ELSEVIER

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition Letters 24 (2003) 3059–3068

Pattern Recognition Letters

www.elsevier.com/locate/patrec

# Detecting pattern-based outliers

Tianming Hu *, Sam Y. Sung

*Department of Computer Science, National University of Singapore, Singapore 117543, Singapore*

Received 11 January 2003; received in revised form 23 June 2003

**Abstract**

Outlier detection targets those exceptional data that deviate from the general pattern. Besides high density clustering, there is another pattern called low density regularity. Thus, there are two types of outliers w.r.t. them. We propose two techniques: one to identify the two patterns and the other to detect the corresponding outliers.

*Keywords:* Outlier detection; Complete spatial randomness; Clustering; Regular spacing

## 1. Introduction

In contrast to traditional pattern recognition that aims to find the general pattern for the majority of data, outlier detection targets the finding of the rare data whose behavior is very exceptional compared to other data. A well-known definition of outlier was given by Hawkins (1980) who defined it as an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism. A similar definition (Barnett and Lewis, 1994) also stated that an outlier is an observation that appears to be inconsistent with the remainder of that set of data. Using the above general definitions, we always imply some pattern w.r.t. which we declare some data points are outliers. This pattern is followed by the global/local majority of the data and is breached by the outliers. In detail, it is embodied by 'other observations' in Hawkins' definition, and by 'the remainder of that set of data' in Barnett and Lewis' definition.

Although outliers are often treated as noise or error in many operations, such as clustering, they may have potential causes and bear useful information that cannot be mined from other data that reside deeply inside clusters. It is not unusual that one man's noise is another man's signal. After identifying possible outliers, we may go further and study the underlying reasons why they happen. This knowledge may be profitable. For instance, outliers may be produced by an incorrect assumption of distribution. In such situations, further investigation for outliers can lead to a more appropriate statistical model, which, in turn, leads to a more appropriate statistical inference.

Given a labeled data set consisting of non-outliers and outliers, the outlier detection is essentially an unsupervised classification problem. In terms of output, current outlier detection techniques can be

* Corresponding author.
*E-mail address:* hutianmi@comp.nus.edu.sg (T. Hu).

0167-8655/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0167-8655(03)00165-X

divided into two categories: hard and soft classifications. Hard classification partitions the data into two crisp sets: non-outliers and outliers. Our approach falls in the soft classification arena that offers a ranking by assigning each datum an outlier factor describing the degree of outlierness. In this paper, we make the following contributions: (1) In addition to the high density pattern clustering, we show there is another pattern called low density regularity, whose corresponding outlier cannot be detected by the local outlier factor (LOF) approach. (2) We propose a technique to identify these two patterns based on the ratio of expected hyper-sphere volume over the observed one. For clustering, it can also tell the minimum cluster size. (3) Based on the sample variance of the hypersphere volume, we develop a technique to detect outliers w.r.t. both high and low density patterns. The distance to the $k$th nearest neighbor is used as the radius of the hyper-sphere and we also offer some heuristic to determine $k$.

The rest of the paper is organized as follows: Related work is reviewed in Section 2. In Section 3, we first show two patterns, high density clustering and low density regularity. Then, under assumption of uniform distribution inside clusters, we propose two techniques: One technique is to identify these two patterns, and the other to detect the corresponding outliers. Experimental evaluation is reported in Section 4 and concluding remarks are given in Section 5.

## 2. Related work

Most outlier detection techniques treat objects with $d$ attributes as points in $\mathfrak{R}^d$ space and these techniques can be divided into two classes based on the dimensionality of data. The first class handles one dimensional data and is mainly developed in the statistics field (Barnett and Lewis, 1994). The other class deals with multi-dimensional data and offers tests based on distance, density etc. These techniques are closely related to the relevant clustering algorithms. In fact, given a clustering algorithm with a function to measure its clustering quality, a naive algorithm for calculating outlier factor can assign each point a value that equals the absolute difference between the original clustering quality and the new clustering quality after removing that point.

