2010 IEEE International Conference on Data Mining

Finding Local Anomalies in Very High Dimensional Space

Timothy de Vries, Sanjay Chawla
School of Information Technologies
The University of Sydney, Australia
timothy.devries@gmail.com, chawla@it.usyd.edu.au

Michael E. Houle
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
meh@nii.ac.jp

***Abstract*—Time, cost and energy efficiency are critical factors for many data analysis techniques when the size and dimensionality of data is very large. We investigate the use of Local Outlier Factor (LOF) for data of this type, providing a motivating example from real world data. We propose Projection-Indexed Nearest-Neighbours (PINN), a novel technique that exploits extended nearest neighbour sets in the a reduced dimensional space to create an accurate approximation for k-nearest-neighbour distances, which is used as the core density measurement within LOF. The reduced dimensionality allows for efficient sub-quadratic indexing in the number of items in the data set, where previously only quadratic performance was possible. A detailed theoretical analysis of Random Projection (RP) and PINN shows that we are able to preserve the density of the intrinsic manifold of the data set after projection. Experimental results show that PINN outperforms the standard projection methods RP and PCA when measuring LOF for many high-dimensional real-world data sets of up to 300000 elements and 102600 dimensions.**

***Keywords*-Anomaly detection; Dimensionality reduction;**

I. Introduction

The amount of data being collected and stored is continually increasing. Time, cost, and energy efficiency are critical factors to consider when such data is analysed. One technique with scalability issues in large, high-dimensional data sets is Local Outlier Factor (LOF) [1], a formula that measures the degree to which a data point is an outlier with respect to its local neighbourhood. The outliers are 'local' in the sense that their determination does not depend on knowledge of the global distribution of the data set. Our goal is to develop an alternative to LOF that is capable of finding meaningful local outliers in large data sets with very high representational dimensionality.

A. Local Outliers vs. Global Outliers

Local outliers are in some sense the most general form of distance-based outliers, as Figure 1 illustrates. This sample data set contains a set of four visibly-distinguishable outliers $P = \{p_1, p_2, p_3, p_4\}$. The question of which of these points will be reported as outliers depends upon the definition we use.

- *Classical*: A point can be declared to be an outlier if its distance from the mean is sufficiently large, a definition that would fail to report p_2.
- *PCA based*: In applications involving the use of Principal Component Analysis (PCA) [2], an outlier is usually declared if the point is sufficiently far away from the subspace spanned by the eigenvectors corresponding to the highest eigenvalues. For this example, p_2 and p_3 would likely not be discovered as they lie in or near the subspace spanned by the dominant eigenvector.
- *Distance based*: A point can be declared to be an outlier if its distance to its k-th nearest neighbour is sufficiently large [3]. In this example, the points p_3 and p_4 would likely not be reported as high-ranking outliers due to the existence of many points in the sparse cluster that are further from their own neighbours. In general, definitions based on k-nearest-neighbour distances have difficulty in discovering outliers within data sets having great variations in density.
- *Local density based*: A point is declared to be an outlier if the distance to its k-th nearest neighbour is large relative to the distances of its neighbours to their own k-th nearest neighbours [1]. Under this definition, all of the elements of P could conceivably be reported as outliers.

Despite their use of k-nearest neighbour distances, local outlier measures are often considered to be density-based rather than distance-based: in the case of LOF, distance

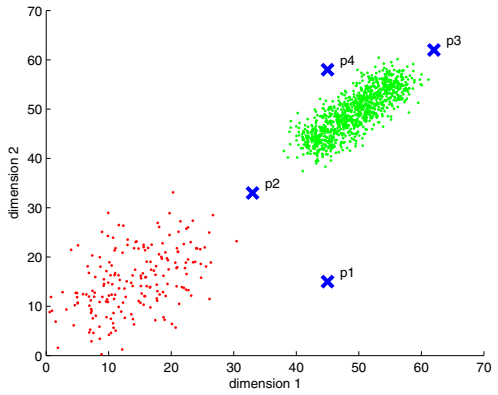

Figure 1. A data set with varying density and four outliers p_1 to p_4.

1550-4786/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDM.2010.151

Figure 2. The top 12 global distance-based outliers from the Amsterdam data set, determined using 20-nearest-neighbour distances.

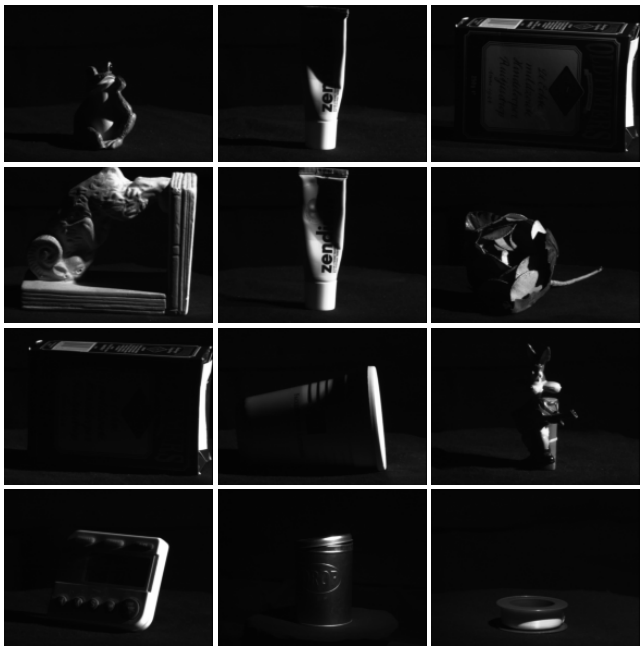

Figure 3. The top 12 local outliers from the Amsterdam data set, ranked by LOF values computed after a random projection to a 20-dimensional space.

values are used to estimate density, which in turn are used to compare the density of points in the vicinity of the query item with the average density of points in the vicinity of the k-nearest neighbours of the query item. If the density value for the query item is significantly less than this average, then the query is deemed to be a local outlier. Local outlier formulations can be thought of as a generalisation of global outlier formulations, as global outliers will typically also be local outliers, but not vice versa. Local outlier methods tend to be more computationally expensive than global methods.

