Similarity Kernels for Nearest Neighbor-Based Outlier Detection

Ruben Ramirez-Padron[1], David Foregger[2], Julie Manuel[3], Michael Georgiopoulos[1], and Boris Mederos[4]

[1] School of Electrical Engineering and Computer Science, University of Central Florida, FL, USA
rramirez@knights.ucf.edu
[2] Wesleyan University, Middletown, CT, USA
[3] University of South Florida, Tampa, FL, USA
[4] Universidad Autónoma de Ciudad Juárez, Ciudad Juárez, Mexico

Abstract. Outlier detection is an important research topic that focuses on detecting abnormal information in data sets and processes. This paper addresses the problem of determining which class of kernels should be used in a geometric framework for nearest neighbor-based outlier detection. It introduces the class of similarity kernels and employs it within that framework. We also propose the use of isotropic stationary kernels for the case of normed input spaces. Two definitions of similarity scores using kernels are given: the k-NN kernel similarity score (kNNSS) and the summation kernel similarity score (SKSS). The paper concludes with preliminary experimental results comparing the performance of kNNSS and SKSS for outlier detection on four data sets. SKSS compared favorably to kNNSS.

Keywords: similarity kernels, similarity scores, outlier detection, nearest neighbors.

1 Introduction

Outlier detection is a growing field within the data mining community. It focuses on detecting abnormal observations in data sets and processes. Detection of credit card fraud, computer networks attacks and suspicious activity in electronic commerce are common applications. Hawkins defined an outlier as "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" [10]. Some of the current definitions of outliers use a global approach, e.g. [12], while others focus on a neighborhood around each data point. The work of Breunig et al. [4] was the first one using scores to describe outliers. Outlier detection methods range from traditional approaches like nearest neighbor-based and statistical methods to more recent approaches such as information theoretic and spectral outlier detection. Our work focuses on the nearest neighbor-based approach, which relies on similarities (or equivalently, on distances) between data points. A recent survey on the different approaches to outlier detection can be found in [5].

P.R. Cohen, N.M. Adams, and M.R. Berthold (Eds.): IDA 2010, LNCS 6065, pp. 159–170, 2010.

The computational complexity of straightforward implementations of nearest neighbor-based outlier detection is $O(n^2)$, where n is the number of data points. Although this is not efficient for large data sets, there are techniques that allow for almost linear complexity without sacrificing much accuracy [3,17,22].

A great number of outlier-detection algorithms are limited to numerical data. They benefit from a variety of statistical techniques and well-established metrics. Currently, there is an increasing interest in working on data sets with non-numerical attributes. Accordingly, several similarity-based outlier detection algorithms have been proposed to deal with non-numerical data. [5].

Kernels methods have given researchers the ability to deal with both numeric and non-numeric data types within a single framework [19]. Several works have proposed the "kernelization" of different outlier detection algorithms [8,19,14,18,15,20]. Kernel functions map input data points to a high dimensional feature space to which simple algorithms are applied implicitly. Of particular interest to our work is the geometric framework proposed in [8]. It applies nearest-neighbor outlier detection techniques to the high dimensional feature space. The distances between two points in the feature space can be easily calculated using the kernel function. However, it is an open problem to determine which classes of kernel functions are well suited to outlier detection problems [21,16].

Our work has three main contributions. First, we introduce the concept of similarity kernels and we argue that similarity kernels should be used for kernel nearest neighbor-based outlier detection. It is noted that our concept of similarity kernel extends the class of isotropic stationary kernels as defined in [9]. For the case of normed input spaces, we propose to restrict kernel nearest neighbor-based outlier detection to the class of isotropic stationary kernels. Those kernels guarantee invariance to translations and rotations in the input space.

Second, two similarity scores using kernels are defined: the k-NN kernel similarity score (kNNSS) and the summation kernel similarity score (SKSS). kNNSS is a characterization in terms of similarity kernels of the well-known k-NN score [5]. Recently, in [1], the outlier score of an observation x was defined as the sum of distances from x to a fixed number of its nearest neighbors. In contrast to that work, SKSS is computed as the sum of the similarities between x and all observations within a ball of a fixed radius centered on x. To our knowledge, SKSS is a new type of density-based score.

