RKOF: Robust Kernel-Based Local Outlier Detection*

Jun Gao[1], Weiming Hu[1], Zhongfei (Mark) Zhang[2], Xiaoqin Zhang[3], and Ou Wu[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{jgao,wmhu,wuou}@nlpr.ia.ac.cn
[2] Dept. of Computer Science, State Univ. of New York at Binghamton, Binghamton, NY 13902, USA
zhongfei@cs.binghamton.edu
[3] College of Mathematics & Information Science, Wenzhou University, Zhejiang, China
xqzhang@wzu.edu.cn

Abstract. Outlier detection is an important and attractive problem in knowledge discovery in large data sets. The majority of the recent work in outlier detection follow the framework of Local Outlier Factor (LOF), which is based on the density estimate theory. However, LOF has two disadvantages that restrict its performance in outlier detection. First, the local density estimate of LOF is not accurate enough to detect outliers in the complex and large databases. Second, the performance of LOF depends on the parameter k that determines the scale of the local neighborhood. Our approach adopts the variable kernel density estimate to address the first disadvantage and the weighted neighborhood density estimate to improve the robustness to the variations of the parameter k, while keeping the same framework with LOF. Besides, we propose a novel kernel function named the Volcano kernel, which is more suitable for outlier detection. Experiments on several synthetic and real data sets demonstrate that our approach not only substantially increases the detection performance, but also is relatively scalable in large data sets in comparison to the state-of-the-art outlier detection methods.

Keywords: Outlier detection, Kernel methods, Local density estimate.

1 Introduction

Compared with the other knowledge discovery problems, outlier detection is arguably more valuable and effective in finding rare events and exceptional cases from the data in many applications such as stock market analysis, intrusion detection, and medical diagnostics. In general, there are two definitions of the

* This work is supported in part by the NSFC (Grant No. 60825204, 60935002 and 60903147) and the US NSF (Grant No. IIS-0812114 and CCF-1017828).

J.Z. Huang, L. Cao, and J. Srivastava (Eds.): PAKDD 2011, Part II, LNAI 6635, pp. 270–283, 2011.

outlier detection: Regression outlier and Hawkins outlier. Regression outlier defines that an outlier is an observation which does not match the predefined metric model of the interesting data [1]. Hawkins outlier defines that an outlier is an observation that deviates so much from other observations as to arouse suspicion that this observation is generated by a different mechanism [2]. Compared with Regression outlier detection, Hawkins outlier detection is more challenging work because of the unknown generative mechanism of the normal data. In this paper, we focus on the unsupervised methods for Hawkins outlier detection. In the rest of this paper, outlier detection refers particularly to Hawkins outlier detection.

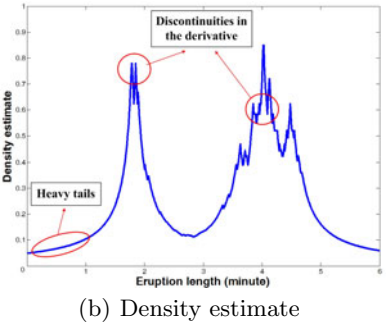
Over the past several decades, the research on outlier detection varies from the global computation to the local analysis, and the descriptions of outliers vary from the binary interpretations to probabilistic representations. Breunig et al. propose a density estimation based Local Outlier Factor (LOF) [4]. This work is so influential that there is a rich body of the literature on the local density-based outlier detection. On the one hand, plenty of local density-based methods are proposed to compute the outlier factors, such as the local correlation integral [5], the connectivity-based outlier factor [8], the spatial local outlier measure [9], and the local peculiarity factor [7]. On the other hand, many efforts are committed to combining machine learning methods with LOF to accommodate the large and high dimensional data [10,14].

Although LOF is popular in use in the literature, there are two major disadvantages restricting its applications. First, since LOF is based on the local density estimate theory, it is obvious that the more accurate the density estimate, the better the detection performance. The local reach-ability density used in LOF is the reciprocal of the average of reach-distances between the given object and its neighbors. This density estimate is an extension of the nearest neighbor density estimate, which is defined as

