C. Ooi, H. T. Shen, C. Yu, and A. Y. Zhou, An adaptive and efficient dimensionality reduction algorithm for high-dimensional indexing, In Proceedings of the 19th International Conference on Data Engineering (ICDE), Bangalore, India, 2003, 87–98.

[51] C. C. Aggarwal and P. S. Yu, On high dimensional indexing of uncertain data, In Proceedings of the 24th International Conference on Data Engineering (ICDE), Cancun, Mexico, 2008, 1460–1461.

[52] S. France and D. Carrol, Is the distance compression effect overstated? Some theory and experimentation, In Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM), Leipzig, Germany, 2009, 280–294.

[53] 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, 482–500.

[54] P. Hall, J. S. Marron, and A. Neeman, Geometric representation of high dimension, low sample size data, J Royal Stat Soc: Ser B 67 (2005), 427–444.

[55] F. Murtagh, The remarkable simplicity of very high dimensional data: application of model-based clustering, J Classif 26 (2009), 249–277.

[56] S. France, D. Carrol, and H. Xiong, Distance metrics for high dimensional nearest neighborhood recovery: compression and normalization, Inform Sci 184 (2012), 92–110.

[57] R. J. Durrant and A. Kabán, When is 'nearest neighbour' meaningful: a converse theorem and implications, J Complex 25(4) (2009), 385–397.

[58] F. Korn, B.-U. Pagel, and C. Faloutsos, On the "dimensionality curse" and the "self-similarity blessing", IEEE Trans Knowledge Data Eng 13(1) (2001), 96–111.

[59] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, Nature Rev Cancer 8 (2008), 37–49.

[60] K. Sim, V. Gopalkrishnan, A. Zimek, and G. Cong, A survey on enhanced subspace clustering, Data Mining Knowledge Disc (2012). DOI:10.1007/s10618-012-0258-x.

[61] H.-P. Kriegel and A. Zimek, Subspace clustering, ensemble clustering, alternative clustering, multiview clustering: What can we learn from each other? In MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with KDD 2010, Washington, DC, 2010.

[62] T. Bernecker, M. E. Houle, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, and A. Zimek, Quality of similarity rankings in time series, In Proceedings of the 12th International Symposium on Spatial and Temporal Databases (SSTD), Minneapolis, MN, 2011, 422–440.

[63] M. Verleysen and D. François, The curse of dimensionality in data mining and time series prediction, In Proceedings of the 8th International Work-Conference on Artificial Neural Networks (IWANN), Barcelona, Spain, 2005, 758–770.

[64] D. Hawkins, Identification of Outliers, Chapman and Hall, New York, 1980.

[65] M. Radovanović, A. Nanopoulos, and M. Ivanović, Nearest neighbors in high-dimensional data: the emergence and influence of hubs, In Proceedings of the 26th International Conference on Machine Learning (ICML), Montreal, Canada, 2009. 865–872.

[66] M. Radovanović, A. Nanopoulos, and M. Ivanović, Hubs in space: Popular nearest neighbors in high-dimensional data, J Mach Learn Res 11 (2010), 2487–2531.

[67] M. Radovanović, A. Nanopoulos, and M. Ivanović, Time-series classification in many intrinsic dimensions, In Proceedings of the 10th SIAM International Conference on Data Mining (SDM), Columbus, OH, 2010, 677–688.

[68] N. Tomašev, M. Radovanović, D. Mladenić, and M. Ivanović, The role of hubness in clustering high-dimensional data, In Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Shenzhen, China, 2011, 183–195.

[69] A.-L. Barabási, Scale-free networks: a decade and beyond, Science 325(5939) (2009), 412–413.

[70] E. M. Knorr and R. T. Ng, A unified approach for mining outliers, In Proceedings of the conference of the Centre for Advanced Studies on Collaborative research (CASCON), Toronto, Canada, 1997, 11–23.

[71] E. M. Knorr, R. T. Ng, and V. Tucanov, Distance-based outliers: algorithms and applications, VLDB J 8(3–4) (2000), 237–253.

[72] 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 Discoverys (PKDD), Helsinki, Finland, 2002, 15–26.

[73] G. Kollios, D. Gunopulos, N. Koudas, and S. Berchthold, Efficient biased sampling for approximate clustering and outlier detection in large datasets, IEEE Trans Knowledge Data Eng 15(5) (2003), 1170–1187.

[74] S. D. Bay and M. Schwabacher, Mining distance-based outliers in near linear time with randomization and a simple pruning rule, In Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington, DC, 2003, 29–38.