Finally, SKSS is compared to kNNSS on two numerical data sets and two categorical data sets. The Gaussian kernel was used in the numerical cases. The Hamming distance kernel [6] and a diffusion kernel [13] were used on the categorical data. It is proved that those categorical kernels are similarity kernels. Preliminary results suggest that SKSS can be a valuable similarity score for nearest neighbor-based outlier detection.

The paper is structured as follows: in Section 2 we provide a brief description of concepts and methods that are relevant to our paper. Our approach is presented in section 3. Our experimental findings are shown in Section 4. Finally, concluding remarks and a few comments regarding future work are provided in Section 5.

2 Related Work

2.1 Kernels

Kernel functions were introduced in machine learning as a way of finding non-linear patterns in data sets through the application of linear methods to a high dimensional representation of the data [7]. Kernels are positive semi-definite functions defined as follows: [19]

Definition 1. *A symmetric function $K : X \times X \rightarrow \mathbf{R}$ is called a positive semi-definite kernel function if and only if for any positive integer n, any choice of n objects $x_1, .., x_n \in X$, and any choice of real numbers $c_1, .., c_n$ the following property holds:*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0 \tag{1}$$

Kernel functions are defined on pairs of objects from a variety of data types. For every kernel function $K : X \times X \rightarrow \mathbf{R}$ there exists a unique mapping ϕ, $\phi : X \rightarrow H$, where H is a high dimensional feature space. Typically, the mapping ϕ is used implicitly through the expression $K(x_i, x_j) =< \phi(x_i), \phi(x_j) >$, where x_i and x_j are data points in X and $<,>$ denotes the inner product on H. Throughout this paper we often refer to a *kernel matrix* K instead of the kernel function K. A kernel matrix is an n x n matrix containing the values of the corresponding kernel function evaluated on all possible pairs of data points from a data set containing n objects. Kernels have been categorized based on whether they are local or not, stationary or non-stationary, separable or non-separable, among other characteristics [9]. Isotropic stationary kernels are of particular interest to our work:

Definition 2. *A kernel function $K : X \times X \rightarrow \mathbf{R}$ is isotropic stationary if there is a function $g_K : \mathbf{R} \rightarrow \mathbf{R}$ such that:*

$$K(x, z) = g_K(\|x - z\|) \tag{2}$$

It follows from this definition that isotropic stationary kernels are invariant to translations and rotations. The well-known Gaussian RBF kernel is an example of an isotropic stationary kernel. On the other hand, the linear kernel $K(x_i, x_j) = x_i^T x_j$ and the polynomial kernel of degree p, defined as $K(x_i, x_j) = (x_i^T x_j + 1)^p$, are examples of kernels that are not isotropic stationary. [9]

2.2 The Hamming Distance Kernel

The Hamming distance kernel [6] is based on the well-known Hamming distance. To follow the notation given in [6], let us assume a categorical data set consisting of m attributes, where the ith-attribute takes values in a finite categorical domain D_i. The cross product over all domains D_i is denoted by D^m, i.e. $D^m = \prod_{i=1}^{m} D_i$.

Definition 3. *The Hamming distance kernel function $K_H(s,t)$ between two input categorical objects s and t is defined as:*

$$K_H(s,t) = \sum_{u \in D^m} \prod_{i=1}^{m} \lambda^{\delta(u_i,s_i)} \lambda^{\delta(u_i,t_i)}$$

where $\lambda \in (0,1)$ and $\delta(s_i,t_i)$ is 0 when $s_i = t_i$, and 1 otherwise.

2.3 Diffusion Kernels

Diffusion kernels are a family of kernel functions defined on graphs [13]. Every data point that could possibly appear on the data set is considered a vertex of the graph. Two vertices are linked if and only if they differ on only one attribute. The following expression shows the diffusion kernel we used in our experiments:

$$K_{DK}(\beta)(x,y) = \prod_{i=1}^{m} \left(\frac{1 - e^{-|D_i|\beta}}{1 + (\mid D_i \mid -1)e^{-|D_i|\beta}} \right)^{\delta(x_i,y_i)} \tag{3}$$

where m is the number of attributes in the data set, $x = (x_1,..,x_m)$ and $y = (y_1,..,y_m)$ are categorical data points, $|D_i|$ is the number of values that the i-th attribute can take, and $\delta(x_i,y_i)$ returns 0 for $x_i = y_i$ and 1 otherwise.