B. Example

While the concept of local outliers is an important one, most explanations appearing in the research literature have made use of examples based on synthetic data. We thus provide a motivating real-world example using the Amsterdam Library of Object Images [4], a data set containing 24000 images of 1000 distinct objects, the 24 images of each object taken from 8 different orientations under 3 different illumination directions. We selected a greyscale image resolution of 192×144, yielding a high representational dimensionality of 27648 pixel features.

Figure 2 shows the top 12 global outliers of this dataset, discovered using a distance-based outlier technique with $k = 20$ [5]. These images contain large brightly-lit areas and unusual shapes, features that make these images stand out distinctly from the rest of the data set as a whole. Figure 3 shows the top 12 local outliers, as ranked using LOF values

Figure 4. 15 images of a single object from the Amsterdam data set, ranked by LOF values computed after random projection to a 20-dimensional space. 9 additional images are not shown as they have the lowest LOF values and are highly similar to the images in the bottom three rows.

computed after a random projection to a 20-dimensional space (see the experiments in Section V for more details). The example illustrates the markedly different characteristics of local outliers as compared to global outliers. In most cases, the 24 images associated with an individual object form a cluster, as the variation among images of an object are much less significant than the variation between images of different objects.

In this example, the highest-ranked local outliers are those object images that least resemble other images of the same object, due to differences in illumination intensity or direction, or occlusion of some areas of the images such as from shadows. Images for one such cluster exhibiting a high degree of image variation, and containing highly-ranked local outliers, are shown in Figure 4. This example suggests that local outlier detection may find uses for object matching or classification in object recognition systems, as the local outlier rankings for the query image, as well as the similarity rankings, would enable a decisive classification under varying lighting conditions.

C. Outlier Detection and Scalability

The example presented above involved a moderate number of small greyscale images that nevertheless have a very high representational dimensionality. Although existing local outlier detection methods could perhaps cope with such a small example, they would be hard-pressed to cope with data sets consisting of many millions (or more) of large-sized colour images. Other application areas where performance problems are frequently encountered due to the large data set sizes and dimensionalities include text and medical data, for which the potential benefits of local outlier analysis has not yet been adequately investigated. In general, inherent scalability difficulties generally prevent the use of standard local outlier techniques to obtain useful results on such large, high-dimensional data sets.

Much research has been carried out into the use of distance-based outlier detection for high dimensional data sets [6]. The greatest challenge is how to deal with the 'curse of dimensionality': as the data dimension grows, distance measures lose their discriminative ability, and conventional indexing techniques are no longer effective in managing the search for neighbours required by distance- and density-based outlier detection. Without an effective index, LOF would require $O(n^2)$ time for a data set of size n, using sequential search for neighbours. In practice, however, there are many real-world data sets that have high representational dimensionality but low intrinsic dimensionality [7].

A basic taxonomy of optimisation solutions for global and local outlier detection at low and high dimensions is shown in Figure 5, where dotted lines indicate areas the literature has not covered. Papadimitriou et. al. [8] have applied a sampling strategy and standard indexing to their local outlier definition which is highly similar to LOF, but

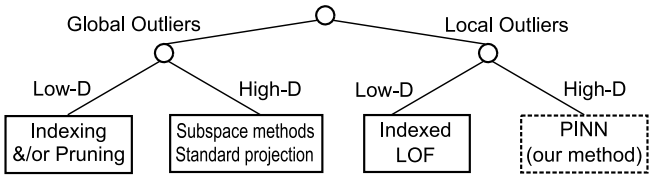

Figure 5. A basic taxonomy for the optimisation of global and local outlier methods on low (Low-D) and high (High-D) dimensional data sets.

only up to 20 dimensions. Jin, Tung and Han [9] have developed a highly-specialised indexing method for LOF, but have tested it only for synthetic data of up to 20 dimensions. Nevertheless, this approach may be applicable in place of other traditional indexing methods such as kD-trees. The pruning approach used by Bay and Schwabacher [5] cannot be readily applied to LOF due to the large amount of overlapping density computation required to find the LOF of a single point. Approximate similarity search methods such as LSH [10] and SASH [11] can potentially be used for the nearest-neighbour search required by distance and density-based methods. Kriegel et. al. [12] describe a subspace LOF method that works well when the relevant outlier subspace is known, but was tested only with sets of up to a few hundred items and dimensions.

D. Contributions

In this paper, we propose a projective local outlier detection method based on LOF, which we call *Projection-Indexed Nearest-Neighbour* (PINN), and compare its effectiveness with alternative methods based on PCA and on Random Projection (RP). PINN goes beyond the simple strategy of 'project and compute': it estimates the distances required by LOF by computing them within a reduced-dimensional projection space, where the computational costs associated with k-nearest-neighbour search can be expected to be significantly smaller than in the original space. We also prove that when the projection matrix is determined randomly, the LOF estimation error may be bounded with very high probability, in terms of a measure of the intrinsic dimension of the target projection space. Our experimentation shows that the use of PINN for LOF causes a dramatic reduction of data loss and therefore a large improvement in accuracy over the existing projection techniques of RP and PCA, while retaining the indexing scalability benefits of these techniques. To the best of our knowledge PINN is the first sub-quadratic heuristic for local anomaly detection in very high dimensional space.

II. Background

A. Principal Component Analysis

One very popular method for preparing high-dimensional sets for data analysis involves the projection of the data to a lower-dimensional subspace. If it is required to reduce a m

dimensional data set to t dimensions, the projection process can be denoted as $Y \leftarrow XR$, where X is the original n by m data matrix, R is an m by t projection matrix, and Y is the resultant n by t matrix. For a given point $x \in X$, we will denote its image under the projection as $x' = xR$.