$$f(p) = \frac{k}{2n} \cdot \frac{1}{d_k(p)} \tag{1}$$

4.7	3.87	4.12	2.72	4.58	1.9
1.68	1.73	4	4.03	3.5	4.08
1.75	3.92	4.93	1.73	4.62	3.43
4.35	3.2	3.68	3.1	4.03	1.77
1.77	2.33	1.85	4.62	1.97	4.5
4.25	4.57	3.83	1.88	4.6	1.8
4.1	3.58	1.85	3.52	4	3.7
4.05	3.7	3.8	3.77	3.75	2.5
1.9	4.25	3.8	3.43	4	2.27
4	3.58	3.33	2	4.33	2.93
4.42	3.67	3.73	3.73	1.82	4.63
1.83	1.9	1.67	4.6	1.67	4
1.83	4.13	4.63	2.93	3.5	1.97
3.95	4.53	1.83	4.65	4.2	3.93
4.83	4.1	2.03	4.18	4.43	4.07
4.5	4.25	4.73	4.4	4.08	4.33
2.25	3.92	3.72	4.58	4.28	4.5
4.13	1.95	3.5	1.8		

(a) Old Faithful data



(b) Density estimate

Fig. 1. (a) Eruption lengths of 107 eruptions of Old Faithful geyser. (b) The density of Old Faithful data based on the nearest neighbor density estimate, redrawn from [3].

where n is the total number of the objects, and $d_k(p)$ is the distance between object p and its kth nearest neighbor. As shown in Fig. 1, the heavy tails of the density function and the discontinuities in the derivative reduce the accuracy of the density estimate. This dilemma indicates that with the LOF method, an outlier is unable to deviate substantially from the normal objects in the complex and large databases. Second, like all other local density-based outlier detection methods, the performance of LOF depends on the parameter k which is defined as the least number of the nearest neighbors in the neighborhood of an object [4]. However, in LOF, the value of k is determined based on the average density estimate of the neighborhood, which is statistically vulnerable to the presence of an outlier. Hence, it is hard to determine an appropriate value of this parameter to ensure the acceptable performance in the complex and large databases.

In order to address these two disadvantages of LOF, we propose a Robust Kernel-based Outlier Factor (RKOF) in this paper. Specifically, the main contributions of our work are as follows:

- We propose a kernel-based outlier detection method which brings the variable kernel density estimate method into the computation of outlier factors, in order to achieve a more accurate density estimate. Besides, we propose a new kernel function named the Volcano kernel which requires a smaller value of the parameter k for outlier detection than other kernels, resulting in less detection time.
- We propose the weighted density estimate of the neighborhood of a given object to improve the robustness of determining the value of the parameter k. Furthermore, we demonstrate that this weighted density estimate method is superior to the average density estimate method used in LOF in robust outlier detection.
- We keep the same framework of local density-based outlier detection with LOF. This makes that RKOF can be directly used in the extensions of LOF, such as Feature Bagging [10], Top-n outlier detection [14], Local Kernel Regression [15], and improve the detection performance of these extensions.

The remainder of this paper is organized as follows. Section 2 introduces our RKOF method with a novel kernel function, named the Volcano kernel, and analyzes the special property of the Volcano kernel. Section 3 shows the robustness and computational complexity of RKOF. Section 4 reports the experimental results. Finally, Section 5 concludes the paper.

2 Main Framework

A density-based outlier is detected by comparing its density estimate with its neighborhood density estimate [4]. Hence, we first introduce the notions of *the local kernel density estimate of object p*, *the weighted density estimate of p's neighborhood*. Then, we introduce the notion of *the robust kernel-based outlier factor of p*, which is used to detect outliers. Besides, we analyze the influences of different kernels to the performance of our method, and propose a novel kernel function named *the Volcano kernel* with its special property in outlier detection.

To make this work self-contains, we introduce the notions of *the k-distance of an object p*, and *the k-distance neighborhood of p*, which are defined in LOF.

Definition 1. *Given a data set D, an object p, and any positive integer k, the k-distance(p) is defined as the distance $d(p, o)$ between p and an object $o \in D$, such that:*

- *for at least k objects $o' \in D\backslash\{p\}$, it holds that $d(p, o') \leq d(p, o)$.*
- *for at most $k-1$ objects $o' \in D\backslash\{p\}$, it holds that $d(p, o') < d(p, o)$.*

Definition 2. *Given a data set D, an object p, and any positive integer k, the k-distance neighborhood of p, named $N_k(p)$, contains every object whose distance from p is not greater than the k-distance(p), i.e., $N_k(p) = \{q \in D\backslash\{p\} | d(p, q) \leq k\text{-distance}(p)\}$, where any such object q is called a k-distance neighbor of p. $|N_k(p)|$ is the number of the k-distance neighbors of p.*

2.1 Robust Kernel-Based Outlier Factor (RKOF)

Let $p = [x_1, x_2, x_3, \ldots, x_d]$ be an object in the data set D, where d is the number of the attributes. $|D|$ is the number of all the objects in D.