2.4 Geometric Framework for Unsupervised Anomaly Detection

Nearest neighbor-based outlier detection techniques are based on the assumption that outliers appear in low density regions while normal data points are located in highly dense neighborhoods. A distance measure is thus required to be able to define neighborhoods for points in the data set. The geometric framework in [8] is based on the fact that the feature space H is a Hilbert space. Therefore, for every data points y_i, y_j in H the inner product $< y_i, y_j >$ is well defined, and it can be calculated as $K(\phi^{-1}(y_i), \phi^{-1}(y_j))$. Given two data points x_i, x_j from X, the distance $d_\phi(x_i,x_j)$ between x_i and x_j is defined as $d(\phi(x_i), \phi(y_j))$. By using the well-known distance equation $d(y_i,y_j) = \sqrt{< y_i,y_i > + < y_j,y_j > -2 < y_i,y_j >}$, the following distance equation is obtained:

$$d_\phi(x_i,x_j) = \sqrt{K(x_i,x_i) + K(x_j,x_j) - 2K(x_i,x_j)} \tag{4}$$

Nearest neighbor-based outlier detection methods can be applied to data sets of arbitrary type, provided there is a kernel for that data type.

3 Proposed Approach

This section is divided in three subsections. In the first subsection, we introduce the class of similarity kernels and discuss some basic properties. In the second subsection, a relationship between the class of similarity kernels and the class of isotropic stationary kernels is presented. The advantage of using isotropic stationary kernels for normed input spaces is discussed as well. In the third subsection we offer a definition for two kernel similarity scores on which outlier detection can be based.

3.1 Similarity Kernels for Nearest Neighbor-Based Outlier Detection

To the best of our knowledge, nothing has been published about the classes of kernels that should be used within the geometric framework proposed in [8]. In this paper, we address that question by establishing a relationship between the concept of similarity and a class of kernel functions that we define as similarity kernels. It is required that any similarity function must fulfill the following three properties: First, for any fixed input space X, the similarity of any object to itself is always equal to a constant c. Second, that constant c is an upper bound to the similarity between any two objects in X. Finally, there exists a number d such that d is a lower bound to the similarity of any two objects in X. Based on these properties, we introduce the following definition:

Definition 4. *A positive semi-definite kernel function K defined on some input space X is a similarity kernel if and only if there exist $c \in \mathbf{R}^+$ such that $K(x,x) = c$ for all $x \in X$.*

Note that given a similarity kernel K, equation 4 can be written as:

$$d_\phi(x_i, x_j) = \sqrt{2c - 2K(x_i, x_j)} \tag{5}$$

Consequently, $K(x_i, x_j)$ can be interpreted as a similarity measure between x_i and x_j. It remains to show that all similarity kernels K satisfy the properties we require from a similarity measure. The first of those properties is fulfilled by definition. To prove that $c = K(x,x)$ is an upper bound for all values of K, let us assume an arbitrary data set D with n objects $\{x_1, x_2, ..., x_n\}$. Let us denote by K the corresponding $n \times n$ kernel matrix. Let x_i and x_j be two arbitrary objects from D, and let z be a vector of length n containing 1 in the position i, -1 in position j and 0 in all remaining positions. Because K is positive semi-definite, we have $z^T K z \geq 0$. Consequently:

$$z^T K z = K(x_i, x_i) + K(x_j, x_j) - 2K(x_i, x_j) = 2c - 2K(x_i, x_j) \geq 0 \tag{6}$$

$$K(x_i, x_j) \leq c \tag{7}$$

Following the same approach, but assuming that z is the vector containing 1 in positions i and j and 0 in all remaining positions, it is concluded that -c is a lower bound for all values of the kernel matrix. As a direct consequence of these properties, similarity kernels can be normalized to the interval $[-1, 1]$ simply by dividing the kernel by c.

It is worthy of noting that the closure properties given in section 3.4.1 of [19], which establish how to obtain new kernels from previously defined kernels, are preserved for the class of similarity kernels, i.e.:

1. The sum of two similarity kernels is also a similarity kernel.
2. The product of a similarity kernel and a positive scalar is a similarity kernel.
3. If $K_1(x,y)$ and $K_2(x,y)$ are similarity kernels then $K(x,y) = K_1(x,y)K_2(x,y)$ is also a similarity kernel.

4. If $K(x, y)$ is a similarity kernel, and $\varphi : X \to \mathbf{R}^n$, then $K(\varphi(x), \varphi(y))$ is a similarity kernel.

