Local Anomaly Descriptor: A Robust Unsupervised Algorithm for Anomaly Detection based on Diffusion Space

Hao Huang, Hong Qin
Department of Computer Science,
Stony Brook University
{haohuang,qin}@cs.stonybrook.edu

Shinjae Yoo, Dantong Yu
Computational Science Center,
Brookhaven National Laboratory
{sjyoo,dtyu}@bnl.gov

ABSTRACT

Current popular anomaly detection algorithms are capable of detecting global anomalies but oftentimes fail to distinguish local anomalies from normal instances. This paper aims to improve unsupervised anomaly detection via the exploration of physics-based diffusion space. Building upon the embedding manifold derived from diffusion maps, we devise Local Anomaly Descriptor (LAD) whose originality results from faithfully preserving intrinsic and informative density-relevant neighborhood information. This robust and effective algorithm is designed with a weighted umbrella Laplacian operator to bridge global and local properties. To further enhance the efficacy of our proposed algorithm, we explore the utility of anisotropic Gaussian kernel (AGK) which can offer better manifold-aware affinity information. Comprehensive experiments on both synthetic and UCI real datasets verify that our LAD outperforms existing anomaly detection algorithms.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; I.5.1 [**Pattern Recognition**]: Models—*Unsupervised Anomaly Detection*

Keywords

Anomaly detection, diffusion space, LAD

1. INTRODUCTION

Anomaly detection or outlier detection is of great significance to many applications [40] [28]. Its primary goal is similar to that of a classification problem except that it further distinguishes normal instances from a small portion of new or abnormal instances (anomalies) [5] [19] [20]. In many applications, anomalies are sparse and quite diverse, learning with the known anomalies [9] [38] [3] may not be necessarily useful in detecting the unknown ones in unseen data [34]. On the other hand, labeling known datasets can be extremely time-consuming for real-life applications and sometimes even unpractical to detect new types of rare events. Therefore, the key challenge of anomaly detection still lies in its robust ability to quantitatively and unsupervisedly characterize the intrinsic and informative density distribution around every instance.

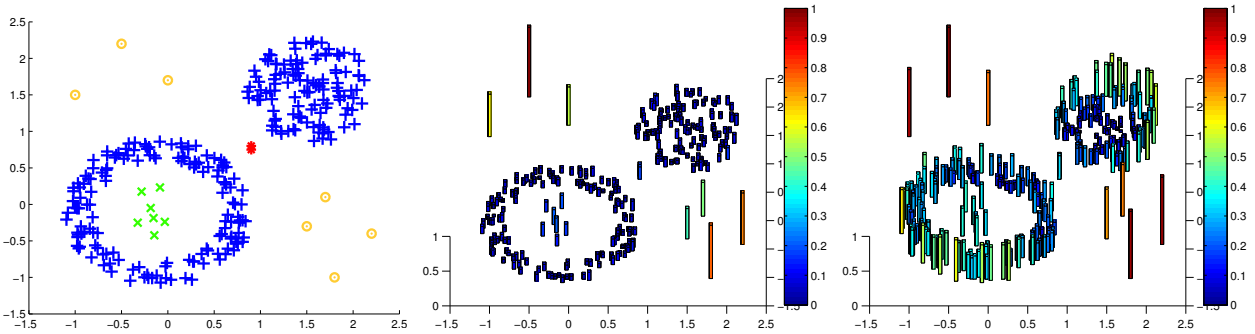
Our proposed unsupervised method in this paper, called Local Anomaly Descriptor (LAD), computes a measurement of anomalousness based on physics-based diffusion theory, which is more **informative** and **intrinsic** compared with the existing algorithms ([4] [25] [19] [36] [1] etc.). First of all, our algorithm projects origin instances onto a diffusion space. In the diffusion space, distance between anomalies and normal instances will be magnified, which makes the density of anomalies even less and therefore more salient compared with those in the original space. However, the perfect diffusion maps are usually unreachable. Oftentimes anomalies are hard to be totally distinguished from the normal instances that are not too far away. To better set anomalies apart from the nearby normal instances, we innovatively apply a weighted Laplace umbrella operator on the projected diffusion space, called Local Anomaly Descriptor (LAD). With the novel LAD which bridges the gap between global and local properties, we can not only obtain intrinsic local density information, but also take the quantity of similar instances into consideration. Therefore the representation is more reliable than original attribute distribution, and more informative since it covers a sufficiently large neighborhood for each instance. Furthermore, LAD provides reasonably stable performance as the scaling parameters vary (the neighborhood size k and Gaussian scaling parameter σ).

In this paper, we employ heat kernel to build the diffusion maps, which offers a statistical description on random walks. However, the pivotal techniques of our algorithm are fundamentally different from the current existing data mining research based on heat kernel theory [15] or other similar diffusion methods [6] [30] [29]. Our proposed LAD has distinctive uniqueness on balancing local and global properties, and its advantage on both performance and robustness on real world datasets for anomaly detection.

1.1 Related Work

According to the most classical definition by Hawkins [12], an anomaly is "an observation which deviates so much from the other observations as to arouse suspicion that it was generated by a different mechanism". However, it is far from trivial to define the quantitative scope of "other observations". As Figure 1(a) illustrates, global anomalies (in yellow) are those with global minimum neighborhood density, and

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.



(a) Synthetic Dataset Example (b) LOF Score ($k = 10$) (c) IForest Score ($nt = 100$, $\rho = 4000$)

Figure 1: 1(a) Synthetic dataset with normal instances (blue), global anomalies (yellow), and local anomalies (red and green); 1(b) LOF score with $k = 10$; 1(c) IForest score. The anomalousness are visualized as height bar over all the instances. We can see that both LOF and IForest fail to totally distinguish local anomalies from normal instances.

distinct with respect to (almost) the entire dataset. While local anomalies (in red and green) are those with local minimum neighborhood density, and distinct with respect to instances in their local neighborhood. Profoundly speaking, local anomalies can be thought of as a generalization of global anomalies, as global anomalies will typically also be local anomalies, but not vice versa [7].

In implementation, kNN-based algorithms such as LOF [4] and LDOF [39] define anomaly if its distance (usually in Euclidean space) to its k-th nearest neighbor (kNN) is large relative to the distances of its neighbors to their own k-th nearest-neighbors. Recent research [7] extended LOF to high-dimensional dataset by using random projection. Two major drawbacks of these approaches are: (1) they tend to miss local anomalies (Figure 1(b)) since it is not peculiar that kNN distances of local anomalies are similar to their normal instance neighbors'; (2) it is of extreme importance to determine the value of k, because k can not be too small to avoid statistical error. In other words, we need to ensure that for each instance, especially those forming micro-cluster of anomalies, it does intend to use a neighborhood size which is large enough to include more normal instances than anomalies. However, too large k will lead to miss some genuine anomalies. In Section 6.2 we will show that LOF is statistically vulnerable by analyzing the sensitivity of k.