Distance-based techniques distinguish potential outliers from others based on the number of points in the neighborhood. They do not assume any prior distribution of the data and limit the counting of points to the neighborhood of each point. Corresponding to clustering algorithms that find convex clusters (Ng and Han, 2002), one such technique is the well-known DB($p, d$)-outlier proposed by Knorr et al. (2000). It defines a point in dataset $T$ is an outlier if at least $p$ fraction of points in $T$ lie greater than distance $d$ from it. The strength of this definition includes simplicity and capture of the basic meaning of Hawkins' definition; however, it cannot handle data with different local patterns (densities). Sometimes, a point could be outlying only in a subspace. Corresponding to clustering algorithms capable of finding clusters in subspace (Agrawal et al., 1998), Aggarwal and Yu (2001) searched all subspaces for low density regions and all points in such regions are declared outliers.

Because we mainly compare our approach against LOF, we introduce it here in some detail. Corresponding to clustering algorithms capable of finding arbitrary shape clusters (Ester et al., 1996), Breunig et al. (2000) proposed the notion of LOF, which measures the degree of outlierness based on the difference in the local density of a point and its $k$ nearest neighbors. Generally speaking, DB-outlier can only find global outliers that lie far away from all spherical clusters. It cannot detect local outliers w.r.t. a neighboring dense cluster in the presence of another very sparse cluster. LOF solves this problem by thinking locally, i.e., comparing local density of the outlier only with those of its neighbors. For cluster points, their density will be similar and their LOF will be close to one. For an outlier outside the cluster, its local density will be lower than those of its neighbors inside the cluster and its LOF will be more than one. The weakness of LOF is that it cannot detect the outlier if its local density is higher, not lower, than those inside the neighboring pattern. Such a pattern may consist of a set of regularly spaced points whose densities are lower than that of their neighboring outlier. The introduction of the outlier significantly breaks the regularity and increases the local densities.

## 3. Pattern-based outliers

### 3.1. Patterns based on complete spatial randomness

According to Webster's dictionary, a pattern is 'a natural or chance configuration, or a reliable sample of traits, acts, tendencies, or other observable characteristics'. Complete spatial randomness (csr) (Cressie, 1993) refers to a lack of structure (pattern) in the spatial point process, where events (points regarded as a realization of events) are uniformly distributed in the study region $A \subset \Re^d$. For any sub-region $B \subset A$, the probability that there is at least one event within it is equal to the ratio of its volume over the total volume, i.e., $|B|/|A|$, where $|\cdot|$ denotes volume. This probability is independent from $B$'s location and shape. This kind of spatial point process is also called a homogeneous Poisson process because $N(B)$, the number of events in $B$, follows a Poisson distribution with mean $\lambda|B|$, where $\lambda$ denotes the constant intensity in the study region. On the other hand, an inhomogeneous Poisson process means the intensity is non-stationary and becomes a function of location. In that case, the intensity can be estimated via various kernel methods (Diggle, 1985).

A particular realization of a homogeneous Poisson process with $N(B) = 100$ and $|B| = 10 \times 10$ is given in Fig. 1(a). At least two patterns exist based on csr: clustering and regularity. A cluster with arbitrary shape can be defined as a set of points with similar densities that are significantly higher than those of points in its immediate surrounding area. Both homogeneous and inhomogeneous Poisson processes have been used for cluster analysis in classification of remote sensing images (Rasson, 1993; Rasson and Granville, 1995). Two clusters $C_1$ and $C_2$ are shown in the