Other classical operations over similarity kernels produce similarity kernels as well. For instance, a polynomial function with positive coefficients composed with a similarity kernel is a similarity kernel. The composition of an exponential function with a similarity kernel is also a similarity kernel.

3.2 Similarity Kernels and Isotropic Stationary Kernels

From definition 2, it is obvious that all isotropic stationary kernels are within the class of similarity kernels. Isotropic stationary kernels must be defined on normed spaces, while similarity kernels are not constrained to any particular input space. Consequently, the class of isotropic stationary kernels is a proper subset of the class of similarity kernels. For any isotropic stationary kernel, equation 4 can be written as follows:

$$d_\phi(x_i, x_j) = \sqrt{2g_K(0) - 2g_K(\|x_i - x_j\|)} \tag{8}$$

When X is a normed space, nearest neighbor-based outlier detection methods should be invariant to translations and rotations in X. From definition 2, it is clear that any nearest neighbor-based outlier detection based on an isotropic stationary kernel will remain invariant to translations and rotations in X. On the other hand, given any non-stationary kernel function K, there exists input values x_1, x_2 and $t \in X$ such that $K(x_1, x_2) \neq K(x_1 + t, x_2 + t)$. However, it is not clear whether the non-stationary quality of K could actually influence the accuracy of outlier detection in the presence of arbitrary translations or rotations. The following example provides a positive answer to that question for the case of an arbitrary translation.

We generated a two-dimensional data set containing 4 outliers and 46 non-outliers. A second data set was obtained by adding the vector (100, 100) to all points in the original data set. Figure 1 shows the performance of k-NN outlier detection [8] on both data sets. The polynomial kernel with $p = 2$ was used. The number of neighbors, k, was set equal to 3, and four data points were selected as outliers. The different results obtained show that outlier detection results can become inconsistent under translations in X when a non-stationary kernel is used. We believe that similar examples could be devised for arbitrary rotations when non-isotropic kernels are used. For the sake of invariance, we suggest the use of isotropic stationary kernels whenever possible.

3.3 Kernel Similarity Scores

Our kernel similarity scores make use of the interpretation of each value $K_{i,j}$ of a $n \times n$ similarity kernel matrix K as a similarity measurement. The goal is to calculate a similarity score vector $\mathbf{s}$, where $\mathbf{s}_i$ provides an estimate of how similar x_i is to other objects within a certain neighborhood. Consequently, for

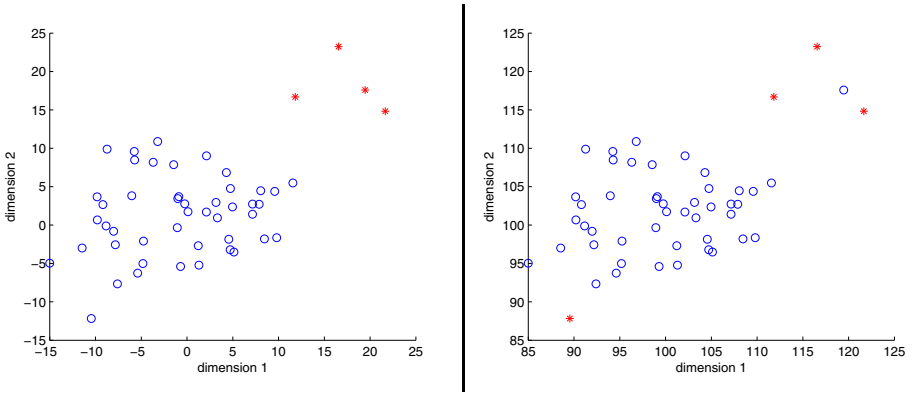

Fig. 1. Outlier detection using k-NN scores with a polynomial kernel of degree 3 and k=3. Detected outliers are denoted by '*'. **Left**: Original data set. **Right**: Original data set translated by (100, 100).

two objects x_i and x_j, if $\mathbf{s}_i < \mathbf{s}_j$ then x_i is more likely to be an outlier than x_j. The vector $\mathbf{s}$ allows ranking objects by their likelihood of being outliers. Those objects having the lowest similarity scores are more likely to be outliers.