Instead of detecting anomalies based on average neighborhood distance, recent approaches such as IForest [19] [20] and Mass [36] are to separate the anomalies from normal instances with their unique attribute distribution. A representative anomaly definition [19] in the aforementioned papers states that anomalies should have "attribute-values that are very different from those of normal instances", and at the same time are "minority consisting of fewer instances". Therefore these approaches have the capacity to handle anomalies with different attribute distribution compared with normal instances [18]. Nonetheless, these approaches may fail to detect some local anomalies when their attributes have similar distribution with some normal instances'. From Figure 1(c) we can see that, even though IForest did a good job on global anomaly detection, it fails to distinguish local anomalies (green and red instances in Figure 1(a)) from the "boundary" instances in the normal instance clusters (blue instances in Figure 1(a)). The reason is that, these anomaly detectors partition instances mainly based on observable attributes, or more precisely, the attribute distribution in original data space. Therefore it will fail miserably when the anomaly distribution becomes far less discriminative by sharing similar attribute range/distribution pattern with parts of the normal instances. In Figure 2 we can see that some anomalies have overlapped distribution on the first four eigenvectors with normal instances in Ionosphere dataset (a popularly used dataset for anomaly detection [19] [13] [23]). Such overlapping also appears at nonclassical multidimensional scaling as well. So, this problem indeed exists in some real world datasets.

A few techniques [1] tried to find an approximation of the data using a combination of attributes that capture the bulk of the variability in the data, and then detect anomalies on the projected space. This kind of approaches adopted by spectral techniques is to determine manifold subspaces in which the anomalous instances can be easily identified [5]. However, the existing algorithms are based on techniques such as isometric feature mapping (ISM) and locally linear embeddings (LLE) [1] which are highly sensitive to density-varying data distribution [16] [37].

1.2 Motivation

Motivated by the aforementioned problems, we re-define anomaly as follows:

Definition: Anomalies are those instances with (1) **local minimum neighborhood density** and (2) **small quantity** compared with normal instances.

To capture anomalies under such definition, we consider the heat equation in diffusion theory, which has intrinsic relationship with manifold reconstruction and built-in robustness of scaling parameters [15]. The reason why we resort to manifold space is that normal instances usually lie on some low dimensional manifold structures in high density regions, moreover, the anomalies deviate from the normal instances which makes them even more discriminative. Diffusion distance is based on Markov matrix which is a stochastic matrix representing random walks on graph [22], it can consider up to t steps out of all the possible paths bridging any two instances, which makes it more robust than Euclidean and geodesic distance [6] [16] [37]. However, different from the

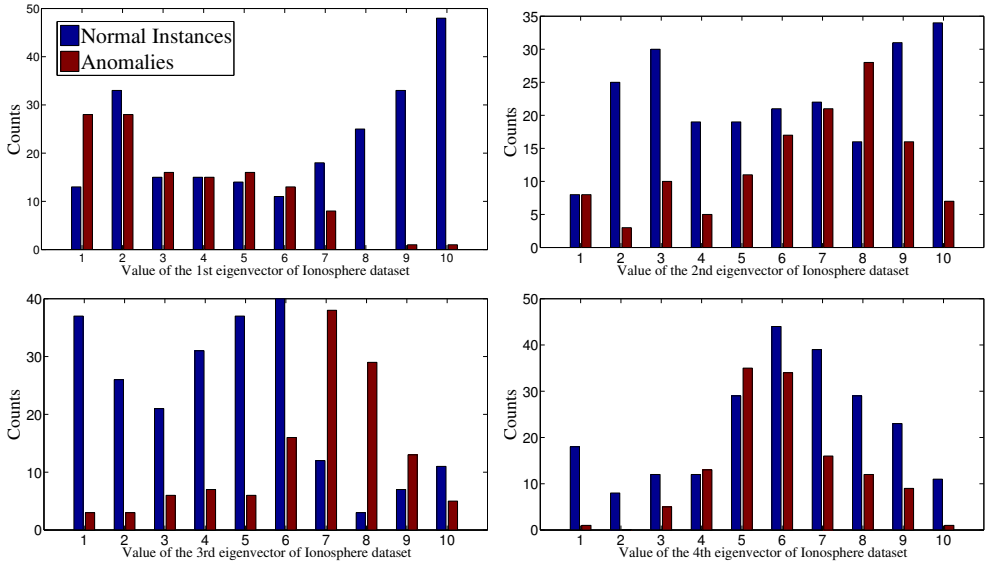


Figure 2: Histogram of anomaly (red) and normal instance (blue) on the first four eigenvectors (*) of Ionosphere dataset (a popular benchmark dataset for anomaly detection [19] [13] [23]). Some anomalies have overlapped distribution with parts of normal instances and therefore it is nontrivial to separate them only by simple attribute distribution. *Since the dataset is high-dimensional, dimension reduction is required to provide a concise illustration. Although eigenvectors do not necessarily show full distribution of the original data, it tends to show certain patterns of partial dimensions in the original space.

current diffusion-related research with the goal of clustering [30] [29] [15], diffusion mapping for anomaly detection has crucial requirement for connecting similar instances and at the same time avoiding excessive-connection. To our best knowledge, this diffusion-based utility on anomaly detection has never been explored yet.

In this paper, we propose Local Anomaly Descriptor (LAD) which offers a natural mechanism to express intrinsic neighborhood density information through heat diffusion process. To offer a solution to the inherent problem from the perspective of heat diffusion, LAD provides a meaningful trade-off between local and global manifold awareness by applying a weighted Laplacian umbrella operator. Experiments show that LAD is immediately useful for a wide variety of anomaly detection applications.

1.3 Contribution

This paper articulates a novel unsupervised anomaly detection algorithm which is intrinsic, informative, and robust to scaling parameters with the following contributions:

(1) We quantitatively characterize local density information based on heat diffusion theory (Section 2) and anisotropic Gaussian kernel (Section 3). This method is both intrinsic and informative in that, it has a more locally adaptive scope of manifold-aware neighborhood and therefore can very well satisfy the first property of the above definition in Section 1.2 more insightfully and intrinsically compared with the existing algorithms.

(2) In order to take the amount of similar instances into account (the second property of the above definition in Section 1.2) which can better separate local anomalies from normal instances, we explore the use of weighted umbrella Laplacian operator (Section 4) which can bridge the gap between local and global information.