Many techniques for determining suitable projection spaces have been proposed; perhaps the most popular and well-established of these approaches involves the use of Principal Component Analysis (PCA) [2]. The basis of an m-dimensional subspace is constructed by computing m eigenvectors corresponding to the m largest eigenvalues of the covariance matrix of the data. A well known drawback of PCA is the high cost associated with the computation of eigenvectors — computing the full set of eigenvectors of a set of n data points in m dimensions using the traditional Cyclic Jacobi method requires $\Theta(m^3+m^2n)$ time [2]. However, the first t eigenvectors can be computed sequentially in $O(m^2tn)$ time using Gram-Schmidt decomposition [13]. For all practical purposes, PCA is generally not feasible for very high dimensional applications.

B. Random Projection

The high computational expense associated with PCA has lead to investigations of alternative methods for determining a basis for data projection. A simple and computationally inexpensive alternative is the use of a random basis of projection [14].

The classical Johnson-Lindenstrauss lemma states that under certain conditions, there exists a projection that approximately preserves pairwise euclidean distances between data points [15] [16]. The dimension of the projection space depends logarithmically on the number of data points, and is independent of the original dimension. More recently, Achlioptas [17] showed that a simple random projection strategy can (asymptotically) achieve this dimensionality reduction with very high probability, as follows. Let the entries of the projection matrix R be generated randomly and independently as

$$r_{ij} = \sqrt{s} \begin{cases} +1 & \text{with probability} \quad \frac{1}{2s} \\ 0 & \text{with probability} \quad 1 - \frac{1}{s} \\ -1 & \text{with probability} \quad \frac{1}{2s}. \end{cases}$$

The parameter s represents sparsity, causing random projection to sample approximately $\frac{1}{s}$ of the total attribute space for each new projected dimension.

Lemma 1 ([17]): If for some choice of $\gamma > 0$ the reduced dimension t satisfies

$$t \geq \frac{4+2\gamma}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}} \ln n,$$

then with probability at least $1 - n^{-\gamma}$, the projection $Y \leftarrow XR$ approximately preserves euclidean distances for all data points in X. More precisely, for all $u, v \in X$ with projection points $u', v' \in Y$,

$$(1-\epsilon) \cdot d(u,v) \leq d(u',v') \leq (1+\epsilon) \cdot d(u,v),$$

where $d(u,v)$ and $d(u',v')$ denote the euclidean distance between the points in their respective spaces.

In [17], Achlioptas argues that this projection scheme (which we will refer to as RP) can be implemented efficiently within a database framework, giving it an advantage over PCA. Projected data may also be incrementally updated for RP (but not PCA) by replacing entries in a sliding window, but not for PCA.

C. Local Outlier Factor

Determination of local density-based outliers can be made using the well-known Local Outlier Factor (LOF) measure [1], which assesses the degree to which a point is an outlier relative to other points in their immediate neighbourhood.

Let $\mathbb{R}^m$ be the m-dimensional euclidean space and $D \subset \mathbb{R}^m$. We denote the euclidean metric between the points p and q as $d(p,q)$. Our objective is to efficiently identify local outliers in D.

Definition 1: For a fixed k, let $d_k(p)$ be the distance of p to its k-th nearest neighbour. Then the k-nearest neighbour set of p is defined as $N_k(p) = \{q \in D \setminus \{p\} \,|\, d(p,q) \leq d_k(p)\}$.

Definition 2: The relative density of a point p is defined as $rd(p) = \frac{1}{d_k(p)}$, the reciprocal of the distance of p to its k-th nearest neighbour. Then the Local Outlier Factor is defined as:

$$LOF(p) = \frac{\frac{1}{|N_k(p)|} \sum_{q \in N_k(p)} rd(q)}{rd(p)}$$

$LOF(p)$ is the ratio of the relative density of p and the average relative density of its k-th nearest neighbours.

D. Intrinsic dimensionality

In [18], Karger and Ruhl introduced a measure of intrinsic dimensionality as a means of analyzing the performance of a local search strategy for handling nearest neighbour queries. We give a slight generalization of their measure here. Let $B_D(v,r) = \{w \in D \mid d(v,w) \leq r\}$ be the set of elements of D contained in the closed ball of radius r centered at $v \in D$. D is said to have (b,c,Δ)-expansion if for all $q \in D$ and $r > 0$, $|B_D(q,r)| \geq b \Longrightarrow |B_D(q,cr)| \leq \Delta \cdot |B_D(q,r)|$. The (b,c)-*expansion rate* of D is the minimum value of Δ such that the above condition holds, subject to the choice of minimum ball set size b . In their analysis, Karger and Ruhl chose $b = \mathbf{O}(\log |D|)$ and $c = 2$.

One can consider the value $\log_2 \Delta$ to be a measure of the intrinsic dimension, by observing that for the euclidean distance metric in $\mathbb{R}^m$, doubling the radius of a sphere would increase its volume by a factor of 2^m, and thus the sphere would contain proportionately as many points

from a uniformly-distributed set. However, as pointed out in [18], low-dimensional subsets in high-dimensional spaces can have very low expansion rates, whereas even for one-dimensional data the expansion rate can be linear in the size of S. The (b, c)-*expansion dimension* $\log_2 \Delta$ is also not a robust measure of intrinsic dimensionality, in that the addition or deletion of even a single point can cause an arbitrarily-large increase or decrease.

III. Density-Preserving Random Projection

Although the use of LOF for outlier detection is well-established, standard implementations are inherently very computationally intensive in high-dimensional settings due to the difficulties surrounding the efficient computation of k-nearest-neighbour sets for high-dimensional data. Projection of the data into a lower-dimensional subspace has the potential for speeding up the computation of neighbourhoods; however, the question arises as to whether neighbourhood information is sufficiently well preserved by the projection. In this section, we address this question in the affirmative, by showing that the random projections considered in [17] not only preserve distances approximately with high probability, but that they also preserve distances from points to their k-nearest neighbors. Moreover, these distances are preserved under the projections even though the neighbor relationships themselves may not be — if v is the k-nearest neighbour of u, v' may not be the k-nearest neighbour of u'.