In this section we define two specific kernel similarity scores: the k-NN kernel similarity score (kNNSS) and the summation kernel similarity score (SKSS). The k-NN kernel similarity score is a characterization of the well-known k-NN score [5]. It assigns scores to data points based on the sum of their similarities to their k nearest neighbors. To our knowledge, SKSS has not been proposed before in the literature. It uses a density-based approach [12,11], but instead of counting the number of nearest neighbors within a neighborhood of each data point x_i, the SKSS score of x_i is equal to the sum of the similarities between x_i and all data points within a ball of a fixed radius implicitly determined by a parameter p. The definitions for kNNSS and SKSSS follow:

Definition 5. *Let x_i be a point in a data set X with n data points, $K \in \mathbf{R}^{n\times n}$ a similarity kernel matrix defined on X, $K_i \in \mathbf{R}^n$ the i-th row of K, and k a positive integer parameter. The k-NN kernel similarity score of x_i, $kNNSS(x_i, K, k)$, is defined as:*

$$kNNSS(x_i, K, k) = \sum_{j=1}^{k} K_{ij}^s \tag{9}$$

where $K_i^s \in \mathbf{R}^{n-1}$ is the row vector obtained by sorting K_i in descending order without including the diagonal element $K_{i,i}$.

Definition 6. *Let x_i be a point in a data set X with n data points, $K \in \mathbf{R}^{n\times n}$ a similarity kernel matrix defined on X, and $K_i \in \mathbf{R}^n$ the i-th row of K. The summation kernel similarity Score (SKSS) of x_i, $SKSS(x_i, K, p)$, is defined by:*

$$SKSS(x_i, K, p) = \sum_{K_{i,j}^s \geq p} (K_{i,j}^s) \tag{10}$$

where p is a real-valued similarity threshold, and $K_i^s \in \mathbf{R}^{n-1}$ is the row vector obtained from K_i by removing the diagonal element $K_{i,i}$.

Considering that similarity kernels can be normalized in the interval [-1, 1] (or [0,1] when all kernel values are non-negative), it is safe to assume that p lies on that normalized interval.

The computational complexity of calculating both kernel similarity scores for a data set X with n data points is $O(n^2)$, assuming the kernel matrix K is given. Consequently, approximation techniques such as the one mentioned in [8] should be used for applications with large data sets. Because our experiments involved small data sets we did not implement any approximation technique in this work. However, it is worth of noting that kNNSS is slightly more expensive to compute than SKSS when using straightforward implementations. That is because of the sorting step involved in determined the k nearest neighbors for each object.

4 Experimental Comparison of Kernel Similarity Scores

4.1 The Data Sets

We used two numerical data sets and two categorical data sets to compare the performance of kNNSS and SKSS for outlier detection. The data sets were obtained from the UCI Machine Learning Repository [2]. The Gaussian kernel was employed in the numerical cases. The Hamming distance kernel and the diffusion kernel from equation 3 were used in the categorical cases. For some data sets, we randomly removed a substantial amount of samples from one of the classes, in order to comply with the assumption that outliers constitute a small percentage of the data. A brief description of each data set follows.

The yeast data set: A data set consisting of 1484 data points with 8 real-valued attributes. The data points correspond to proteins taken from the SWISS-PROT database. They are classified into ten different classes of localization sites. We chose the 463 proteins with cytosolic or cytoskeletal localization as non-outliers. The 20 proteins corresponding to peroxisomal localization were labeled as outliers.

The breast cancer Winsconsin (diagnostic) data set: It consists of 569 data points with 32 real-valued attributes. It categorizes points as benign or malignant. We used the 357 benign cases from the data set as non-outliers. A random sample of 15 malignant cases were kept as outliers.

The lymphography data set: It contains 148 instances with 18 categorical attributes. Those instances corresponding to the metastases and malign lymph classes form the majority of the data set. The other 6 points, related to the normal and fibrosis categories, were considered outliers.

The post-operative data set: A categorical data set containing 148 instances with 18 attributes. Instances are classified according to where patients should go after surgery: the ICU, home, or the general hospital. The 64 instances where the patients were sent to the general hospital constituted our non-outliers observations. We randomly chose 4 entries from the rest of the data to be outliers.

4.2 The Similarity Kernels

Here we prove that the Hamming distance kernel and the diffusion kernel satisfy our definition of similarity kernels.

From the definition of the diffusion kernel given in section 2.3, it is clear that $K_{KD}(x,x) = 1$ for all $x \in X$. Because all diffusion kernels are positive semi-definite, it is concluded that the diffusion kernel used in this work is a similarity kernel.