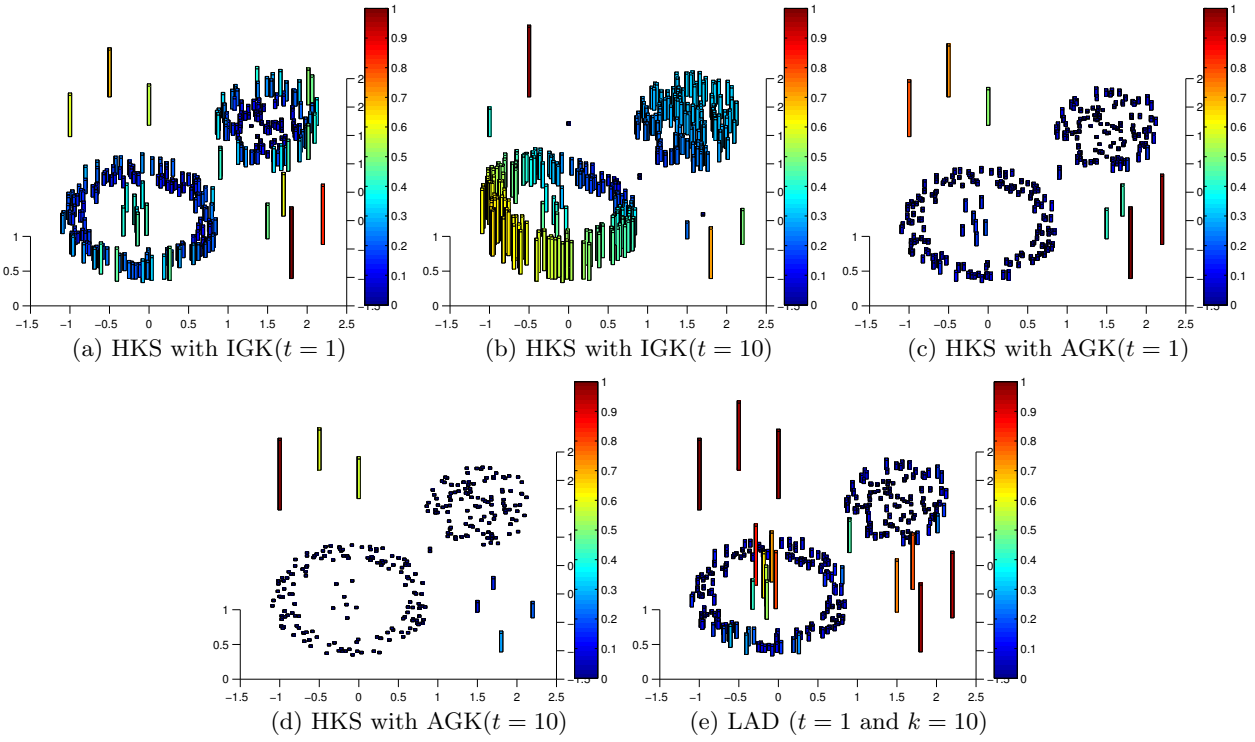
(3) We systematically evaluate the proposed algorithm with several closely-related baseline algorithms on a number of benchmark datasets (Section 6). Our algorithm shows not only better average performance but also more stable results than the other popular algorithms. Moreover, our algorithm affords robustness for parameter selections (neighborhood size and Gaussian scaling parameters).

2. HEAT KERNEL SIGNATURE BASED ON DIFFUSION SPACE

Our proposed work is strongly inspired by heat diffusion theory [14] in that it can provide information intimately related to local density. Heat theory can be interpreted as the transition density function of Brownian motion [33], which is the most fundamental continuous time Markov process. Laplace operator is intimately related to heat diffusion, connecting geometry of a manifold with the properties of the heat flow. Using the discrete Laplace operator, the heat equation can be simplified, and generalized to matrix operation over spaces with an arbitrary number of dimensions. In practice the heat equation is often associated with random walk graph Laplacian [6], L_{rw}. Random walk is a stochastic process which randomly jumps from vertex to vertex. Heat equation therefore can be defined by

$$\frac{\partial H_t}{\partial t} = -L_{rw} H_t, \tag{1}$$

where $H_t = e^{-tL_{rw}}$ is the heat kernel on Riemannian manifold $\mathcal{M}$ and t is the time scaling parameter [11]. For $L_{rw} = \psi' \lambda \psi$ (ψ and λ are the eigenvectors and eigenvalues of L_{rw}



(a) HKS with IGK($t = 1$) (b) HKS with IGK($t = 10$) (c) HKS with AGK($t = 1$)

(d) HKS with AGK($t = 10$) (e) LAD ($t = 1$ and $k = 10$)

Figure 3: HKS score with IGK (Isotropic Gaussian Kernel) and AGK and LAD score of the synthetic dataset in Figure 1(a). We can see that LAD is the most sensitive to both global and local anomalies.

), the heat kernel can be re-formulated as follows:

$$H_t(i,j) = \sum_{p=1}^{N} [e^{-\lambda_p t} \psi_p(i) \psi_p(j)], \tag{2}$$

where $H_t(i,j)$ represents the amount of heat being transferred from i to j in time t given a unit heat source at i. The scaling parameter t in heat kernel is used to control the transitive connectivity: small t makes the loosely-connected graph into slightly stronger connection within t connections, while large t makes the graph tend to be more strongly-connected.

In 2009, Sun et.al [33] proposed a concise form given by the heat kernel from one instance to itself

$$HKS_t(i) = H_t(i,i) = \sum_{p=1}^{N} [e^{-\lambda_p t} \psi_p^2(i)], \tag{3}$$

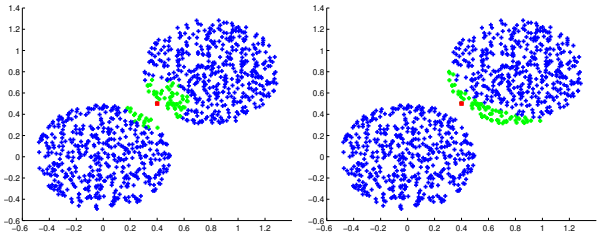
which is called Heat Kernel Signature (HKS). The physical meaning of HKS is the amount of heat each instance keeps within itself in time t. **The property of heat diffusion process states that heat tends to diffuse slower at instances with more sparse neighborhood and faster at instances with denser neighborhood. Therefore HKS can intuitively depict the local density of each instance (the first property in our anomaly definition).** Besides, HKS also has the following properties which make it a very lucrative candidate for local density measurement: 1) it is intrinsic to the local manifold structure; 2) it is informative since it contains density information about the whole neighborhood in t scale; and 3) the probabilistic interpretation of heat diffusion can well support the stableness of HKS against small perturbation in the neighborhood.

However, Heat equation is assumed to build on the underlying manifold. But in most applications, the underlying manifold is unknown. In practice, HKS is usually built on isotropic Gaussian kernel (IGK) on observed space. Although graph Laplacian normalizations [6] based on simple IGK on observed space can recover manifold structure to certain extent, non-uniformly sampled instances tend to show unpreserved density distribution on the reconstructed manifold. HKS on IGK will fail to reveal local density faithfully in such reconstructions. Figure 3(a) and 3(b) shows the performance of HKS on anomaly detection with $t = 1$ and $t = 10$ based on simple IGK and random walk graph Laplacian normalization. When $t = 10$ (Figure 3(b)) the heat is extremely easy to dissipate, which blends both local and a few global anomalies into normal instances. Meanwhile many marginal instances of the two normal instance clusters stand out due to the fact that the HKS on IGK fails to show manifold-aware properties. When $t = 1$ (Figure 3(a)), although the short period of heat dissipation has salient effect on global anomalies, HKS on IGK still fails to distinguish local anomalies from normal instances on the boundary area of normal clusters. Therefore an alternative way is indispensable to build better manifold-aware affinity matrix. One of the most preferable candidates is anisotropic Gaussian kernel (AGK) [31] [32].