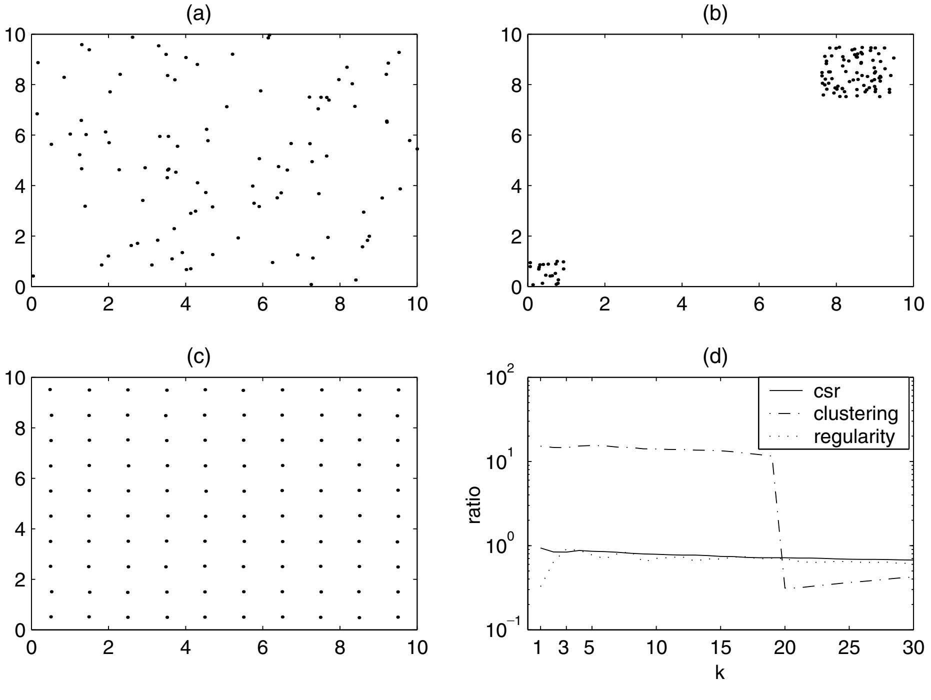

Fig. 1. (a) Complete spatial randomness (csr), (b) clustering with two clusters, (c) regularity with a small Gaussian disturbance, (d) ratio vs. $k$.

lower left and upper right corners of Fig. 1(b), where $C_1$ has 20 points uniformly distributed in a $1 \times 1$ area and $C_2$ has 80 points uniformly distributed in a $2 \times 2$ area. Compared to csr, clustering means that points tend to attract one another, and consequently the average nearest neighbor distance is smaller than that of csr. Fig. 1(c) illustrates 100 points regularly spaced at one interval in both horizontal and vertical directions, with a small Gaussian disturbance. Despite the Gaussian noise, the difference between it and csr in Fig. 1(a) is still obvious. In a way, regular spacing can be regarded as a special clustering in that the points are distributed so uniformly that it shows too little randomness. Compared to csr, regularity means that points tend to push one another. As a result, the nearest neighbor distance is approximately the same for all points and is larger than its counterpart in csr. Besides, for each point and some small $j$ (e.g., $j = 4$ in Fig. 1(c)), its $k$th ($k \leqslant j$) nearest neighbor distances are usually also the same.

### 3.2. Identifying clustering and regularity

Let $V_k$ denote the random variable of the hypersphere volume centered at a randomly chosen point in the study area $B \subset \Re^d$, with radius equal to $R_k$, its distance to the $k$th nearest neighboring object. Note that it does not matter whether there is an event (object) happening at that point location. By assuming the distribution of the objects is csr with constant intensity $\lambda$, we note the random variable $V_k$ is actually from a gamma distribution with parameter $(k, \lambda)$ (Ross, 1998). If the randomly chosen point above is replaced by a randomly chosen object, the distribution of the corresponding random variable remains the same, with expectation $E(V_k) = k/\lambda$ and variance $\mathrm{Var}(V_k) = k/(\lambda^2)$.

Based on the expectation, we propose a technique to identify the data structure by telling us whether it is csr, clustering, or regularity. Furthermore, in the case of clustering with csr inside each cluster, it can tell the minimum cluster size. Given a dataset $\{x_i \in \Re^d\}_{i=1}^n$, after collecting the volume $V_k = \pi^{d/2} R_k^d / \Gamma(1 + d/2)$ for each datum and estimating the total intensity $\lambda$, we can compute the ratio of the expectation of $V_k$ over the observed one, as in Eq. (1):

$$R(k) \equiv \frac{k/\lambda}{\frac{1}{n}\sum_{i=1}^{n} V_k}. \tag{1}$$

We can draw a plot of $R(k)$ vs. $k$ and identify the structure based on the following three properties:

(1) If $R(k)$ is close to 1 at all $k$s, the data structure is csr.
(2) If $R(k)$ is significantly less than 1 at small $k$, e.g., $k = 1, 2$, the pattern is regularity. Because the nearest neighbor distance of regularity is larger than csr, such relation also holds for the volume.
(3) If $R(k)$ is significantly greater than 1 at many $k$s, especially at small ones, the pattern is clustering. The reason is that its nearest neighbor distances are smaller than those in csr, which also leads to smaller volume. Besides, if there are multiple clusters, $R(k)$ will initially remain nearly constant, and drop sharply when $k$ reaches the minimum cluster size.