Definition 3. (Local kernel density estimate of object p)
The local kernel density estimate of p is defined as

$$kde(p) = \frac{\sum_{o \in N_k(p)} \left\{ h^{-\gamma} \lambda_o^{-\gamma} K(h^{-1} \lambda_o^{-1} (p - o)) \right\}}{|N_k(p)|}$$

$$\lambda_o = \{f(o)/g\}^{-\alpha} \quad \log g = |D|^{-1} \sum_{q \in D} \log\left(f(q)\right)$$

where h is the smoothing parameter, γ is the sensitivity parameter, $K(x)$ is the multivariate kernel function and λ_o is the local bandwidth factor. $f(x)$ is a pilot density estimate that satisfies $f(x) > 0$ for all the objects, α is the sensitivity parameter that satisfies $0 \leq \alpha \leq 1$, and g is the geometric mean of $f(x)$.

$kde(p)$ is an extension of the variable kernel density estimate [3]. $kde(p)$ not only retains the adaptive kernel window width that is allowed to vary from one object to another, but also is computed locally in the k-distance neighborhood of object p. The parameter γ equals the dimension number d in the original variable kernel density estimate [3]. For the local kernel density estimate, the larger γ, the more sensitive $kde(p)$. However, the high sensitivity of $kde(p)$ is not always a merit for the local outlier detection in high dimensional data. For example, if λ_o is always very small for all the objects in a sparse and high dimensional data set, $(\lambda_o)^{-\gamma}$ always equals infinity. This makes $kde(p)$ lack of the capacity to discriminate between outliers and normal data. We give γ a default value 2 to obtain a balance between the sensitivity and the robustness.

In this paper, we compute the pilot density function $f(x)$ by the approximate nearest neighbor density estimate according to Equation 1.

$$f(o) = \frac{1}{k\text{-}distance(o)} \tag{2}$$

Substituting Equation 2 into $kde(p)$ in Definition 3, we obtain Equation 3, where the default values of C and α are 1. In the following experiments, we estimate the local kernel density of object p as follows:

$$kde(p) = \frac{\sum\limits_{o \in N_k(p)} \frac{1}{(C \cdot k\text{-}distance(o)^{\alpha})^2} K(\frac{p-o}{C \cdot k\text{-}distance(o)^{\alpha}})}{|N_k(p)|} \qquad C = h \cdot g^{\alpha} \tag{3}$$

Definition 4. (Weighted density estimate of object p's neighborhood) *The weighted density estimate of p's neighborhood is defined as*

$$wde(p) = \frac{\sum\limits_{o \in N_k(p)} \omega_o \cdot kde(o)}{\sum\limits_{o \in N_k(p)} \omega_o} \qquad \omega_o = \exp\left\{ -\frac{(\frac{k\text{-}distance(o)}{min_k} - 1)^2}{2\sigma^2} \right\}$$

where ω_o is the weight of object o in the k-distance neighborhood of object p, σ is the variance with the default value 1*, and $min_k = \min_{o \in N_k(p)}(k\text{-}distance(o))$.*

In the majority of local density-based methods, outlier factor is computed by the ration of the neighborhood's density estimate to the given object's density estimate. The neighborhood's density is generally measured by the average value of all the neighbors' local densities in the neighborhood. In this estimate approach, the detection performance is sensitive to the parameter k. The larger the value of k, the larger the scale of the neighborhood. When k is large enough that the majority in the neighborhood are normal objects, outliers have the chance to be detected. In the weighted neighborhood density estimate, the weight of the neighbor object is a monotonically decreasing function of its k-distance. The neighbor object with the smallest k-distance has the largest weight 1. Compared with the average neighborhood density estimate, the weighted neighborhood density estimate makes that outliers can be detected accurately even if the number of outliers in the neighborhood equals the number of normal objects. This means that the interval of the acceptable k in the weighted neighborhood density estimate is much larger than that of the average neighborhood density estimate, and our method is more robust to the variations of the parameter k.

Definition 5. (Robust kernel-based outlier factor of object p) *The robust kernel-based outlier factor of p is defined as*

$$RKOF(p) = \frac{wde(p)}{kde(p)}$$

where $wde(p)$ is the density estimate of the k-distance neighborhood of p, and $kde(p)$ is the local density estimate of p.