3. ANISOTROPIC GAUSSIAN KERNEL

In this section we integrate anisotropic Gaussian kernel (AGK) [31] into HKS to achieve better manifold reconstruc-

tion. In Figure 4 we can see the 70 nearest neighbors of red instance when using IGK (Figure 4(a)) and AGK (Figure 4(b)), which shows that the intra-manifold distances are much shorter than the inter-manifold by using AGK. To further support this idea, in Figure 3(c) and 3(d) we show that anomaly detection can directly benefit from the use of AGK. In Figure 3(d) with $t = 10$, the global anomalies are highlighted even though the local anomalies are latent (compared with Figure 3(b)). This is because if the manifold is well reconstructed, global anomalies should be separated far away from normal instances even with large t scale. Furthermore, in the small scope of $t = 1$ (Figure 3(c)), both local and global anomalies can be detected, which illustrates that with the support from AGK, HKS is capable of revealing the density information of the intrinsic manifold structure.



(a) (b)

Figure 4: 70 nearest neighbors (in green) of red instance on IGK (a) and AGK (b), which shows that AGK has better manifold-aware property than IGK.

In the rest of this section we briefly introduce AGK on the observed space Y that approximates the isotropic Gaussian kernel on the underlying manifold X. The idea is to approximate the Euclidean distance between instances $x^{(j)}$ in the manifold space X using covariance matrix $C = JJ^T$ where J is the Jacobian matrix [31] and the instances $y_j = f(x_j)$ in the observable space Y. Let x, ϵ be two instances in the manifold space X and $y = f(x)$, $\eta = f(\epsilon)$ be their mapping to the observable space Y. Let $g : Y \to X$ be the inverse mapping of $f : X \to Y$, that is, $g(f(x)) = x$ and $f(g(y)) = y$, $\forall x \in X$, $\forall y \in Y$. Expanding the functions $x = g(y)$ in a Taylor series at the instance y gives

$$\epsilon_i = x_i + \Sigma_j g_{i(j)}(y)(\eta_j - y_j) + \frac{1}{2}\Sigma_{kl} g_{i(kl)}(y)(\eta_k - y_k)(\eta_l - y_l) + O(\| \eta - y \|^3). \quad (4)$$

where $g_{i(j)} = \frac{\partial g_i}{\partial y_j}$. Therefore,

$$\| \epsilon - x \|^2 = \Sigma_{ijk} g_{i(j)}(y) g_{i(k)}(y)(\eta_j - y_j)(\eta_k - y_k) + \frac{1}{2}\Sigma_{ijkl} g_{i(j)}(y) g_{i(kl)}(y)(\eta_j - y_j)(\eta_k - y_k)(\eta_l - y_l) + O(\| \eta - y \|^4). \quad (5)$$

A similar expansion can be built at instance η and the average of these two equations can be produced as

$$\| \epsilon - x \|^2 = \frac{1}{2}(\eta - y)^T [(JJ^T)^{-1}(y) + (JJ^T)^{-1}(\eta)](\eta - y) + O(\| \eta - y \|^4), \quad (6)$$

given that the Jacobian of the inverse g is the inverse of the Jacobian J (a detailed description of calculation can be referred to [31]). So we can construct the anisotropic Gaussian kernel

$$AGK(y_i, y_j) = e^{\{-\frac{\| J^{-1}(y_i)(y_i - y_j) \|^2 + \| J^{-1}(y_j)(y_j - y_i) \|^2}{\sigma^2}\}}, \quad (7)$$

where $i, j = 1, ..., N$.

AGK has the desired attributes that it is separable, and its first (nontrivial) eigenfunctions are monotonic functions of the independent parameters [32]. It also has been proved that the eigenvectors of AGK reveal the independent components [31]. HKS, built on such approximation, can better capture the manifold structure of data as shown in Figure 3(c) and 3(d), which is difficult or even impossible to achieve by using IGK or other similar techniques.

4. LOCAL ANOMALY DESCRIPTOR

Although HKS on AGK has the capability to offer desirable local density information, it is of importance to select the right time scaling parameter t, which provides a trade-off between the effects of local and global information. However, it is hard to get the "best of both worlds" with single setting for this parameter. Even with better manifold reconstruction, if t is large the heat is still easy to dissipate regardless of normal instances or local anomalies, although not for global anomalies, which is shown in Figure 3(d). This is because with large t scale, the distance between local anomalies and the normal instances around them would still be close. So local anomalies cannot retain their heat. On the other hand, if t is small, the heat diffusion runs for only a short period of time, and the resulting anomalousness are usefully local, but almost carry the same value for instances with similar density inside a very restrained neighborhood, which is the major reason why it sometimes assimilates local anomalies into some normal instances. In Figure 3(c) we can see HKS assigns similar scores to the local anomalies and some of the boundary normal instances. Intuitively speaking, HKS on AGK still fails to take the amount of similar instances into account with off-the-sweet-spot t setting.

In order to handle the above problem, we propose to use the umbrella operator [35] [8] to consider the quantity of similar instances in neighborhood by bridging the gap between global and local properties. The main motivation for using this operator is to compute the average difference between a point (x_i) and its k neighbors ($nb(i,k)$).

$$\Delta x_i = \sum_{x_j \in nb(x_i,k)} W_{i,j}(x_j - x_i), \quad (8)$$

where $W_{i,j}$ is the weight between x_i and x_j. If we use $AGK(i,j)$ for $W_{i,j}$, then we may define the Local Anomaly Descriptor (LAD) for a point i as follows:

$$LAD(i) = HKS_t(i) - \frac{1}{k}\Sigma_{j \in nb(i,k)} HKS_t(j) \cdot AGK(i,j). \quad (9)$$

The geometric meaning of LAD is illustrated in Figure 5 where we measure the difference between a single $HKS_t(i)$ and its neighborhood's average $HKS_t(j)$ value. Note that the heat kernel signature value is always positive and it means the degree of global anomaly level and thus $LAD(i)$ indicate the level of both global and local properties.