This preservation of k-nearest neighbour distance then allows us to state a lemma that gives conditions under which the k-nearest-neighbour set of a point p in the original space $\mathbb{R}^m$ is contained in a larger neighbourhood set based at p' in the projection space $\mathbb{R}^t$. These conditions depend on the measure of the intrinsic dimensionality of the projection space introduced by Karger and Ruhl. The lemma will be then serve as the foundation of a fast, probabilistically-correct estimation of the LOF for the original space, to be presented in Section IV.

A. Preservation of k-nearest-neighbour distance under RP

We start by showing that random projections can preserve k-nearest neighbor distances with high probability, regardless of whether the neighbour sets themselves are preserved by the projection. For a given point $p \in D$ associated with projection point $p' \in D'$ (where D' is the image of D under the projection), let $N_k(p)$ be the k-nearest neighbour set of p in D, and let $N_k(p')$ be the k-nearest neighbour set in D'.

Lemma 2: Consider any projection satisfying the Johnson-Lindenstrauss distance bounds

$$(1-\epsilon)\cdot d(x,y) \le d(x',y') \le (1+\epsilon)\cdot d(x,y)$$

for all $x, y \in D$. Then for all points $p \in D$ with associated projection point p', and for any choice of $k \ge 1$,

$$(1-\epsilon)\cdot d_k(p) \le d_k(p') \le (1+\epsilon)\cdot d_k(p).$$

Proof: Let v_k be the k-th nearest neighbour of p, and w'_k be the k-th nearest neighbour of p'.
We first show that $d_k(p') \le (1+\epsilon)\cdot d_k(p)$. There are two cases to consider:

1) $d(p', w'_k) \le d(p', v'_k)$.
   Since $d(p', v'_k) \le (1+\epsilon)\cdot d(p, v_k)$, we have $d_k(p') = d(p', w'_k) \le (1+\epsilon)\cdot d_k(p)$.
2) $d(p', w'_k) > d(p', v'_k)$.
   Since w'_k is the k-th nearest neighbour of p', then at most $k-1$ members of $N_k(p)$ can have image points under the projection with distances to p' strictly less than $d(p', w'_k)$. Therefore, there exists $v \in N_k(p)$ such that $d(p', w'_k) \le d(p', v')$. This implies that

$$\begin{aligned} d_k(p') &= d(p', w'_k) \\ &\le d(p', v') \\ &\le (1+\epsilon)\cdot d(p, v) \\ &\le (1+\epsilon)\cdot d(p, v_k) \\ &\le (1+\epsilon)\cdot d_k(p) \end{aligned}$$

Using a symmetric argument and noting that under the assumptions of Lemma 1 we have

$$d(x,y) \le \frac{1}{1-\epsilon}\cdot d(x',y')$$

holding for all $x, y \in D$, we can show that

$$d_k(p) \le \frac{1}{1-\epsilon}\cdot d_k(p')$$

Combining the bounds, the lemma follows. ■

The distance bounds established by the lemma indicate that the relative density is also preserved by the random projection with very high probability.

Corollary 1: Consider a random projection satisfying the conditions of Lemma 1. Then with probability at least $1 - n^{-\gamma}$ (where $n = |D|$), for all points $p \in D$ with associated projection point p',

$$\frac{1}{1+\epsilon}\cdot rd(p) \le rd(p') \le \frac{1}{1-\epsilon}\cdot rd(p).$$

B. Preservation of neighbour sets

In the following discussion, for the purposes of measuring the intrinsic dimension of a set D, we will assume that the distances from any point $p \in D$ to the remaining points of D are all distinct. In practice, the distinctness of distances can be realized using a symbolic perturbation scheme, where ties are broken in some systematic manner. More details on such perturbation schemes can be found in [19].

Let RP be a random projection selected according to the conditions of Lemma 1. For any point $p \in D$, let us consider a neighbourhood $N_h(p')$ surrounding the projection $p' \in D'$, where D is the image of D under the projection. The items of $N_h(p')$ are the images under RP of some subset of D, which we will denote by

$$\mathrm{RP}^{-1}(N_h(p')) = \{x \in D | \, \mathrm{RP}(x) \in N_h(p')\}.$$

Stated another way, the original k-nearest neighbourhood of p is mapped by the projection to a subset of the h-nearest neighbourhood of p'. The following lemma shows that if h chosen to be sufficiently large, the set $\text{RP}^{-1}(N_h(p'))$ captures all of the k-nearest neighbours of p with very high probability. The conditions on the choice of h can be expressed in terms of the intrinsic dimensionality of D'.

Lemma 3: Consider a random projection RP satisfying the conditions of Lemma 1. For a given value of $k \leq 1$, let Δ be the $(k, \frac{1+\epsilon}{1-\epsilon})$-expansion rate of D', the image of D under RP. Then with probability at least $1 - n^{-\gamma}$ (where $n = |D|$), for all points $p \in D$ with associated projection point p',

$$N_k(p) \subseteq \text{RP}^{-1}(N_h(p')),$$

where $h = \lfloor \Delta k \rfloor$.