RKOF is computed by dividing the weighted density estimate of the neighborhood of the given object by its local kernel density estimate. The larger RKOF, the more probable to be an outlier the given object. It is obvious that the smaller the object p's local kernel density, and the larger the weighted density of its neighborhood, the larger its outlier factor.

2.2 Choice of Kernel Functions

In LOF method, for most objects in a cluster, their outlier factors are approximately equal to 1; for most outliers isolated from the cluster, their outlier factors are much larger than 1 [4]. This property makes it easy to distinguish between outliers and normal objects.

The multivariate Gaussian and Epanechnikov kernel functions are commonly used in the kernel density estimation, whose formulations are defined as follows:

$$K(x) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\|x\|^2\right) \tag{4}$$

$$K(x) = \begin{cases} (3/4)^d(1 - \|x\|^2), & \textbf{if } \|x\| \leq 1 \\ 0, & \textbf{otherwise} \end{cases} \tag{5}$$

where $\|x\|$ denotes the norm of a vector x and it can be used to compute the distances between objects.

Our RKOF method with the Gaussian kernel cannot ensure that outlier factors of the normal objects in a cluster are approximately equal to 1. Then, we need to determine the threshold value of outlier factors in addition. The Epanechnikov kernel function equals zero when $\|x\|$ is larger than 1. Hence, for most of outliers and normal objects lying in the border of clusters, their outlier factors equal infinity.

In order to achieve the same property with LOF, we define a novel kernel function called the Volcano kernel as follows:

Definition 6. *The Volcano kernel is defined as*

$$K(x) = \begin{cases} \beta, & \textbf{if } \|x\| \leq 1 \\ \beta g(\|x\|), & \textbf{otherwise} \end{cases}$$

where β assures that $K(x)$ integrates to one, and $g(x)$ is a monotonically decreasing function, lying in a close interval $[0, 1]$ and equal to zero at the infinity. Unless otherwise specified, we use $g(x) = e^{-|x|+1}$ as the default function for our experiments.

Fig. 2 shows the curve of the Volcano kernel for the univariate data. When $\|x\|$ is not larger than 1, the kernel value equals a constant value β. This generates that outlier factors of the objects deeply in the cluster are approximately equal to 1. When $\|x\|$ is larger than 1, the kernel value is the monotonically decreasing

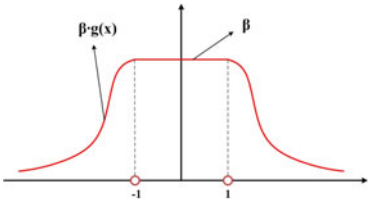


Fig. 2. The curve of the Volcano kernel for the univariate data

function of $\|x\|$ and less than 1. This not only makes outlier factors continuous and finite, but also makes outlier factors of outliers much larger than 1. Hence, RKOF method with the Volcano kernel can capture outliers much easier, and sort all the objects according to their RKOF values.

3 Robustness and Computation Complexity of RKOF

In this section, we first analyze the robustness of RKOF to the parameter k. Then, we analyze the computation complexity of RKOF in detail.

Compared with the average neighborhood density estimate used in LOF, the weighted neighborhood density estimate defined in Definition 4 is more robust to the parameter k and it substantially helps improve the detection performance. As shown in Theorem 1, if the weighted neighborhood density estimate replaces the average neighborhood density estimate in the computation of outlier factors, any local density-based outlier detection method following the framework of LOF can be less sensitive to the parameter k.

Theorem 1. *Let $N_k(p)$ be the neighborhood of object p, and p be an outlier in a data set D. Let r be the proportion of the outliers in $N_k(p)$. Suppose that these outliers have the same local density estimate (DE) α and k-distance α' with p. Also suppose that the normal data in $N_k(p)$ have the same local density estimate β and k-distance β', with $\alpha < \beta$ and $\alpha' > \beta'$. The Outlier Factor (OF) can be computed based on any local density-based outlier detection method that follows the framework of LOF. Then for the average density estimate, it holds that:*

$$OF(p) = r + (1 - r)\rho$$

For the weighted density estimate, it holds that:

$$OF(p) = \frac{(\rho - w)r - \rho}{(1 - w)r - 1}$$

where $\rho = \beta/\alpha$ and ω is the weight of the outlier in $N_k(p)$.