If an instance is globally anomalous, its HKS would be already high enough to discriminate itself to the other instances. While it is locally anomalous, its HKS is likely to be similar to some normal instances' with similarly sparse

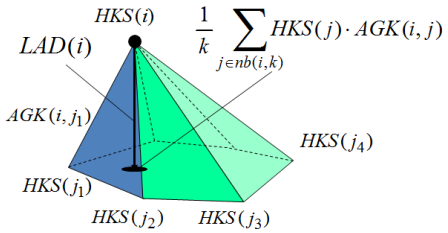


Figure 5: Illustration of Local Anomaly Descriptor which calculates weighted average of neighbor differences. It is one of the ways taking into consideration the neighborhood structures [35].

neighborhood. However, the amount of similar instances, if it is taken into account, can serve to recognize the local anomalies from normal instances. Since local anomaly only has a small amount of neighbors with close HKS, but normal instances, on the other hand, have more such neighbors. **LAD has a very lucrative property in considering the amount of similar instances (the second property in our anomaly definition): with similar intra-cluster density, AGK tends to assign larger affinity value to two instances inside a larger manifold/cluster, and less value to those inside a smaller manifold/cluster, and much less value to those not inside the same manifold/cluster [31].** So even though k is not large enough to include the whole appropriate neighborhood, LAD can still capture the information related to the amount of similar instances.

The benefits of LAD in comparison with HKS can be seen in Figure 3(e), which shows the LAD score of synthetic dataset (Figure 1(a)). Note that HKS with AGK possesses good global properties for long enough time ($t = 10$ in Figure 3(d)) and good local properties for small time ($t = 1$ in Figure 3(c)), but not both. Nevertheless, Figure 3(e) shows that our proposed LAD has a penetrating awareness on both global and local anomalies primarily because of the power of our proposed umbrella operator.

We now justify the LAD utility by briefly documenting its theoretic connections with a few existing methods, which also lays a solid foundation for LAD's attractive properties in practical use.

Biharmonic Operator. HKS itself is directly derived from the Laplace operator and its eigen-decomposition, so that HKS is intrinsically a second-order property relevant to the Laplace's equation. The aforementioned derivation of LAD can be intuitively related to the biharmonic process, because the Laplace operator is essentially applied twice (to compute both HKS and its umbrella operator of HKS). It provides a good balance in the sense that it decays slowly in small cluster around the source instance and fast enough to be structurally inherent in dense areas. This specific "balancing" is intimately derived from the biharmonic equation with properties such as local support, global informative, and shape-aware [17].

Signal Processing. LAD also has strong connection to signal processing . In lowpass filtering, the divergence of a sample from its average neighborhood is the easiest way to pinpoint those inconsistent instances if the desired signal has significant high frequency content. As in traditional signal processing [35], it is possible for LAD to quantify the frequency response by computing an adjoining sum of the Laplacian operator in its immediate vicinity. As a result, this enables LAD to distinguish between normal instances and inconsistent instances (anomalies) with greater precision.

kNN-based Approaches. kNN-based methods [4] [7] [39] approach local density for each instance using its neighborhood information. Like LAD, they require scaling parameters to capture a reasonably large neighborhood, and the density information is based on this prescribed local region. However, kNN-based methods has strictly local context in that they simply fix the neighborhood size with k. In contrast, LAD employs locally adaptive neighborhood size directly benefited from the physics-inspired properties of heat diffusion. Moreover, Euclidean distance in kNN-based methods is a pair-wise local quantity, while heat kernel used in LAD considers all the possible paths between two instances within time t, therefore LAD is more stable than kNN-based methods.

Attribute-based Approaches. Attribute-based methods [19] [20] [36] try to compute local density by adding up a sequence of values from an attribute-based function [36], equivalent to a kernel density function such as heat kernel. The global instance distribution is based on each attribute and how deviated each instance is from the other instances in that specific attribute, which indeed is more informative than kNN-based approaches. However, the strong emphasis on attribute distribution along its dimension is also a "double-sided sword": on the one hand it is much faster without any distance calculation, on the other hand, such distribution based on attributes still fails to consider local anomalies.

Diffusion-based Clustering. Some recent research [30] [29] [15] proposed the unified probabilistic clustering approach based on diffusion map. By integrating all time scales of kernel function into one single term, this kind of techniques completely removes the time scaling parameter of diffusion dissipation, therefore it has the built-in robustness to data perturbation and scaling parameter modification [15]. However, as a side-effect, this process of "integration" assimilates local anomalous instances into normal instance clusters since the excessive-diffusion tends to connect everything together. LAD, in sharp contrast, is built upon kernel function with small time scale and weighted umbrella operator instead of integrating all scales together. Therefore it avoids the above-mentioned excessive-connection problem.

5. ALGORITHMIC FRAMEWORK

After investigating some attractive properties of LAD, it now sets a stage for us to introduce a novel anomaly detection algorithm, which is sensitive to both global and local anomalies. Let X be a matrix of size $n \times m$, where n is the number of instances and m the number of dimensions, our framework is detailed in Algorithm 1. This algorithm undergoes a kind of data warping process by using AGK (Step 1) and Laplacian Random Walk normalization (Step 2). Then we perform the eigen-decomposition (Step 3) and construct HKS for each instance (Step 4). Equation 9 is used as the last step to compute Local Anomaly Descriptor for the final measurement of anomalousness.

In Step 4, we adopt the notation in [10] by normalizing the time scale $t \leftarrow t/(2\lambda_1)$ to achieve scale invariance. Henceforth, with little abuse of notation, heat diffusion time in

Algorithm 1: LocalAnomalyDescriptor(X, σ, t, k)

Input: Input data $X \in R^{n \times m}$; σ the Gaussian scaling parameter; t the time scaling parameter; k the neighborhood size

Output: LAD score for each instance

1 Construct anisotropic Gaussian kernel W using Equation 7 and σ;
2 Construct Laplacian random walk normalization on W ;
3 Compute generalized eigenvectors $\psi(i)$ and corresponding eigenvalues λ_i, $i = 1, 2, ..., n.$;
4 Construct Heat Kernel Signature in time scale t using Equation 3 ;
5 Compute Local Anomaly Descriptor using Equation 9 with Heat Kernel Signature and anisotropic Gaussian kernel in the k neighborhood for each instance ;

our paper will actually denote $t/(2\lambda_i)$, where λ_i is the first non-trivial eigenvalue.

Regarding computational complexity, eigen-decomposition (Step 3) is the most time-consuming step, which will dominate our computation. There are many iterative methods to conduct eigenvalue decomposition, but in general finding the eigenvalues reduces to matrix multiplication by computing a symbolic determinant, which gives a running time of $O(n^3 + n^2 log^2 n)$ [24]. An alternative way of estimating the heat kernel $K_t = e^{-1L_{rw}} D^{-1}$ is to use a partial sum of infinite series with

$$e^{-tL_{rw}} = \Sigma_{i=0}^{\infty} \frac{(-tL_{rw})^i}{i!}. \quad (10)$$

This method would be especially attractive for small values of t, since only a few terms would be needed to obtain an accurate estimation of $e^{-tL_{rw}}$ [2].