The Hamming distance kernel is defined in [6] in the framework of positive semi-definite kernels. It is not immediate from its definition whether it has a constant diagonal. However, the Hamming distance kernel is characterized by the following recursive formulas: [6]

$$K^0(s,t) = 1 \tag{11}$$

$$K^j(s,t) = \left[\lambda^2(\mid D_j \mid -1 - \delta(s_j,t_j)) + (2\lambda - 1)\delta(s_j,t_j) + 1\right] K^{j-1}(s,t) \tag{12}$$

$$K_H(s,t) = K^m(s,t) \tag{13}$$

where $1 \leq j \leq m$ and m is the number of attributes in the data set.

Table 1. Experimental results. The columns labeled as 'q' indicate the number of data points to return as outliers. Values in the other columns represent how many of the q tentative outliers were true outliers. The last row of each table shows the sum of true outliers detected through each score.

	Yeast data set		Breast cancer data set	
	Gaussian kernel		Gaussian kernel	
	$\sigma = 3$	$\sigma = 0.5$	$\sigma = 200$	$\sigma = 50$
	kNNSS	SKSS	kNNSS	SKSS
q	k = 16	p = 0.25	k = 10	p = 0.1
5	3	5	5	5
10	8	9	8	9
15	9	9	11	10
20	11	11	12	12
25	11	11	12	12
sum:	42	45	48	48

	Lymphography data set				Post-operative data set			
	Hamming kernel		Diffusion kernel		Hamming kernel		Diffusion kernel	
	$\lambda = 0.2$	$\lambda = 0.4$	$\beta = 0.3$	$\beta = 0.5$	$\lambda = 0.1$	$\lambda = 0.1$	$\beta = 0.9$	$\beta = 0.3$
	kNNSS	SKSS	kNNSS	SKSS	kNNSS	SKSS	kNNSS	SKSS
q	k = 8	p = 0.2	k = 16	p = 0.05	k = 2	p = 0.05	k = 2	p = 0.35
2	2	2	2	2	1	1	1	1
4	4	4	4	4	1	2	1	1
6	5	6	4	4	1	2	1	2
8	5	6	6	5	1	2	2	3
10	6	6	6	6	2	2	2	3
sum:	22	24	22	21	6	9	7	10

It is easy to see that $K_H(s,s)$ is an algebraic expression that depends only on λ and D_j. Consequently, $K_H(s,s)$ is equal to a constant value regardless of the particular data point s. To show that this constant is positive, we set $K^0(s,s) = 1$ as a base step and prove by induction that if $K^j(s,s) > 0$ then $K^{j+1}(s,s) > 0$ for $1 \leq j \leq m$:

$$K^{j+1}(s,s) = \left(\lambda^2(\mid D_{j+1} \mid -1) + 1\right) K^j(s,s) \tag{14}$$

$$= \left(\lambda^2(\mid D_{j+1} \mid) + (1-\lambda^2)\right) K^j(s,s) > 0 \tag{15}$$

4.3 Experimental Results

A tuning process allowed us to obtain the best values for each kernel and score parameter. The best value of the parameter σ for the Gaussian kernel was obtained from the interval $[0.1, 1000]$ for each similarity score. The best value for the parameters of the categorical kernels were determined, for each similarity score, from the set $\{0.1, 0.2, ..., 0.9\}$. The values for the score parameter k were chosen from the set $\{1, 2, 3, 4, 5, 6, 8, 10, 13, 16\}$. The values of p were restricted to the interval $[0.05, 0.5]$. The experimental results are shown in table 1.

For the breast cancer and the lymphography data sets, both similarity scores performed about equally. For the yeast and post-operative data sets the highest detection rates corresponded to SKSS. Overall, SKSS showed slightly better results than kNNSS.

5 Conclusions

In this paper, the concept "similarity kernel" was introduced by giving a formal definition for it. Similarity kernels should be used for unsupervised nearest neighbor-based outlier detection. For any similarity kernel K, the distance between two points x_i and x_j in the feature space are inversely proportional to $K(x_i, x_j)$. Consequently, the kernel values can be considered as similarity values. Additionally, our definition of similarity kernels satisfy desirable properties of similarity measures. Isotropic stationary kernels should be used for nearest neighbor-based outlier detection whenever possible, in order to maintain invariance to translations and rotations. The class of isotropic stationary kernels is a proper subset of the class of similarity kernels. It would be interesting to determine whether both classes of kernels are the same when constrained to normed spaces.