Proof: *For the average density estimate:*

$$OF(p) = \frac{\sum\limits_{o \in N_k(p)} DE(o)}{|N_k(p)|DE(p)} = \frac{r|N_k(p)|\alpha + (1 - r)|N_k(p)|\beta}{|N_k(p)|\alpha} = r + (1 - r)\rho$$

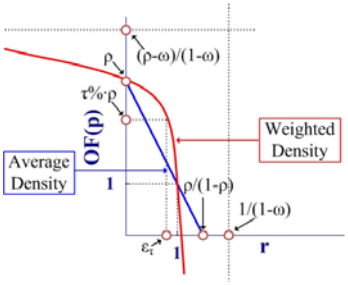


Fig. 3. The curves of $OF(p)$ for the average and the weighted density estimates

For the weighted density estimate:

Let ω_{o_i} *and* ω_{o_j} *be the weights of the outlier and the normal object, respectively. According to Definition 4,* $\omega_{o_i} < 1$ *and* $\omega_{o_j} = 1$ *because* $\alpha' > \beta'$*. Replacing* ω_{o_i} *with* ω*, we have*

$$OF(p) = \frac{\sum\limits_{o \in N_k(p)} \omega_o \cdot DE(o)}{DE(p) \cdot \sum\limits_{o \in N_k(p)} \omega_o} = \frac{r|N_k(p)|\omega\alpha + (1-r)|N_k(p)|\beta}{\alpha(r|N_k(p)|\omega + (1-r)|N_k(p)|)}$$
$$= \frac{r\omega + (1-r)\rho}{r\omega + (1-r)} = \frac{(\rho-\omega)r - \rho}{(1-\omega)r - 1}$$

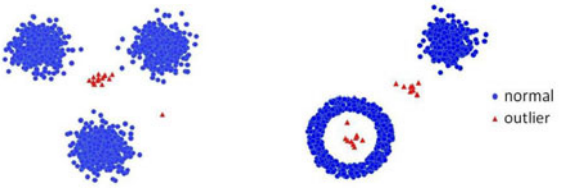
According to Theorem 1, $OF(p)$ is a function of the parameter r while ρ has a fixed value. r is determined by the parameter k. As shown in Fig. 3, for the average neighborhood density estimate, $OF(p)$ is a monotonically decreasing linear function when r increases. For the weighted neighborhood density estimate, $OF(p)$ is a quadratic curve of r. When $r \in [0, 1]$, $OF(p)$ of the average method is always much less than that of the weighted method. Fig. 3 shows that $OF(p)$ of the weighted method is larger than $\tau\%$ of the maximum of $OF(p)$ when $r \in [0, \epsilon_\tau]$. ϵ_τ depends on ρ and the weights of the outliers in the neighborhood of p. More importantly, $OF(p)$ is approximately a constant in the interval $[0, \epsilon_\tau]$. This property indicates that the weighted method makes the local outlier detection more robust to the variations of the parameter k.

Since RKOF shares the same framework with LOF, RKOF has the same computational complexity as that of LOF. To compute the RKOF values with the parameter k, the RKOF algorithm includes two steps. In the first step, the k-distance neighbors for each object need to be found with their distances to the given object computed in the data set D of n objects. The computational complexity of this step is $O(n \log n)$ by using the index technology for k-nn queries, which has been used in LOF [4]. In the second step, the $kde(p)$, $wde(p)$, and $RKOF(p)$ values are computed by scanning the whole data set. Since both $kde(p)$ and $wde(p)$ are computed in the k-distance neighborhood of the given object, the computational complexity of this step is $O(nk)$. Hence, the total

computational complexity of the RKOF algorithm is $O(n \log n + nk)$. Clearly, the larger k is, the more the running time is consumed.

4 Experiments

In this section, we evaluate the outlier detection capability of RKOF based on different kernel functions and compare RKOF with the state-of-the-art outlier detection methods on several synthetic and real data sets.



(a) Synthetic-1 (b) Synthetic-2

Fig. 4. The distributions of the synthetic data sets

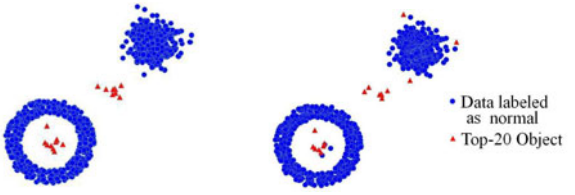
4.1 Synthetic Data

As shown in Fig. 4, the Synthetic-1 data set consists of 1500 normal objects and 16 outliers with two attributes. The normal objects distribute in three Gaussian clusters with 500 normal objects in each cluster with the same variance, respectively. Fifteen outliers lie in a dense Gaussian cluster, and the other outlier is isolated from the others. The Synthetic-2 data set consists of 500 normal objects uniformly in the annular region, 500 normal objects in a Gaussian cluster, and 20 outliers in two Gaussian clusters.