6. EXPERIMENTAL ANALYSIS

6.1 Experimental Setup

Dataset. To demonstrate the performance of our proposed method, we evaluate our algorithm on nine UCI benchmark datasets including three medical datasets (WDBC, Pima, and Arrhythmia), three biological datasets (Ecoli, Yeast, and Abalone), and three physics datasets (Glass, Ionosphere, and Magic), whose statistics are summarized in Table 1. All these data have been popularly used in anomaly detection research (related references for each dataset are listed in Table 1). Anomalies in some of the datasets (WDBC, Pima, Arrhythmia, etc.), although carrying a large number of instances, have scattered and sparse distribution (Figure 6). Therefore the anomalies in these datasets should be treated as a combination of many small anomalous clusters instead of one or a few normal clusters with high density [7] [23], which has nothing inconsistent with our definition about anomaly in Section 1.2. Such diverse combination of data is intended for our comprehensive studies. In the data preprocessing step, all nominal (including binary) attributes or attributes with missing value are removed.

Baselines. We choose six states of the art competitors in three categories to show the outstanding performance of our proposed LAD. For kNN-based algorithms, we choose Local Outlier Detection (LOF) [4] and Local Correlation Integral (LOCI) [25]. Specially, LOCI provides an automatic,

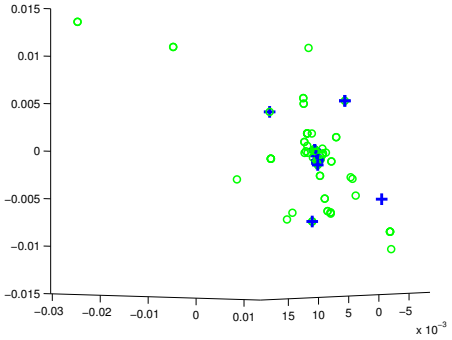


Figure 6: Anomalous instances in green (37.3%) are more scattered and sparse than normal instances in blue (62.7%) in WDBC dataset (shown with the first three eigenvectors). Therefore these anomalies, although have a large amount of instances, should be treated as many small abnormal clusters.

data-dictated cut-off to determine whether an instance is an anomaly based on probabilistic reasoning. For attribute-based methods, we include IForest [19] and Mass [36]. For manifold-based methods, we refer readers to two different manifold-based techniques in [1] including locally linear embeddings (LLE), and isometric feature mapping (ISM), followed by LOF to obtain anomalousness measurement.

Evaluation Metrics. Since we have the ground truth of labels for each data, we compare our anomaly detection results with labels. Due to space limitation, AUC (Area under Receiver Operating Characteristics Curve) is used as the only listed evaluation metric in this paper because it is commonly used to evaluate anomaly detectors and it is cut-off independent. Detailed definition of AUC can be referred to [21]. In our paper we also show that our LAD has the most robust and stable performance for all the datasets by using macro paired t-tests [41] against each competitor respectively. Note that a score of macro paired t-tests (p-value) should be no more than 0.05 to be considered statistically significant.

Parameters. Our proposed algorithm has three scaling parameters, namely Gaussian kernel scaling parameter σ, time scaling parameter t, and the size of neighborhood k. We set $t = 1$ which makes our proposed LAD capable of depicting local minimum density. As the default setting for most of the other algorithms, we fix $\sigma = 1$ and $k = 10$ for the experiments in Table 2 and 3. But the LAD robustness to the change of these two parameters will be shown in Figure 7. For LOF we try $k = 10$, 25, 50 for all the datasets. Since in practice, single setting of k for LOF may introduce statistical errors [25]. As for LOCI, we set $k = 10$ and $k = 50$ due to its instinctive stability on k which comes from a multi-granularity deviation factor [25]. Radius coefficient is set as $\alpha = 0.5$ in LOCI, which is the same to their paper [25]. As for IForest, even though in their paper [19] Liu et al. claimed that a small sub-sampling size ρ provides high AUC and a further increase of ρ is not necessary, in practice when ρ increases, anomalies in-between data groups will become more detectable by IForest [18]. To conduct safe and fair comparison, we set $\rho = 4000$ and the number of trees $nt = 100$ since they are the recommended settings in the authors' technical report [18]. Similarly, in Mass we

Table 1: Statistics of our evaluation datasets.

	Dataset	# Instance	# Attribute	% Anomalies(classes)	References
1	WDBC	569	29	37.3% (malignant)	[39] [23]
2	Pima	768	8	34.9% (positives)	[19] [23]
3	Arrhythmia	452	279	45.0% (abnormal)	[7] [19] [23]
4	Ecoli	336	7	2.7% (omL,imL and imS)	[13] [23]
5	Yeast	1484	8	3.7% (vac, pox and erl)	[23] [27]
6	Abalone	4177	7	8.0% ($age < 5$ or > 15)	[23] [26]
7	Glass	214	9	4.2% (tableware)	[13] [23]
8	Ionosphere	351	34	35.9% (bad)	[19] [13] [23]
9	Magic	19020	10	35.2% (hadron)	[23]

Table 2: Comparison of LAD and other four popular methods (with different k for LOF and LOCI) on nine datasets using AUC metrics. The bold-faced numbers indicate the best method on a particular dataset; the numbers in parentheses indicate the ranks of our LAD. Average is the average of performance across all the datasets respectively. A $*$ indicates a p-value of 5% or lower and $**$ indicates a p-value of 1% or lower in the statistical significance test for performance comparison w.r.t. LAD.

Dataset	LOF(k=10)	LOF(k=25)	LOF(k=50)	LOCI(k=10)	LOCI(k=50)	Mass	IForest	LAD
WDBC	0.5874	0.6192	0.7784	0.7573	0.8119	0.7974	0.7760	**0.8864** (1)
Pima	0.4847	0.5270	0.6003	0.5946	0.6089	0.5812	0.6630	**0.6936** (1)
Arrhythmia	0.7419	0.7547	0.7482	0.7482	0.7483	0.6363	0.7456	**0.7558** (1)
Ecoli	0.8260	0.8614	0.8454	0.8641	0.8549	0.7699	**0.8754**	0.8692 (2)
Yeast	0.4159	**0.6253**	0.5986	0.6076	0.6035	0.5712	0.6159	0.6183 (2)
Abalone	0.5724	0.6056	0.6525	0.6932	0.7058	0.6923	**0.7466**	0.7398 (2)
Glass	0.7474	0.7008	0.7528	0.7480	0.7593	**0.8922**	0.6933	0.8612 (2)
Ionosphere	0.9064	0.8709	0.7982	0.8512	0.7923	0.8269	0.8467	**0.9240** (1)
Magic	0.6420	0.6804	0.6970	0.5672	0.6825	0.6984	0.7506	**0.7516** (1)
Average	0.6582**	0.6939**	0.7190**	0.7146**	0.7297**	0.7184**	0.7459*	**0.7889** (1)

set the number of mass estimation $ne = 100$ and the sub-sampling size as $\#instance$ of dataset. On the other hand, IForest and Mass are based on random sub-sampling which makes their performance very unstable. In an attempt to get more stable results, for each dataset we run 30 times for both IForest and Mass and use the average AUC in the final comparison. For LLE and ISM, we conduct experiments on size of neighborhood $k = 10$, 50 and 100 with the best number of dimensions d in $[2, 30]$ respectively, in order to compare their performance and robustness in k.