Proof: Let v_k be the k-th nearest neighbour of p, and let w'_k be the k-th nearest neighbour of p'. Then for all $v \in N_k(p)$,

$$\begin{aligned} d(p', v') &\leq (1+\epsilon) \cdot d(p, v) \text{ (from Lemma 1)} \\ &\leq (1+\epsilon) \cdot d(p, v_k) \\ &\leq \frac{1+\epsilon}{1-\epsilon} \cdot d(p', w') \text{ (from Lemma 2)} \\ &\leq \frac{1+\epsilon}{1-\epsilon} \cdot d_k(p'). \end{aligned}$$

Since $v' \in B(p', \frac{1+\epsilon}{1-\epsilon} \cdot d_k(p'))$ for all $v \in N_k(p)$, the expansion rate bound applies, yielding $k \leq |B(p', \frac{1+\epsilon}{1-\epsilon} \cdot d_k(p'))| \leq \Delta |B(p', d_k(p'))|$. This implies that $v' \in B(p', \frac{1+\epsilon}{1-\epsilon} \cdot d_k(p')) \subseteq N_h(p')$, and therefore $v \in \text{RP}^{-1}(N_h(p))$ as required. ■

IV. Projection-Indexed Nearest-Neighbour (PINN)

As stated earlier, Lemma 2 does not in itself guarantee that neighbourhood sets are preserved by the random projection. Since $LOF(p)$ computes an average of relative densities based at the neighbours of p, changes to the membership of neighbourhood set could have a large impact on the value. For this reason, it may not be appropriate to approximate $LOF(p)$ by computing and using the value of $LOF(p')$. As an example, consider the situation shown in Figure 6. Here, the distance between p and one of p's neighbours, q_2, may increase by $1+\epsilon$ under RP, and the distance between p and q_3 may decrease by $1-\epsilon$. After projection, p would have q_3 as a neighbour instead of q_2, resulting in a substantial increase in the relative density associated with that neighbour. If many neighbours are replaced in this manner as a result of the progression, the value of $LOF(p')$ could vary widely from that of $LOF(p)$.

As an alternative to the estimation of $LOF(p)$ by $LOF(p')$, we instead propose that the calculation of $LOF(p)$ be computed in the original space, but estimated using neighbourhood memberships and distances as determined after projection. Theorem 1 suggests that instead of paying the high cost of k-nearest-neighbour computation in the original space, we can instead determine a larger neighbourhood within the projection space, and reverse the mapping to obtain estimates of the neighbours in the original space. If a sufficient number of neighbours are considered in the projection space (where the number depends on the intrinsic dimension of D'), the true original set of k neighbours can be recuperated with very high probability.

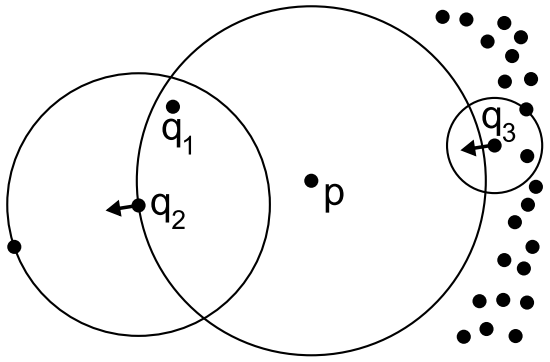

Figure 6. An illustration of possible neighbour replacement under RP for 2-nearest-neighbour LOF of p. The difference in radius around q_2 and q_3 show that the density change due to neighbour replacement may be arbitrarily large.

The proposed approach is summarized in Algorithm 1 and illustrated in Figure 7. Corollary 1 and Lemma 3 together imply the following main result:

Theorem 1: Consider a random projection RP satisfying the conditions of Lemma 1. For a given value of $k \leq 1$, let Δ be the $(k, \frac{1+\epsilon}{1-\epsilon})$-expansion rate of D', the image of D under RP. If h is chosen to be $\lfloor \Delta k \rfloor$, then with probability at least $1 - n^{-\gamma}$ (where $n = |D|$), for every $p \in D$, the estimate

$$\overline{LOF}(p) = \frac{\frac{1}{k} \sum_{q \in \overline{N}_k(p)} rd(q)}{rd(p)}$$

computed by Algorithm 1 satisfies

$$\frac{1-\epsilon}{1+\epsilon} \cdot LOF(p) \leq \overline{LOF}(p) \leq \frac{1+\epsilon}{1-\epsilon} \cdot LOF(p).$$

In practice, the expansion rates of the data set are too expensive to compute, and are generally treated as unknown. However, the theorem does indicate that for a fixed choice of h, the algorithm performs best when the intrinsic dimension of the data set is small. As we shall see in the next section, even the small fixed values of $h = \{2k, 3k\}$ worked to great effect in our experimentation. The parameter h can be increased for an improvement in accuracy, albeit with diminishing returns, at the cost of increased processing time linear with h. Personal preference can therefore be used to choose a value of h, with potential variation in benefits for

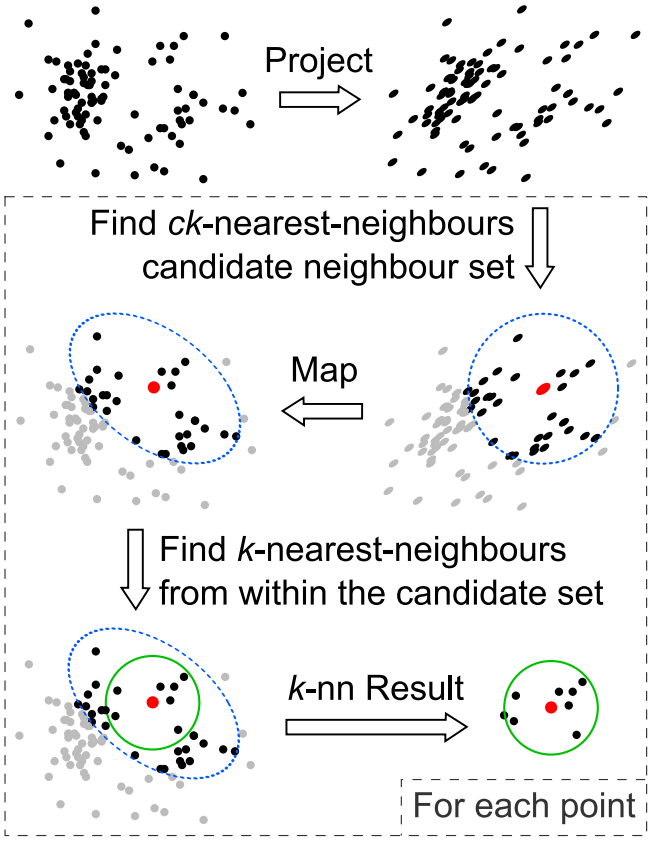

Figure 7. The stages of the PINN algorithm (see Algorithm 1 for details).

different kinds of data sets, visible as the gap between $2k$ and $3k$ in the accuracy graphs to follow.