The ratio for three datasets in Fig. 1(a)–(c) is illustrated in Fig. 1(d) with $\hat{\lambda} = 100/(10 \times 10)$. As expected, $R(k)$ for csr in Fig. 1(a) is close to 1 for all $k$s. For regularity in Fig. 1(c), $R(k)$ is significantly smaller than 1 at $k = 1, 2$ and close to 1 at $k = 3$, which means under csr, the average distance to the third nearest neighbor is approximately 1. For clustering in Fig. 1(b), $R(k)$'s curve is relatively flat as $k < 20$, and drops radically at $k = 20$, the smaller cluster $C_1$'s size. The reason is that at $k = 20$, the 20th nearest neighbor of every point in $C_1$ is in $C_2$, which means their $V_k$ no longer follows a gamma distribution with parameter ($k = 20$, $\lambda = 20/(1 \times 1)$). Generally, suppose the dataset consists of $m$ disjointed clusters $\{C_j(n_j, \lambda_j)\}_{j=1}^m$, where $n_j$ and $\lambda_j$ denote the $j$th cluster size and intensity and $n_1 \leqslant \cdots \leqslant n_m$. Under the assumption of csr inside every cluster, we can approximate the sample mean of $V_k$, the denominator in Eq. (1), with the denominator in Eq. (2), i.e., replace the sum of $V_k$ in every cluster with the expected value. Consequently, $R(k)$ in Eq. (2) is independent of $k$ and remains constant till the replacement is no longer valid at $k = n_1$, when the $k$th nearest neighbor of every point in $C_1$ is no longer in $C_1$ and the corresponding $V_k$ no longer follows the gamma distribution with parameters $(k, \lambda_1)$.

$$R(k) \approx \frac{k/\lambda}{\frac{1}{n}\sum_{j=1}^{m}\frac{n_j k}{\lambda_j}} \tag{2}$$

### 3.3. Detecting pattern-based outliers

A data point could be outlying w.r.t. a nearby high density pattern cluster because its own density is relatively low. This case is shown in Fig. 2(a), where there are two clusters, one dense $C_1$ and a sparse one $C_2$. Densities illustrated in Fig. 2(b) are obtained with a Gaussian kernel function

$$f(x) = \sum_{i=1}^{n} \exp(-d^2(x, x_i)/(2\sigma^2)),$$

where $\sigma = 1$ and $d(x, y)$ denotes the Euclidean distance between $x$ and $y$. Point $O_2$ is a global outlier because its density is lower than both clusters and it can be detected by both DB-outlier and LOF. Point $O_1$ is a local outlier w.r.t. $C_1$, for its density is lower than $C_1$ but comparable to $C_2$. Only LOF can detect it, as shown in Fig. 2(c). On the other hand, a data point could also be outlying w.r.t. a nearby low density pattern regularity because its own density is higher than neighboring points belonging to the regularity. This situation is shown in Fig. 2(d), where two outliers, $O_1$ and $O_2$, have densities higher than most points in the pattern, a 3 × 3 grid, as demonstrated in Fig. 2(e) with the same kernel function. Fig. 2(f) proves that LOF cannot detect them by making their outlier factors simultaneously higher than those of all regularity points. In fact, $R_2$'s LOF is consistently higher than that of both at all $k$s except 2 where $R_1$ takes the lead.

Combining the two situations, we can conclude that a point may be outlying because its density is