6.2 Algorithm Performance Comparison

In this section we evaluate our proposed LAD and the other six anomaly detection algorithms. Table 2 documents the anomaly detection comparison result (in AUC and p-value) of LAD and other four popular algorithms: LOF, LOCI, Mass, and IForest. While the manifold-based methods comparison including LLE and ISM (all followed by LOF), and our proposed LAD are also listed in Table 3.

In Table 2 LAD shows the best average performance (0.7889) across all the datasets. For each dataset LAD has the best or very close to the best performance. Specifically, LAD is the top-ranked one for all the three medical datasets and has almost unbeatable performance for the three physics datasets. Although LAD ranks the second among all the methods on the three biological datasets. The AUC score are actually very close to the best one (no more than 0.008 difference in AUC). As for Glass dataset the AUC score of our LAD (0.8612) is still comparable to the best one (0.8922) by Mass, meanwhile beats the third best (LOCI with $k = 50$) for more than 13%. IForest shows the second best (0.7459) average performance of real datasets, which supports the argument that it is able to take both global and local contexts into consideration [18]. This is different from kNN-based methods (LOF and LOCI) which only concern with instance-wise local context. Compared with LOF, LOCI performs robustly when k varies. It ranks the third (0.7297) when $k = 50$ and fifth (0.7146) when $k = 10$. This moderately stable performance comes from the built-in concept of a multi-granularity deviation factor. LOF, although has the third best score (0.7190) when $k = 50$, shows seriously unstable performance as k changes, which can be explained as follows: LOF is based on a direct normalization of anomaly scores for a very limited neighborhood. Although Mass (0.7184) only keeps the fourth record, it has the fastest computation speed compared with the other competitors, especially LAD.

Table 3 shows performance of three different manifold-based algorithms. Generally, LAD evidently outperforms the other methods with average AUC 0.7889. In terms of stability, although LLE has similar average score with $k = 10$, 50 and 100, it shows fluctuation for some datasets especially Ecoli, Glass and Ionosphere. Part of the reason comes from that LLE assumes the data manifold is sufficiently smooth and densely sampled that it is locally approximately linear, while this is not the true story for many real world datasets. Similarly, ISM's AUC score varies as k changes in several datasets (Ecoli, Yeast, Glass etc.). It is because ISM is highly vulnerable to the local data perturbation since the embedding given by the ISM tends to recovers the geodesic distances between points on the man-

Table 3: Comparison of LAD and other two manifold-based methods (with different k) on nine datasets using AUC metrics. The bold-faced numbers indicate the best method on a particular dataset; the numbers in parentheses indicate the ranks of our LAD. Average is the average of performance across all the datasets respectively. A $**$ indicates a p-value of 1% or lower in the statistical significance test w.r.t. LAD.

Dataset	LLE(k=10)	LLE(k=50)	LLE(k=100)	ISM(k=10)	ISM(k=50)	ISM(k=100)	LAD
WDBC	0.7572	0.7561	0.8517	0.5903	0.6359	0.6942	**0.8864** (1)
Pima	0.5183	0.5815	0.6270	0.5234	0.5925	0.5546	**0.6936** (1)
Arrhythmia	0.5564	0.6174	0.6234	0.5694	0.5878	0.6184	**0.7558** (1)
Ecoli	**0.9205**	0.8709	0.5739	0.6259	0.7934	0.8406	0.8692 (3)
Yeast	0.5994	0.6289	**0.6464**	0.4924	0.5802	0.5965	0.6183 (3)
Abalone	0.5764	0.6054	0.6039	0.5520	0.6637	0.6513	**0.7398** (1)
Glass	0.8385	0.6797	0.7100	0.5453	0.6607	0.7957	**0.8612** (1)
Ionosphere	0.6125	0.4526	0.4731	0.4444	0.4188	0.4837	**0.9240** (1)
Magic	0.5638	0.6152	0.6010	0.5742	0.5981	0.6328	**0.7516** (1)
Average	0.6603**	0.6453**	0.6344**	0.5464**	0.6146**	0.6520**	**0.7889** (1)

ifold, which is very locally sensitive compared with random walk [16] [37].

To systematically manifest the robustness of our proposed LAD on different neighborhood size k and Gaussian scaling parameter σ, we test our algorithm respectively on a series of k and σ on seven small datasets: WDBC, Pima, Arrhythmia, Ecoli, Yeast, Glass, and Ionosphere due to limited space, and also with the reason that in theory, datasets with smaller number of instances are more sensitive to the change of k and σ. Therefore these seven datasets are the more effective choices to show whether our LAD is robust to these two parameters. For k our test range is in $[10, 100]$. As for σ, the test range is in $[0.1, 8]$, with 0.05 as step size between 0.1 to 1 and 0.5 as step size between 1 to 8. From Figure 7(a) we can see that our new LAD algorithm has more stable performance than LOF, LLE and ISM on different k (shown in Table 2 and Table 3). Similarly, Figure 7(b) shows that our proposed LAD retains certain level of robustness as σ changes. The stability of LAD has an inherent relationship with diffusion maps and random walk.

As for the macro paired t-tests across all the datasets, compared with LOF, LOCI, Mass, LLE and ISM respectively, LAD has extremely small p-value (less than 1%). Compared with IForest, LAD has p-value less than 5%. This, once again, proves that our LAD has the most stable average performance. Overall, LAD outperforms the selected kNN-based, attribute-based and manifold-based algorithms in that it is more intrinsic, informative, and manifold-aware.

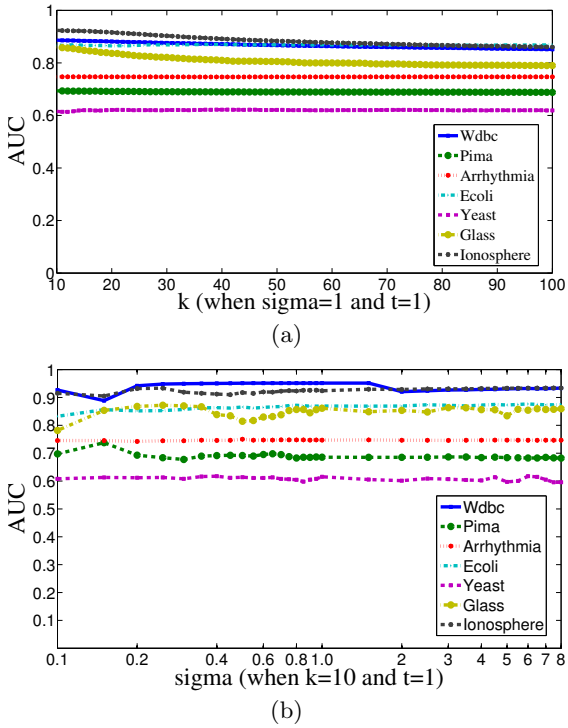


Figure 7: LAD robustness on different k and σ.