Table 1 exhibits the outlier detection results of LOF and RKOF on the Synthetic-1 data set, respectively, where σ is the parameter of the weight in RKOF. Top-16 objects are the sixteen objects that have the largest outlier factors in the synthetic data set. Obviously, if all top-16 objects are outliers, the

Table 1. Outlier detection for the Synthetic-1 data set

k	Number of outliers in the top-16 objects (*coverage*)		
	LOF	RKOF($\sigma = 0.1$)	RKOF($\sigma = 1$)
26	1(6.25%)	15(93.75%)	15(93.75%)
27	2(12.5%)	16(100%)	15(93.75%)
30	4(25%)	16(100%)	15(93.75%)
31	5(31.25%)	16(100%)	16(100%)
59	15(93.75%)	16(100%)	16(100%)
60	16(100%)	16(100%)	16(100%)
70	16(100%)	16(100%)	16(100%)



(a) RKOF: k=14 (b) LOF: k=20

Fig. 5. The best performances of RKOF and LOF on the Synthetic-2 data (Top-20)

detection rate is 100% and the false alarm rate is zero. *coverage* is the ratio of the number of the detected outliers to the 16 total outliers. RKOF($\sigma = 0.1$) can identify all the outliers when $k \geq 27$. RKOF($\sigma = 1$) can detect all the outliers when $k \geq 31$. Clearly, the parameter σ directly relates to the sensitivity of the outlier detection for RKOF. LOF is unable to identify all the outliers until $k = 60$. Table 1 indicates that the available k interval of RKOF is larger than that of LOF, which means that RKOF is less sensitive to the parameter k.

As shown in Fig. 5, RKOF with $k = 14$ captures all the outliers in Top-20 objects. LOF obtains its best performance with $k = 20$, whose detection rate is 85%. Compared with RKOF, LOF can not detect all the outliers whatever the value of k is. It is obviously that the annular cluster and the Gaussian cluster pose an obstacle to the choice of k. This result indicates that RKOF is more adapted to the complex data sets than LOF.

4.2 Real Data

We compare RKOF with several state-of-the-art methods, including LOF [4], LDF [6], LPF [7], Feature Bagging [10], Active Learning [11], Bagging [12], and Boosting [13], on the real data sets. The performance of RKOF with the Gaussian, Epanechnikov, and Volcano kernels is also compared. In the real data sets, the features of the original data include discrete features and continuous features. All the data are processed using the standard text processing techniques following the original steps of the methods [7,11,10].

These real data sets consist of the KDD Cup 1999, the Mammography data set, the Ann-thyroid data set, and the Shuttle data set, all of which can be downloaded from the UCI database except the Mammography data set[1]. The KDD Cup 1999 is a general data set condensed for the intrusion detection research. 60593 normal records and 228 U2R attack records labeled as outliers are combined as the KDD outlier data set. All the records are described by 34 continuous features and 7 categorical features. The Mammography data set includes 10923 records labeled 1 as normal data and another 260 records labeled 2 as outliers; all the records consist of 6 continuous features. The Ann-thyroid data set consists of 73 records labeled 1 as outliers and 3178 records labeled 3

[1] Thank Professor Nitesh.V.Chawla for providing this data set.

Table 2. The AUC values and the running time in parentheses for RKOF and the comparing methods on the real data sets by the k-d tree method [17]. Since LPF has the higher complexity and is unable to complete the data sets in the reasonable time, the accurate running time for LPF is not given in this table.

Methods \ Data	KDD	Mammography	Ann-thyroid	Shuttle (average)
RKOF[a]	0.962 (1918.1s)	0.871 (15.8s)	0.970 (4.9s)	0.990 (36.4s)
RKOF[b]	0.961 (2095.2s)	0.870 (19.8s)	0.970 (5.2s)	0.990 (36.9s)
RKOF[c]	0.944 (2363.7s)	0.855 (48.2s)	0.965 (13.2s)	0.993 (36.7s)
LOF	0.610 (2160.1s)	0.640 (28.8s)	0.869 (5.9s)	0.852 (42.0s)
LDF	0.941 (2214.9s)	0.824 (36.4s)	0.943 (7.2s)	0.962 (37.1s)
LPF	0.98 ($\gg$2363.7s)	0.87 ($\gg$48.2s)	0.97 ($\gg$13.2s)	0.992 ($\gg$42.0s)
Bagging	0.61($\pm$0.25)	0.74($\pm$0.07)	0.98($\pm$0.01)	0.985($\pm$0.031)
Boosting	0.51($\pm$0.004)	0.56($\pm$0.02)	0.64	0.784($\pm$0.13)
Feature Bagging	0.74($\pm$0.1)	0.80($\pm$0.1)	0.869	0.839
Active Learning	0.94($\pm$0.04)	0.81($\pm$0.03)	0.97($\pm$0.01)	0.999($\pm$0.0006)