[75] Y. Pei, O. Zaïane, and Y. Gao, An efficient reference-based approach to outlier detection in large datasets, In Proceedings of the 6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China, 2006, 478–487.

[76] 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, 535–548.

[77] 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, 293–298.
[78] 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, 577–593.
[79] C.-C. Szeto and E. Hung, Mining outliers with faster cutoff update and space utilization, Pattern Recogn Lett 31(11) (2010), 1292–1301.
[80] G. H. Orair, C. Teixeira, Y. Wang, W. Meira, Jr., and S. Parthasarathy, Distance-based outlier detection: consolidation and renewed bearing, Proc VLDB Endowment 3(2) (2010), 1469–1480.
[81] A. Guttman, R-Trees: a dynamic index structure for spatial searching, In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Boston, MA, 1984, 47–57.
[82] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, The R*-Tree: an efficient and robust access method for points and rectangles, In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Atlantic City, NJ, 1990, 322–331.
[83] A. Gionis, P. Indyk, and R. Motwani, Similarity search in high dimensions via hashing, In Proceedings of the 25th International Conference on Very Large Data Bases (VLDB), Edinburgh, Scotland, 1999, 518–529.
[84] P. Indyk, Nearest neighbors in high-dimensional spaces. In Handbook of Discrete and Computational Geometry, chapter 39, (2nd ed.), J. E. Goodman and J. O'Rourke, eds. CRC Press, Boca Raton, 2004, 877–892.
[85] A. Andoni and P. Indyk, Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions, In Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS), Berkeley, CA, 2006, 459–468.
[86] W. B. Johnson and J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space, In Conference in Modern Analysis and Probability, volume 26 of Contemporary Mathematics, American Mathematical Society, 1984, 189–206.
[87] J. Matoušek, On variants of the Johnson–Lindenstrauss lemma, Random Struct Algor 33(2) (2008), 142–156.
[88] D. Achlioptas, Database-friendly random projections, In Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Santa Barbara, CA, 2001.
[89] D. Achlioptas, Database-friendly random projections: Johnson–Lindenstrauss with binary coins, J Comput Syst Sci 66 (2003), 671–687.
[90] S. Venkatasubramanian and Q. Wang, The Johnson–Lindenstrauss transform: an empirical study, In Proceedings of the Workshop on Algorithm Engineering and Experiments (ALENEX) SIAM, San Francisco, CA, 2011, 164–173.
[91] A. Kabán, On the distance concentration awareness of certain data reduction techniques, Pattern Recogn 44(2) (2011), 265–277.
[92] G. Peano, Sur une courbe, qui remplit toute une aire plane, Mathe Ann 36(1) (1890), 157–160.
[93] G. M. Morton, A computer oriented geodetic data base and a new technique in file sequencing, Technical report, International Business Machines Co., 1966.
[94] D. Hilbert, Ueber die stetige Abbildung einer Linie auf ein Flächenstück, Math Ann 38(3) (1891), 459–460.
[95] I. Kamel and C. Faloutsos, Hilbert R-tree: an improved R-tree using fractals, In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago de Chile, Chile, 1994, 500–509.
[96] N. H. Vu and V. Gopalkrishnan, Efficient pruning schemes for distance-based outlier detection, In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Bled, Slovenia, 2009, 160–175.
[97] A. Ghoting, S. Parthasarathy, and M. E. Otey, Fast mining of distance-based outliers in high-dimensional datasets, Data Mining Knowledge Disc 16(3) (2008), 349–364.
[98] J. A. Hartigan, Clustering Algorithms, John Wiley & Sons, New York, 1975.
[99] 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, 410–421.
[100] T. de Vries, S. Chawla, and M. E. Houle, Finding local anomalies in very high dimensional space, In Proceedings of the 10th IEEE International Conference on Data Mining (ICDM), Sydney, Australia, 2010, 128–137.
[101] 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.
[102] 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, 315–326.
[103] F. Angiulli and C. Pizzuti, Outlier mining in large high-dimensional data sets, IEEE Trans Knowledge Data Eng 17(2) (2002), 203–215.
[104] B. Moon, H. V. Jagadish, C. Faloutsos, and J. H. Saltz, Analysis of the clustering properties of the Hilbert space-filling curve, IEEE Trans Knowledge Data Eng 13(1) (2001), 124–141.
[105] S. Liao, M. A. Lopez, and S. T. Leutenegger, High dimensional similarity search with space filling curves, In Proceedings of the 17th International Conference on Data Engineering (ICDE), Heidelberg, Germany, 2001, 615–622.
[106] L. Yang, Distance-preserving dimensionality reduction, Wiley Interdiscip Rev: Data Mining Knowledge Disc 1(5) (2011), 369–380.
[107] R. Gnanadesikan and J. R. Kettenring, Robust estimates, residuals, and outlier detection with multiresponse data, Biometrics 28(1) (1972), 81–124.
[108] I. T. Jolliffe, Principal Component Analysis (2nd ed.), Springer, New York, 2002.
[109] P. Filzmoser, R. Maronna, and M. Werner, Outlier identification in high dimensions, Comput Stat Data Anal 52(3) (2008), 1694–1711.
[110] J. A. Lee and M. Verleysen, Quality assessment of dimensionality reduction: rank-based criteria, Neurocomputing 72(7–9) (2009), 1431–1443.
[111] N. H. Vu and V. Gopalkrishnan, Feature extraction for outlier detection in high-dimensional spaces, J Mach Learn Res, Proc Track 10 (2010), 66–75.