7. CONCLUSION

This paper has documented an original unsupervised anomaly detection algorithm, called Local Anomaly Descriptor (LAD), which is based on the physics-inspired diffusion space and weighted umbrella operator. Compared with the existing algorithms, our proposed LAD has demonstrated many important properties such as intrinsic, informative to local density, and stable to small parameter perturbation. Together with its more manifold-aware property for the goal of anomaly detection, we expect it to be useful for any type of data distribution. Nonetheless, much more extensive experiments are still required to validate this conjecture, which is part of our near-future research. Another direction is to investigate the possible connection with global structure and pattern mining such as clustering and feature classification.

8. ACKNOWLEDGEMENTS

We gratefully thank all the anonymous reviewers for constructive suggestions toward paper improvement. This research is supported in part by NSF grants IIS-0949467, IIS-1047715, and IIS-1049448. It is also supported by United States Department of Energy, Grant No. DE-SC0003361, funded through the American Recovery and Reinvestment Act of 2009. In addition, this project is also supported in part by DOE Systems Biology Knowledgebase (DE-AC02-98CH10886).

9. REFERENCES

[1] A. Agovic, A. Banerjee, A. R. Ganguly, and V. Protopopescu. Anomaly detection in transportation corridors using manifold embedding. *the 1st*

International Workshop on Knowledge Discovery from Sensor Data, 2007.

[2] R. Badeau, B. David, and G. Richard. Fast approximated power iteration subspace tracking. *IEEE Signal Processing*, pages 2931–2941, 2005.

[3] G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *JMLR*, 11:2973–3009, 2010.

[4] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. *ACM SIGMOD*, pages 93–104, 2000.

[5] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: a survey. *ACM Computing Surveys*, 2009.

[6] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[7] T. de Vires, S. Chawla, and M. Houle. Finding local anomalies in very highdimensional space. *IEEE ICDM*, pages 128–137, 2010.

[8] M. Desbrun, M. Meyer, P. Schrder, and A. Barr. Implicit fairing of arbitrary meshes using diffusion and curvature flow. *ACM SIGGRAPH*.

[9] J. Gao, H. Cheng, and P. N. Tan. Semi-supervised outlier detection. *ACM SAC*, pages 635–636, 2006.

[10] F. Goes, S. Goldenstein, and L. Velho. A hierarchical segmentation of articulated bodies. *Symposium on Geometry*, 2008.

[11] A. Grigoryan. Estimates of heat kernels on riemannian manifolds. *ICMS*, pages 140–225, 1999.

[12] D. Hawkins. *Identification of outliers*. Chapman and Hall, London, 1980.

[13] K. Hempstalk, E. Frank, and I. H. Witten. One-class classification by combining density and class probability estimation. *European Conference on Machine Learning and Knowledge Discovery in Databases*, 2008.

[14] E. Hsu. Stochastic analysis on manifolds. *Graduate Studies in Mathematics*, 2002.

[15] H. Huang, S. Yoo, H. Qin, and D. Yu. A robust clustering algorithm based on aggregated heat kernel mapping. *IEEE ICDM*, pages 270–279, 2011.

[16] S. Lafon, Y. Keller, and R. R. Coifman. Data fusion and multi-cue data matching by diffusion maps. *IEEE TPAMI*, 28(11):1784–1797, 2006.

[17] Y. Lipman, R. Rustamov, and T. Funkhouser. Biharmonic distance. *ACM Transactions on Graphics*, 2010.

[18] F. T. Liu and K. M. Ting. Can isolation-based anomaly detectors handle arbitrary multi-modal patterns in data? *Technical Report*, 2010.

[19] F. T. Liu, K. M. Ting, and Z. H. Zhou. Isolation forest. *IEEE ICDM*, pages 413–422, 2008.

[20] F. T. Liu, K. M. Ting, and Z. H. Zhou. Isolation-based anomaly detection. *ACM TKDD*, 2011.

[21] C. Marzban. A comment on the roc curve and the area under it as performance measures. *Technical Report*, 2004.

[22] M. Meila and J. Shi. A random walks view of spectral segmentation. *the 8th International Workshop on Artificial Intelligence and Statistics*, 2001.

[23] K. Noto, C. E. Brodley, and D. Slonim. Anomaly detection using an ensemble of feature models. *IEEE ICDM*, 2010.

[24] V. Y. Pan and Z. Q. Chen. The complexity of the matrix eigenproblem. *ACM STOC*, pages 507–516, 1999.

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

[26] D. Pelleg and A. W. Moore. Active learning for anomaly and rare-category detection. *NIPS'05*, pages 1073–1080, 2005.

[27] C. Plant, C. Bohm, B. Tilg, and C. Baumgartner. Enhancing instance-based classification with local density: a new algorithm for classifying unbalanced biomedical data. *Bioinformatics, Oxford Journals*, pages 981–988, 2010.

[28] B. Pogorelc and M. Gams. Discovery of gait anomalies from motion sensor data. *IEEE ICTAI*, pages 331–336, 2010.

[29] H. Qiu and E. R. Hancock. Clustering and embedding using commute times. *IEEE TPAMI*, 29(11):1873–1890, 2007.

[30] J. W. Richards, P. E. Freeman, A. B. Lee, and C. M. Schafer. Accurate parameter estimation for star formation history in galaxies using sdss spectra. *MNRAS*, pages 1044–1057, 2009.

[31] A. Singer and R. R. Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226–239, 2008.

[32] A. Singer and R. R. Coifman. Anisotropic diffusion on sub-manifolds with application to earth structure classification. *Technical Report*, 2011.

[33] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. *SGP*, 2009.

[34] Z. Syed and I. Rubinfeld. Unsupervised risk stratification in clinical datasets: Identifying patients at risk of rare outcomes. *ICML*, pages 1023–1030, 2010.

[35] G. Taubin. A signal processing approach to fair surface design. *ACM SIGGRAPH*, 1995.

[36] K. M. Ting, G. T. Zhou, F. T. Liu, and J. S. Tan. Mass estimation and its applications. *ACM KDD*, 2010.

[37] L. van der Maaten, E. Postma, and J. van der Herik. Dimensionality reduction: A comparative review. *Technical report*, 2009.

[38] M. Wu and J. Ye. A small sphere and large margin approach for novelty detection using training data with outliers. *IEEE TPAMI*, 31(11):2088–2092, 2009.

[39] K. Zhang, M. Hutter, and H. Jin. A new local distance-based outlier detection approach for scattered real-world data. *Advances in Knowledge Discovery and Data Mining*, pages 813–822, 2009.

[40] X. Zhu, X. Wu, and C. Zhang. Vague one-class learning for data streams. *IEEE ICDM*, pages 657–666, 2009.

[41] D. W. Zimmerman. A note on interpretation of the paired-samples t test. *Journal of Educational and Behavioral Statistics*, 22(3):349–360, 1997.