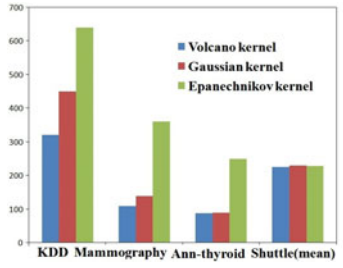
a. Using Volcano kernel b. Using Gaussian kernel c. Using Epanechnikov kernel

as normal data. There are 21 attributes where 15 attributes are binary and 6 attributes are continuous. The Shuttle data set consists of 11478 records with label 1, 13 records with label 2, 39 records with label 3, 809 records with label 5, 4 records with label 6, and 2 records with label 7. We divide this data set into 5 subsets: label 2, 3, 5, 6, 7 records vs label 1 records, where the label 1 records are normal, and others are outliers.

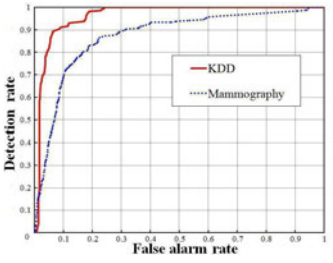
All the comparing outlier detection methods are evaluated using the ROC curves and the AUC values. The ROC curve represents the trade-off between the detection rate as y-axis and the false alarm rate as x-axis. The AUC value is the surface area under the ROC curve. Clearly, the larger the AUC value, the better outlier detection method.

The AUC values for RKOF with different kernels and all other comparing methods are given in Table 2. Also shown in Table 2 are the running time data for RKOF with different kernels as well as those of the other three local density-based methods; since the AUC values for other comparing methods are directly obtained from their publications in the literature, the running time data for these methods are not available and thus are not included in this table.

From Table 2, we see that different RKOF methods using different kernels receive similar AUC values on all the data sets, especially the Volcano and Gaussian kernels. The k values with the best detection performance for all the three kernels on all the data sets are shown in Fig. 6(a). Clearly, the k values for the Volcano kernel are always smaller than those of the other kernels, and the k values for the Epanechnikov kernel are the largest among three kernels. This experiment supports one of the contributions of this work that the proposed Volcano kernel achieves the least computation time among the existing kernels.

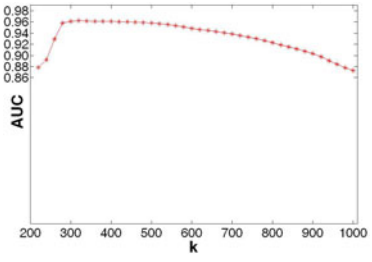


(a) k for different kernels

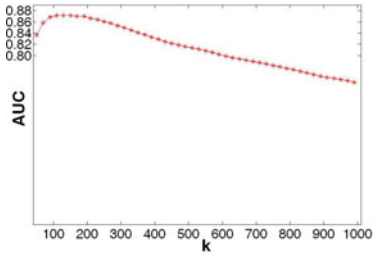


(b) ROC curves

Fig. 6. (a)The k values with the best performance for different kernels in RKOF. (b) ROC curves for RKOF based on the Volcano kernel on the KDD and the Mammography data sets.



(a) KDD



(b) Mammography

Fig. 7. AUC values of RKOF based on the Volcano kernel with different k values for the KDD and Mammography data sets

It indicates that different kernels used in RKOF do not significantly influence the detection performance, but they dramatically change the minimal k value with the acceptable performance and consequently the running time.