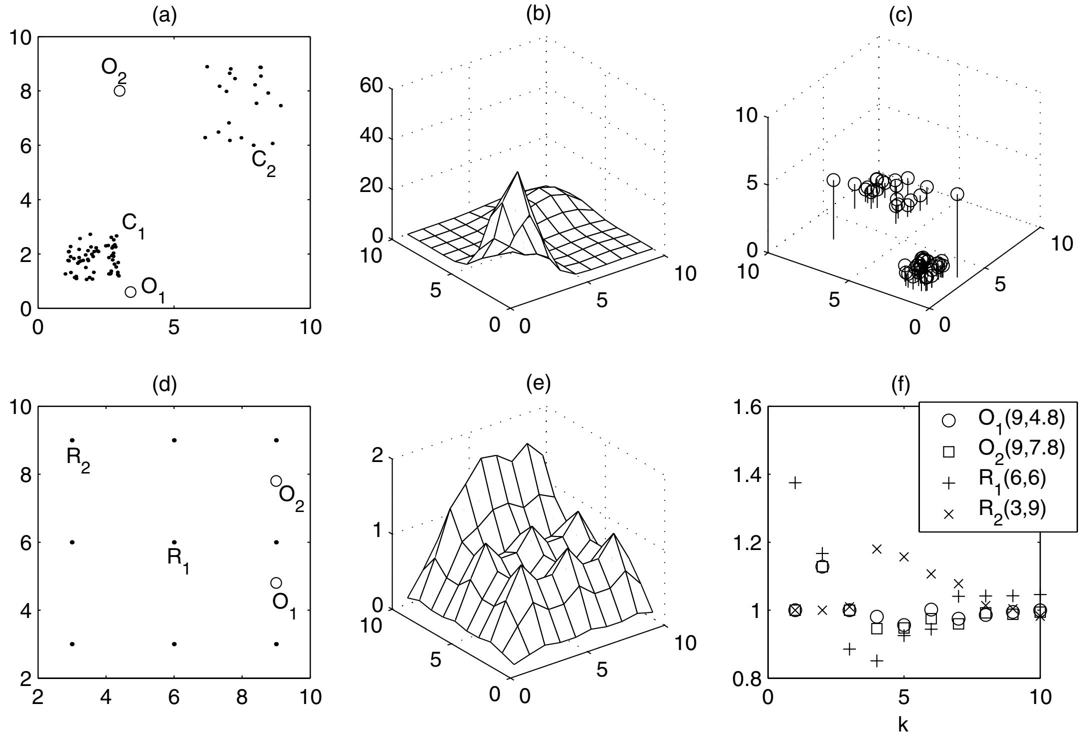

Fig. 2. (a) Cluster-based global/local outliers, (b) density of clustering, (c) LOF with $k = 2$, (d) regularity-based outliers, (e) density of regularity, (f) LOF with $k = 1, \ldots, 10$.

lower (higher) than a nearby high (low) density pattern. In other words, it is outlying because its density is significantly different from that of its neighbors belonging to the pattern. Thus, the sample variance of $R_k$ together with $V_k$ is expected to be high. This observation leads to our approach for detecting local outliers based on variance of volume (VOV). For a dataset $X = \{x_i\}_{i=1}^{n}$, if we denote $x_i$'s $k$th nearest neighbor distance by $d_k(x_i)$ and its $k$th order neighborhood by $N_k(x_i) \equiv \{x : x \in X - \{x_i\}, d(x_i, x) \leqslant d_k(x_i)\}$, our local outlier factor VOV can be computed as follows: first, for each data point $x_i$, retrieve its augmented $k$th neighborhood $N_k^+(x_i) \equiv x_i \cup N_k(x_i)$ and compute $V_k$. Then, compute the sample variance of $V_k$ over $N_k^+(x_i)$ and assign it as the VOV outlier factor to $x_i$. The resulting definition of VOV is given in Eq. (3), where

$$\overline{V}_k(x_i) \equiv \sum_{x \in N_k^+(x_i)} V_k(x)/|N_k^+(x_i)|,$$

$$\mathrm{VOV}(x_i) \equiv S^2(x_i) = \frac{\sum_{x \in N_k^+(x_i)} (V_k(x) - \overline{V}_k(x_i))^2}{|N_k^+(x_i)| - 1}. \qquad (3)$$

### 3.4. Properties of VOV

The sample variance $S^2$ is itself a random variable. For data belonging to the pattern, it is preferred that $E(S^2)$ be smaller than those of outliers. Besides, $\mathrm{Var}(S^2)$ is also preferred small, which is achieved by using reachability distance instead of pure distance in LOF. If the pattern is regularity, for some appropriately chosen small $k$, VOV is 0 for pattern points. If the pattern is clustering, for simplicity, we assume $|N_k^+(x_i)| = k + 1$. In that case, for cluster (csr inside with intensity $\lambda$) points, $E(S^2) = k/\lambda^2$. If $k$ is relatively large, gamma distribution can be approximated by Gaussian distribution and $\lambda^2 S^2$ follows a chi-squared distribution $\chi_k^2$ with $k$ degrees of freedom (Ross, 1998), so $\mathrm{Var}(S^2) \approx 2k/\lambda^4$.