Two kernel similarity scores were defined in this work: kNNSS and SKSS. The first one is a characterization of the well-known k-NN score in terms of kernels. The second one is a new density-based similarity score. The two scores were compared on four data sets, where SKSS compared favorably to kNNSS. Although these are preliminary results, they suggest that SKSS might be a good alternative to the kNN approach for unsupervised outlier detection. The fact that no sorting procedure is needed to calculate SKSS is another point favoring the use of SKSS.

We believe that the kernel nearest-neighbor approach is an excellent option for domains with a large number of attributes. It could be particularly useful for input spaces with complex data structures for which effective similarity kernels could be defined. However, the values of kernel and score parameters need to be determined in order to obtain good detection accuracy. Consequently, an interesting follow up to this work would be to devise methods for automatic estimation of those parameters. Another interesting path would be to determine which other kernel functions are also similarity kernels.

Acknowledgments. This paper is based upon work/research supported in part by the National Science Foundation under Grant No. 0647120 and Grant No. 0647018. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The lymphography domain was obtained from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklic for providing the data. The first author gratefully acknowledges the advice of Dr. Avelino Gonzalez from the University of Central Florida, and the support of the College of Engineering and Computer Science and the I2Lab at the University of Central Florida.

References

1. Angiulli, F., Pizzuti, C.: Fast outlier detection in high dimensional spaces. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) PKDD 2002. LNCS (LNAI), vol. 2431, pp. 43–78. Springer, Heidelberg (2002)
2. Asuncion, A., Newman, D.: UCI Machine Learning Repository, University of California Irvine, School of Information and Computer Science (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
3. Bay, S., Schwabacher, M.: Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 29–38. ACM Press, New York (2003)
4. Breunig, M., Kriegel, H., Ng, R., Sander, J.: LOF: Identifying density-based local outliers. In: International Conference on Management of Data, pp. 1–12 (2000)
5. Chandola, V., Banerjee, A., Kumar, V.: Anomaly Detection: A Survey. ACM Computing Surveys 41, 15:1–15:58 (2009)
6. Couto, J.: Kernel K-Means for Categorical Data. In: Famili, A.F., Kok, J.N., Peña, J.M., Siebes, A., Feelders, A. (eds.) IDA 2005. LNCS, vol. 3646, pp. 46–56. Springer, Heidelberg (2005)
7. Cristianini, N., Shawe-Taylor, J.: An introduction to support Vector Machines: and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
8. Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S.: A geometric framework for unsupervised anomaly detection. In: Proceedings of the Conference on Applications of Data Mining in Computer Security, pp. 78–100. Kluwer Academics, Dordrecht (2002)
9. Genton, M.G.: Classes of kernels for machine learning: a statistics perspective. Journal of Machine Learning Research 2, 299–312 (2001)

10. Hawkins, D.: Identification of Outliers. Chapman and Hall, Boca Raton (1980)
11. Knorr, E.M., Ng, R.T., Tucakov, V.: Distance-based outliers: algorithms and applications. The VLDB Journal 8(3), 237–253 (2000)
12. Knorr, E.M., Ng, R.T.: Algorithms for Mining Distance-Based Outliers in Large Datasets. In: Proceedings of the 24rd International Conference on Very Large Data Bases, pp. 392–403 (1998)
13. Kondor, R., Lafferty, J.: Diffusion Kernels on Graphs and Other Discrete Structures. In: Proceedings of the 19th International Conference on Machine Learning, pp. 315–322 (2002)
14. 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)
15. Oh, J.H., Gao, J.: A kernel-based approach for detecting outliers of high-dimensional biological data. BMC Bioinformatics 10(Suppl. 4), S7 (2009)
16. Petrovskiy, M.I.: Outlier detection algorithms in data mining systems. Programming and Computer Software 29(4), 228–237 (2003)
17. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 427–438. ACM Press, New York (2000)
18. Roth, V.: Kernel fisher discriminants for outlier detection. Neural computation 18(4), 942–960 (2006)
19. Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. Cambridge University Press, Cambridge (2004)
20. Shen, Y.: Outlier Detection Using the Smallest Kernel Principal Components. PhD dissertation, Department of Statistics, Temple University (2007)
21. Schölkopf, B., Smola, A.J.: Learning with kernels. MIT Press, Cambridge (2002)
22. Wu, M., Jermaine, C.: Outlier detection by sampling with accuracy guarantees. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 767–772 (2006)