[112] 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.

[113] A. Lazarevic and V. Kumar, Feature bagging for outlier detection, In Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Chicago, IL, 2005, 157–166.

[114] H. V. Nguyen, H. H. Ang, and V. Gopalkrishnan, Mining outliers with ensemble of heterogeneous detectors on random subspaces, In Proceedings of the 15th International Conference on Database Systems for Advanced Applications (DASFAA), Tsukuba, Japan, 2010, 368–383.

[115] J. Gao and P.-N. Tan, Converting output scores from outlier detection algorithms into probability estimates, In Proceedings of the 6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China, 2006. 212–221.

[116] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, Interpreting and unifying outlier scores, In Proceedings of the 11th SIAM International Conference on Data Mining (SDM), Mesa, AZ, 2011, 13–24.

[117] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel, On evaluation of outlier rankings and outlier scores, In Proceedings of the 12th SIAM International Conference on Data Mining (SDM), Anaheim, CA, 2012, 1047–1058.

[118] E. M. Knorr and R. T. Ng, Finding intensional knowledge of distance-based outliers, In Proceedings of the 25th International Conference on Very Large Data Bases (VLDB), Edinburgh, Scotland, 1999, 211–222.

[119] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, Density-based clustering, Wiley Interdiscip Rev: Data Mining and Knowledge Disc 1(3) (2011), 231–240.

[120] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago de Chile, Chile, 1994, 487–499.

[121] I. Färber, S. Günnemann, H.-P. Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, and A. Zimek, On using class-labels in evaluation of clusterings, In MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with KDD 2010, Washington, DC, 2010.

[122] H.-P. Kriegel, E. Schubert, and A. Zimek, Evaluation of multiple clustering solutions, In 2nd MultiClust Workshop: Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with ECML PKDD 2011, Athens, Greece, 2011, 55–66.

[123] C. C. Aggarwal and P. S. Yu, An effective and efficient algorithm for high-dimensional outlier detection, VLDB J 14(2) (2005), 211–221.

[124] J. Zhang, M. Lou, T. W. Ling, and H. Wang, HOS-miner: a system for detecting outlying subspaces of high-dimensional data, In Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), Toronto, Canada, 2004, 1265–1268.

[125] 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, 138–152.

[126] I. Assent, R. Krieger, E. Müller, and T. Seidl, Subspace outlier mining in large multimedia databases, In Parallel Universes and Local Patterns, 2007.

[127] I. Assent, R. Krieger, E. Müller, and T. Seidl, DUSC: dimensionality unbiased subspace clustering, In Proceedings of the 7th IEEE International Conference on Data Mining (ICDM), Omaha, NE, 2007, 409–414.

[128] I. Assent, R. Krieger, E. Müller, and T. Seidl, EDSC: efficient density-based subspace clustering, In Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM), Napa Valley, CA, 2008, 1093–1102.

[129] 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, 831–838.

[130] 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, Canada, 2010, 1629–1632.

[131] E. Müller, M. Schiffer, and T. Seidl, Statistical selection of relevant subspace projections for outlier ranking, In Proceedings of the 27th International Conference on Data Engineering (ICDE), Hannover, Germany, 2011, 434–445.

[132] H.-P. Kriegel, P. Kröger, M. Schubert, and Z. Zhu, Efficient query processing in arbitrary subspaces using vector approximations, In Proceedings of the 18th International Conference on Scientific and Statistical Database Management (SSDBM), Vienna, Austria, 2006, 184–190.