PINN uses the projected data as a form of index to gather candidate neighbour points that are likely to include the true nearest neighbour set. This method avoids the $O(n^2m)$ computation required for the full high-dimensional nearest-neighbour search and substitutes it for a preprocessing Random Projection step of $O(mtn)$, an indexable projected nearest-neighbour step of $O(tn \log n)$ (when $t \lessapprox 20$), and a candidate-limited full-dimensional nearest-neighbour step of $O(hmn)$, in which h is typically a small multiple of k ($2k$ or $3k$). PINN is therefore well-suited for large, high dimensional problems, with the projection overhead making it unsuitable for smaller ones.

PINN can use any projection function as an input (including PCA), and is applicable to any method that requires k-nearest-neighbour distances. We test PINN in our experiments by replacing the standard nearest-neighbour step of LOF with PINN, utilising a random projection result obtained by preprocessing the original data. This combination uses the projected data as an index to the original full-dimensional data, and LOF is calculated using the data in the original unprojected space.

Algorithm 1 RP + PINN + LOF

Input: The n by m matrix X of data in the original space.
Output: The Local Outlier Factor (LOF) score and ranking for each point in X.

RP:
Project X to a n by t matrix Y, $t < m$, using the random projection scheme described in Section II-B.

PINN:
Define h as the parameter for defining the size of the set of candidate nearest-neighbours used, where $h \geq k$.
For each point $p \in X$:

1) Find h-nearest-neighbours of p' in the projected space Y, forming the candidate nearest-neighbour set $N_h(p')$.
2) Map the points in the candidate set $N_h(p')$ back to the original space (X), forming the set $\text{RP}^{-1}(N_h(p'))$.
3) Find the k items of $\text{RP}^{-1}(N_h(p'))$ closest to p. Call this set $\overline{N}_k(p)$

LOF:
For each point $p \in X$, estimate $LOF(p)$ by computing

$$\overline{LOF}(p) = \frac{\frac{1}{k}\sum_{q \in \overline{N}_k(p)} rd(q)}{rd(p)}.$$

V. Experiments

We measured the accuracy and performance of LOF when RP and PCA are used as preprocessing steps (denoted by RP + LOF and PCA + LOF respectively). We also tested the use of PINN as a replacement for the nearest-neighbour step of full-rank LOF, with candidate sets of size $2k$ and $3k$ (denoted by RP + 2knn-PINN + LOF for candidate set size $2k$). We used the data sets described in Table I, sourced from [4] [7] [20], with a subset of size n chosen for Ads and Reuters. The data sets were chosen for their varying types, their relatively high dimensionality, differences in size, and varying intrinsic dimensionality. Most data sets were selected to be small enough for standard unoptimised LOF to be run, so that accuracy could be calculated by comparing the approximation techniques against the ground truth. We also carried out a large-scale performance analysis for the largest data set, Reuters. The real world data sets were preprocessed by removing nominal attributes and attributes containing missing values. We were not able to perform PCA on some data sets due to their high dimensionality.

A. Accuracy experiments

We carried out accuracy experiments to measure the effectiveness of each optimisation technique against the LOF ground truth scores. We varied the dimension of projection (t) for the dimensionality reduction step of each method.

Name	Type	n	m	#1 Eigenvector
Reuters	Text	300000	102600	Unavailable
Amsterdam	Image	24000	27648	Unavailable
Yale face	Image	699	4096	Unavailable
Ads subset	Mixture	1000	1554	Unavailable
Colon	Medical	2000	62	74.4%
Spam	Text	4601	57	92.7%

Table I
DATA SETS USED IN OUR EXPERIMENTS, SHOWING THE DATA TYPE, THE NUMBER OF INSTANCES (n), THE NUMBER OF ATTRIBUTES (m), AND THE PERCENTAGE OF TOTAL VARIANCE CAPTURED BY THE FIRST EIGENVECTOR (#1 EIGENVECTOR). SOME EIGENVECTOR MEASUREMENTS ARE UNAVAILABLE DUE TO THE PROHIBITIVELY HIGH TIME AND SPACE COMPLEXITY OF PCA.

PINN is included with RP chosen as the projection method for scalability, and candidate nearest neighbour set sizes were chosen as $2k$ and $3k$. For all results shown, we fix k to 20. Similar results were found for alternative values of k. We selected $s = 1$ (no sampling) for the sampling parameter in all experiments. The accuracy measure is calculated by initially analysing each data set to determine the set of top true significant outliers P. The accuracy is then measured as $\frac{|P \cap Q|}{|P|}$, where Q is the resultant set of top outliers for the relevant approximation algorithm used. This measure also takes into account the stability of the LOF result: when the LOF values vary significantly from the ground truth, the ordering of the LOF ranking changes, and the accuracy value drops accordingly.

B. Performance experiments

We carried out experiments testing the performance of each method while varying the number of data set items (n). In order to support exact k-nearest-neighbour computation, a kD-tree was used for indexing. When selecting the number of projected dimensions t, it is important to ensure that t is both higher than the intrinsic dimensionality of the data set used, and small enough to ensure good indexing performance. For this reason, we selected $t = 20$, $k = 20$ and $s = 1$ in the experiments. The results shown include the time taken for the preprocessing step of RP, except for the large-scale Reuters experiment, where it is shown separately.

VI. RESULTS AND EVALUATION

These results exhibit variation in accuracy over different data sets due to differing intrinsic dimensionality as well as different levels of outlier distinctness. PCA was not able to be run for the Yale face and Ads data sets due to high dimensionality, but we were able to obtain results for the relatively low dimensional data sets of Colon and Spam.