From $S^2$'s expectation and variance, we can see that $k$ cannot be large. On the other hand, $k$ cannot be too small. Suppose there are two outliers closest to each other, then their VOV are both 0 at $k = 1$. A method to choose $k$ is to use the figure of ratio vs. $k$ in Eq. (1). Based on that figure, we can find the minimum cluster size and set $k$ at a value a little less than the minimum cluster size, but still larger than the outlier cluster size, if multiple outliers really lie together. At that $k$, cluster points' $k$th nearest neighbors are still in the same cluster and hence $V_k$ still follows a gamma distribution. For outliers, their $k$th nearest neighbors are expected to lie in the nearby clusters and $V_k$ does not follow a gamma distribution; otherwise, those outliers themselves form a cluster of size $k + 1$ and it is not reasonable to regard them as outliers.

At times, a point inside csr may look outlying locally and it is interesting to see how VOV works in that case. Because VOV only offers a ranking and, in general, it is impossible to determine an optimal cutoff value to partition the data, we need other tests to check if some of those top VOV points are false outliers. For example, we can collect $V_k$ in their neighborhoods and perform Kolmogorov–Smirnov tests with null hypothesis that $V_k$ follows a gamma distribution. Under the assumption that gamma distribution can be approximated by Gaussian distribution at large $k$, we propose a simple test by computing their absolute $z$-score, $|Z| \equiv |(V_k(x_i) - \overline{V}_k(x_i))|/\sqrt{S^2(x_i)}$, where the mean and variance are over $N_k(x_i)$ to exclude the possible outlier $x_i$. For $X \sim N(\mu, \sigma^2)$, $(X - \mu)/\sigma \sim N(0, 1)$. With significance level $\alpha = 0.05$, if $|Z| > 1.96$ $(X \sim N(0, 1), P(|X| > 1.96) = 0.05)$, we reject the hypothesis and keep the point in the candidate outlier set; otherwise, we accept the hypothesis and filter out the point. Because cluster and regularity-based outliers have extreme $V_k$, they are unlikely to be filtered out. During a simulation, 100 samples are randomly drawn, each consisting of 100 points uniformly distributed in a $10 \times 10$ area. From each sample, top $p$ points with the highest VOV are selected as candidate outliers. Fig. 3 illustrates the average of the fraction of candidate outliers that have an absolute $z$-score greater than 1.96. Clearly, the fraction decreases with $k$, which suggests that in the case of csr, a large $k$ is preferred to decrease false detection probability.

As for time complexity, VOV is similar to LOF and takes $\mathrm{O}(n \times (k\mathrm{NN} + k))$ time, where $k$NN

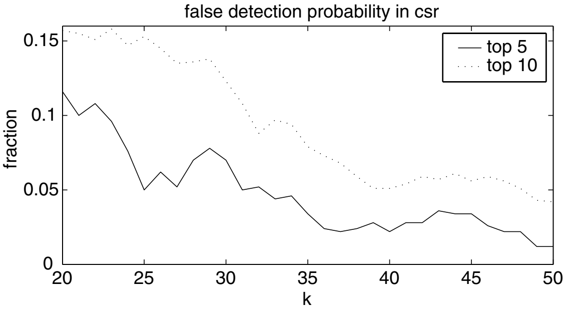

Fig. 3. In csr, the fraction of top $p$ VOV outliers that have an absolute $z$-score greater than 1.96.

denotes the time for a $k$ nearest neighbors query. The dominant part, $\mathrm{O}(n \times k\mathrm{NN})$, is spent in collecting $V_k$ and it depends on the particular implementation of $k$ nearest neighbors query. The remaining part $\mathrm{O}(nk)$ is used for computing the sample variance of $V_k$.

## 4. Experimental evaluation

### 4.1. Evaluating outlier detection approaches

The criteria evaluating outlier detection approaches can be divided into two parts: efficiency and effectiveness. Good efficiency means the technique should be applicable not only to small databases of just a few thousand objects, but also to larger databases with millions of objects. As for effectiveness, considering that the final user is human, a good approach should require as few input parameters from the user as possible and these parameters should have intuitive meaning (such as $k$).