[133] W. Müller and A. Henrich, Faster exact histogram intersection on large data collections using inverted VA-files, In Proceedings of the 3rd International Conference on Image and Video Retrieval (CIVR), Dublin, Ireland, 2004, 455–463.

[134] X. Lian and L. Chen, Similarity search in arbitrary subspaces under L_p-norm, In Proceedings of the 24th International Conference on Data Engineering (ICDE), Cancun, Mexico, 2008, 317–326.

[135] T. Bernecker, T. Emrich, F. Graf, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, and A. Zimek, Subspace similarity search using the ideas of ranking and top-k retrieval, In Proceedings of the 26th International Conference on Data Engineering (ICDE) Workshop on Ranking in Databases (DBRank), Long Beach, CA, 2010, 4–9.

[136] T. Bernecker, T. Emrich, F. Graf, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, and A. Zimek, Subspace similarity search: Efficient k-nn queries in arbitrary subspaces, In Proceedings of the 22nd International Conference on Scientific and Statistical Database Management (SSDBM), Heidelberg, Germany, 2010, 555–564.

[137] T. Bernecker, F. Graf, H.-P. Kriegel, C. Moennig, and A. Zimek, BeyOND –unleashing BOND, In Proceedings of the 37th International Conference on Very Large Data Bases (VLDB) Workshop on Ranking in Databases (DBRank), Seattle, WA, 2011, 34–39.

[138] 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, 1649–1652.

[139] V. A. Epanechnikov, Non-parametric estimation of a multivariate probability density, Theory Prob Appl 14(1) (1969), 153–158.

[140] A. Zimek, Correlation clustering, PhD thesis, Ludwig-Maximilians-Universität München, Munich, Germany, 2008.
[141] A. Zimek, Correlation clustering, ACM SIGKDD Explor 11(1) (2009), 53–54.
[142] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek, Deriving quantitative models for correlation clusters, In Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA, 2006, 4–13.
[143] E. Müller, M. Schiffer, P. Gerwert, M. Hannen, T. Jansen, and T. Seidl, SOREX: Subspace outlier ranking exploration toolkit, In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Barcelona, Spain, 2010, 607–610.
[144] E. Müller, S. Günnemann, I. Assent, and T. Seidl, Evaluating clustering in subspace projections of high dimensional data, In Proceedings of the 35th International Conference on Very Large Data Bases (VLDB), Lyon, France, 2009, 1270–1281.
[145] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explor 11(1) (2009), 10–18.
[146] E. Achtert, H.-P. Kriegel, L. Reichert, E. Schubert, R. Wojdanowski, and A. Zimek, Visual evaluation of outlier detection models, In Proceedings of the 15th International Conference on Database Systems for Advanced Applications (DASFAA), Tsukuba, Japan, 2010, 396–399.
[147] E. Achtert, A. Hettab, H.-P. Kriegel, E. Schubert, and A. Zimek, Spatial outlier detection: data, algorithms, visualizations. In Proceedings of the 12th International Symposium on Spatial and Temporal Databases (SSTD), Minneapolis, MN, 2011, 512–516.
[148] K. Zhang, M. Hutter, and H. Jin, A new local distance-based outlier detection approach for scattered real-world data, In Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand, 2009, 813–822.
[149] S. Shekhar, C.-T. Lu, and P. Zhang, A unified approach to detecting spatial outliers, GeoInformatica 7(2) (2003), 139–166.
[150] S. Chawla and P. Sun, SLOM: a new measure for local spatial outliers, Knowledge Inform Syst (KAIS) 9(4) (2006), 412–429.
[151] F. Chen, C.-T. Lu, and A. P. Boedihardjo, GLS-SOD: a generalized local statistical approach for spatial outlier detection, In Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington, DC, 2010, 1069–1078.
[152] E. Achtert, S. Goldhofer, H.-P. Kriegel, E. Schubert, and A. Zimek, Evaluation of clusterings –metrics and visual support, In Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, DC, 2012.
[153] C. C. Aggarwal and P. S. Yu, Redefining clustering for high-dimensional applications, IEEE Trans Knowledge Data Eng 14(2) (2002), 210–225.
[154] G. Moise, A. Zimek, P. Kröger, H.-P. Kriegel, and J. Sander, Subspace and projected clustering: experimental evaluation and analysis, Knowledge and Inform Syst (KAIS) 21(3) (2009), 299–326.