Figure 8 shows a large scale experiment on the Reuters data set, where an accuracy result could not be obtained as unoptimised standard LOF could not be run. The time complexity is almost linear in n, and far from the quadratic complexity that would be required by standard LOF. This result demonstrates the scalability of our proposed technique in terms of the number of data items.

Figures 9, 10, 11, and 14 show the accuracies for the Yale face, Ads, Colon and Spam data sets, respectively. The results for the Yale face and ADS data sets indicate that for data sets with large feature space and low intrinsic dimensionality, the use of RP allows for highly accurate results even when the dimension of projection is very small. The use of PINN improves the result even further: for the Colon data set, PINN out-performs PCA, while RP is comparable to PCA and slightly better at lower values of t. This shows that PCA is not able to preserve density or LOF values for common distributions of real-world data to the extent that RP can, for the purposes of outlier detection. This may be explained by the fact that many outliers may have strong components in some of the smaller eigenvectors under PCA, and these are discarded when preserving the largest amount of variance for the whole data set including non-outliers. All methods have comparable accuracy on the Spam data set, which was chosen for its suitability for PCA.

Figures 12 and 13 show the performance results while varying data set size (n) for the Yale face and Ads data sets. The results on these sets indicate that for data sets with high representational dimensionality but low intrinsic dimensionality, huge improvements in performance are possible. The performance results of Colon and Spam are omitted, as all methods including unoptimised LOF obtained a highly similar result, as was to be expected due to the low dimensionality and small data set sizes causing the constant factors and factor of m to obscure the effects of scalability with respect to n.

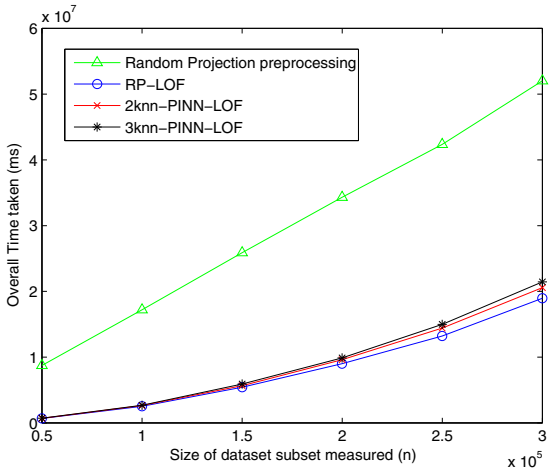

Figure 8. Performance results using the kD-Tree indexed Reuters data set with a dimensionality of 102600. The performance of the common RP preprocessing step is shown separately for clarity.

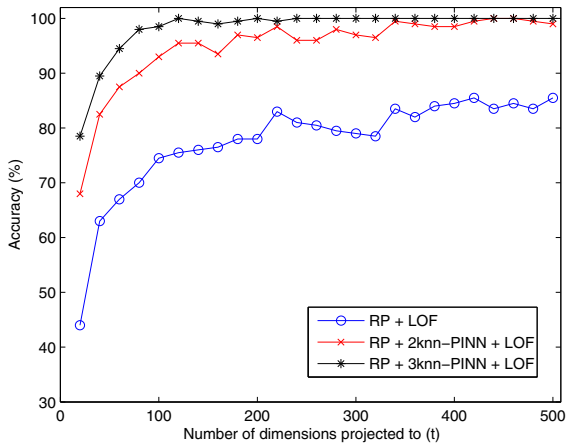

Figure 9. Accuracy results for the top 10 outliers in the Yale face data set.

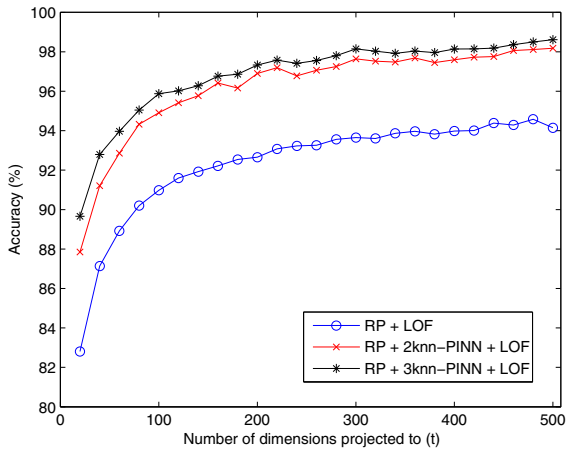

Figure 10. Accuracy results for the top 640 outliers in the Ads subset data set.

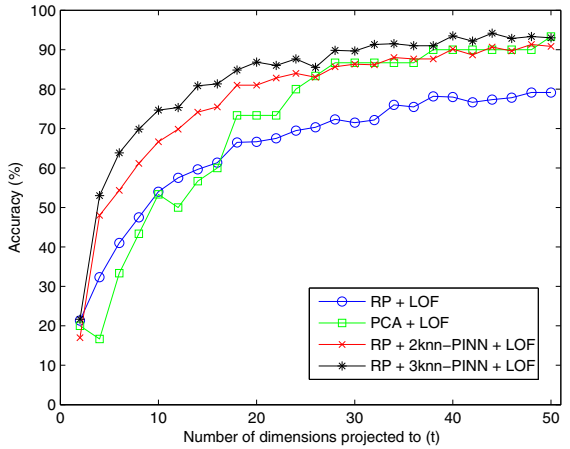

Figure 11. Accuracy results for the top 30 outliers in the Colon data set.

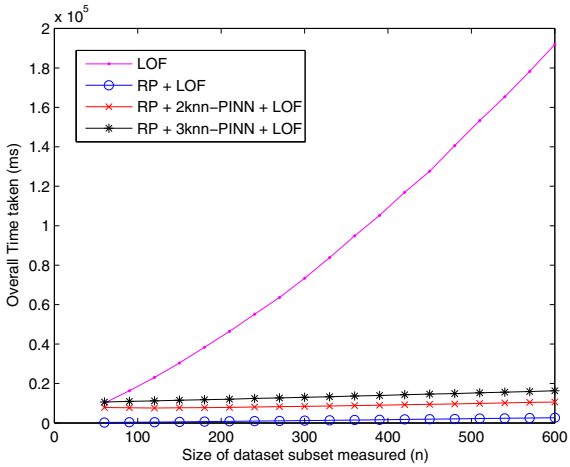

Figure 12. Performance results over n using the kD-Tree indexed Yale face data set.