Fig. 6(b) shows the ROC curves of RKOF based on the Volcano kernel for the KDD data set ($k = 320$) and the Mammography data set ($k = 110$). Fig. 7 shows the AUC values of RKOF based on the Volcano kernel with different k values for the KDD and Mammography data sets. The AUC values for the KDD data set are larger than 0.941, when k varies from 280 to 700; the AUC values for the Mammography data set are larger than 0.824, when k changes from 40 to 460. Clearly, the detection performance of RKOF for any k in these interval is better than that of the other comparing methods except LPF. For the Mammography data set, RKOF is more effective than the other comparing methods with $k = 110$, compared with $k = 11183$ for LPF. For the KDD data set, RKOF achieves the second best performance with $k = 320$. The best AUC value is achieved by LPF, but this AUC value is obtained when $k = 13000$. The complexity of RKOF is $O(n \log n + nk)$, compared with $O(nd \log n + ndk)$ for LPF, where d is the dimensionality of the data. It is clear that under the same circumstances LPF takes much longer time than RKOF while the AUC value of RKOF is very close to this best value. For the Ann-thyroid data set, RKOF

achieves the acceptable performance that is very close to the best performance. The AUC value of the Shuttle data set is the average AUC of all the five subsets, where the AUC values of the subsets with the label 5, label 6, and label 7 are all approximately equal to 1. RKOF also obtains the acceptable performance that is very close to the best performance for the Shuttle data set. Overall, while there is no winner for all the cases, RKOF always achieves the best performance or is close to the best performance in all the data sets with the least running time. In particular, RKOF achieves the best performance or is close to the best performance for the KDD and the Mammography data sets with much less running time, which are the two large data sets of all the four data sets. This demonstrates the high scalability of the RKOF method in outlier detection. Specifically, in all the cases RKOF always has less running time than LOF, LDF and LPF. Though the running time data for the other comparing methods are not available, from the theoretic complexity analysis it is clear that they would all take longer running time than RKOF.

5 Conclusions

We have studied the local outlier detection problem in this paper. We have proposed the RKOF method based on the variable kernel density estimate and the weighted density estimate of the neighborhood of an object, which have addressed the existing disadvantages of LOF and other density-based methods. We have proposed a novel kernel function named the Volcano kernel, which is more suitable for outlier detection. Theoretical analysis and empirical evaluations on the synthetic and real data sets demonstrate that RKOF is more robust and effective for outlier detection at the same time taking less computation time.

References

1. Rousseeuw, P.J., Leroy, A.M.: Robust Rgression and Outlier Detection. John Wiley and Sons, New York (1987)
2. Hawkins, D.: Identification of Outliers. Chapman and Hall, London (1980)
3. Silverman, B.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)
4. Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: Lof: Identifying density-based local outliers. In: SIGMOD, pp. 93–104 (2000)
5. Papadimitriou, S., Kitagawa, H., Gibbons, P.: Loci: Fast outlier detection using the local correlation integral. In: ICDE, pp. 315–326 (2003)
6. Latecki, L.J., Lazarevic, A., Pokrajac, D.: Outlier Detection with Kernel Density Functions. In: Perner, P. (ed.) MLDM 2007. LNCS (LNAI), vol. 4571, pp. 61–75. Springer, Heidelberg (2007)
7. Yang, J., Zhong, N., Yao, Y., Wang, J.: Local peculiarity factor and its application in outlier detection. In: KDD, pp. 776–784 (2008)
8. Tang, J., Chen, Z., Fu, A.W.-c., Cheung, D.W.: Enhancing effectiveness of outlier detections for low density patterns. In: Chen, M.-S., Yu, P.S., Liu, B. (eds.) PAKDD 2002. LNCS (LNAI), vol. 2336, pp. 535–548. Springer, Heidelberg (2002)

9. Sun, P., Chawla, S.: On local spatial outliers. In: KDD, pp. 209–216 (2004)
10. Lazarevic, A., Kumar, V.: Feature bagging for outlier detection. In: KDD, pp. 157–166 (2005)
11. Abe, N., Zadrozny, B., Langford, J.: Outlier detection by active learning. In: KDD, pp. 504–509 (2006)
12. Breiman, L.: Bagging predictors. J. Machine Learning 24(2), 123–140 (1996)
13. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55(1), 113–139 (1997)
14. Jin, W., Tung, A., Ha, J.: Mining top-n local outliers in large databases. In: KDD, pp. 293–298 (2001)
15. Gao, J., Hu, W., Li, W., Zhang, Z.M., Wu, O.: Local Outlier Detection Based on Kernel Regression. In: ICPR, pp. 585–588 (2010)
16. Barnett, V., Lewis, T.: Outliers in Statistic Data. John Wiley, New York (1994)
17. Bentley, J.L.: Multidimensional binary search trees used for associative searching. J. Communications of the ACM 18(9), 509–517 (1975)