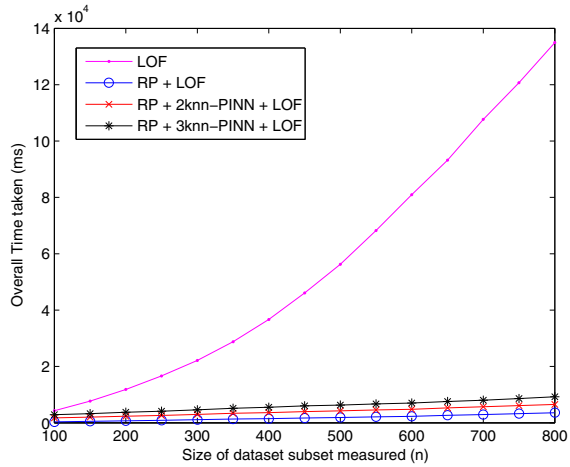

Figure 13. Performance results over n using the kD-Tree indexed Ads subset data set.

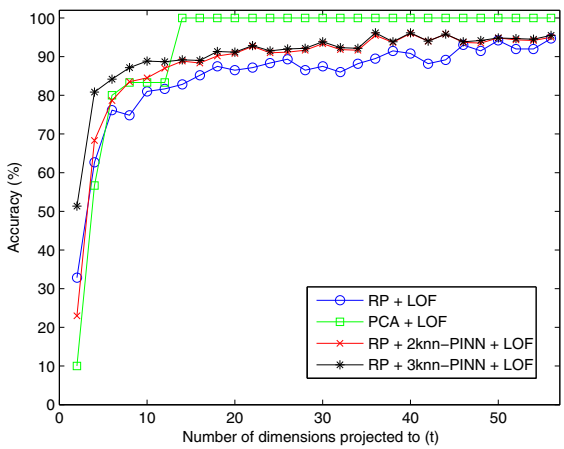

Figure 14. Accuracy results for the top 30 outliers in the Spam data set.

VII. Conclusion

In this paper we have demonstrated the effectiveness of random projection as a stable and robust dimensionality reduction technique in conjunction with LOF. We also presented a new outlier detection strategy, PINN, that allows for highly-accurate and efficient approximation of the LOF, as well as a theoretical analysis of the approximation error. The projections allow the dimensionality to be reduced down to the level where an efficient indexing structure can be applied, resulting in a reduction of the computational complexity of LOF from $O(n^2m)$ to $O(mn \log n)$. Unlike traditional methods based on PCA, the projection-based techniques presented readily scale to the very high-dimensional data sets that are commonly found in areas such as text mining or image processing.

Future work can be carried out to test PINN on other nearest-neighbour-based methods.

Acknowledgment

Timothy de Vries and Sanjay Chawla acknowledge the financial support of the Capital Markets CRC. Michael Houle acknowledges the financial support of the JST ERATO Discrete Structure Manipulation System Project.

References

[1] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," *SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, 2000.

[2] G. H. Golub and C. F. Van Loan, *Matrix computations (3rd ed.)*. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[3] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *The VLDB Journal*, vol. 8, no. 3-4, pp. 237–253, 2000.

[4] J.-M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "The amsterdam library of object images," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112, 2005.

[5] S. D. Bay and M. Schwabacher, "Mining distance-based outliers in near linear time with randomization and a simple pruning rule," in *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2003, pp. 29–38.

[6] A. Hinneburg, C. C. Aggarwal, and D. A. Keim, "What is the nearest neighbor in high dimensional spaces?" in *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 506–515.

[7] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[8] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, "Loci: Fast outlier detection using the local correlation integral," in *ICDE*, 2003.

[9] W. Jin, A. K. H. Tung, and J. Han, "Mining top-n local outliers in large databases," in *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2001, pp. 293–298.

[10] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *STOC '98: Proceedings of the thirtieth annual ACM symposium on Theory of computing*. New York, NY, USA: ACM, 1998, pp. 604–613.

[11] M. E. Houle and J. Sakuma, "Fast approximate similarity search in extremely high-dimensional data sets," in *ICDE*, 2005, pp. 619–630.

[12] 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.

[13] A. Sharma and K. K. Paliwal, "Fast principal component analysis using fixed-point algorithm," *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1151–1155, 2007.

[14] S. Deegalla and H. Bostrom, "Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification," in *ICMLA '06: Proceedings of the 5th International Conference on Machine Learning and Applications*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 245–250.

[15] W. B. Johnson and J. Lindenstrauss, "Extensions of lipschitz mappings into a hilbert space," in *Conference in modern analysis and probability (New Haven, Conn.)*. Amer. Math.Soc., 1982, pp. 189–206.

[16] S. Dasgupta and A. Gupta, "An elementary proof of the johnson-lindenstrauss lemma," International Computer Science Institute, Berkeley, CA, Technical Report TR-99-006, 1999.

[17] D. Achlioptas, "Database-friendly random projections," in *20th ACM Symposium on Principles of Database Systems*. ACM, 2001, pp. 274–281.

[18] D. R. Karger and M. Ruhl, "Finding nearest neighbors in growth-restricted metrics," in *STOC '02: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. New York, NY, USA: ACM, 2002, pp. 741–750.

[19] C. K. Yap, "A geometric consistency theorem for a symbolic perturbation scheme," in *SCG '88: Proceedings of the fourth annual symposium on Computational geometry*. New York, NY, USA: ACM, 1988, pp. 134–142.

[20] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html