At this time, we should discuss some formal criteria. Given a labeled dataset $D = D_\mathrm{O} \cup D_\mathrm{N}$ partitioned into outliers $D_\mathrm{O}$ and non-outliers $D_\mathrm{N}$, for any outlier detection method $M(\theta)$ ($\theta$ denotes its parameters), we say $M(\theta)$ is consistent with $D$ if we can find some particular estimate $\hat{\theta}$ such that $M(\hat{\theta})$ can correctly partition $D$. Hence, LOF is not consistent with the dataset in Fig. 2(d). Apparently, we prefer a method $M$ that is consistent with more labeled datasets. Many other concepts in computational learning theory can also be applied here. For instance, $M(\theta)$ is said to shatter an unlabeled dataset $D$ if for any binary partition of $D$, we can always find some $\hat{\theta}$ such that $M(\hat{\theta})$ is consistent with that partition. Thus, we can define $M$'s VC-dimension as the maximum size of $D$ that can be shattered by $M$. However, high VC-dimension is not always preferred, for among the $2^{|D|}$ partitions, many are unreasonable, e.g., $|D_\mathrm{O}| = |D| - 1$. So, a practical requirement for $M$ may be that it can detect a finite number of fixed patterns and allow the user to specify the patterns on which he/she hopes the detected outliers will be based.

### 4.2. Synthetic data

A dataset is illustrated in Fig. 4(a) with a cluster in the top right corner and a regularity in the bottom left corner. In addition, there are three outliers, including a global outlier $O_1$, a cluster-based

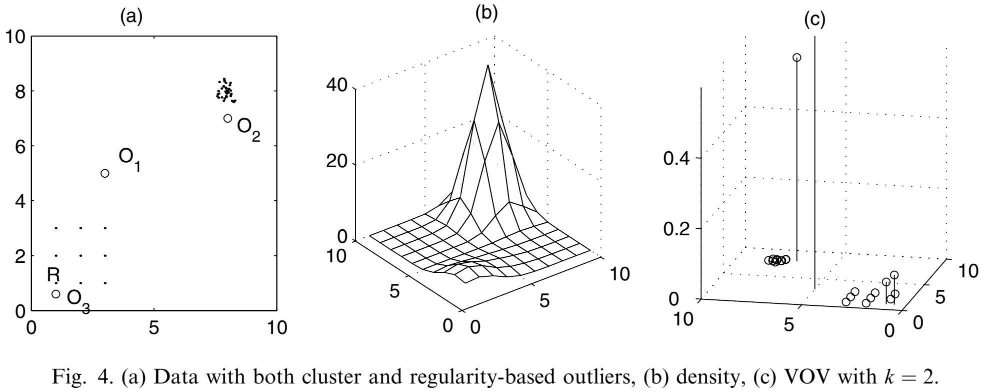

Fig. 4. (a) Data with both cluster and regularity-based outliers, (b) density, (c) VOV with $k = 2$.

local outlier $O_2$, and a regularity-based local outlier $O_3$. The density with Gaussian kernel is shown in Fig. 4(b) and the VOV outlier factors are shown in Fig. 4(c) with $k = 2$. We can see that VOV successfully separates three outliers from pattern points and is consistent with this labeled dataset. Pattern point $R(1,1)$ has the largest VOV over pattern points and this is reasonable, for its density is greatly increased by the presence of the neighboring outlier $O_3$.

### 4.3. Real data

One way to evaluate the performance of outlier detection techniques on real data is to use them to discover the data from rare classes (Aggarwal and Yu, 2001). We choose from the UCI repository (Murphy and Aha, 1994) three binary class datasets: ionosphere, Wisconsin diagnostic breast cancer, and Pima Indian diabetes. All data from the majority classes are treated as non-outliers. To make the ratio of non-outliers over outliers 9:1 in each of three resulting datasets, we select 25 data from the class of bad radar returns in the ionosphere dataset, 40 data from the malignant cancer class in the cancer dataset, and 56 data from the tested-positive class in the diabetes dataset.

First, we draw the figure of ratio vs. $k$ in Fig. 5(a), (d) and (g). Compared to the corresponding csr with the same bounding region, we can see that the ratio is far lower than 1, i.e., the average $k$ nearest neighbor distances are much larger than those under csr. For ionosphere data, the maximum ratio is achieved at $k = 7$. For the other two, the ratio keeps decreasing. We choose $k = 3, 7$ for subsequent comparisons.

After choosing $k$, both VOV and LOF provide a ranking of data in decreasing order of outlier factors. We can choose top $100p\%$ data $T(p)$, compare them to the true outliers $O$ (those 10%)

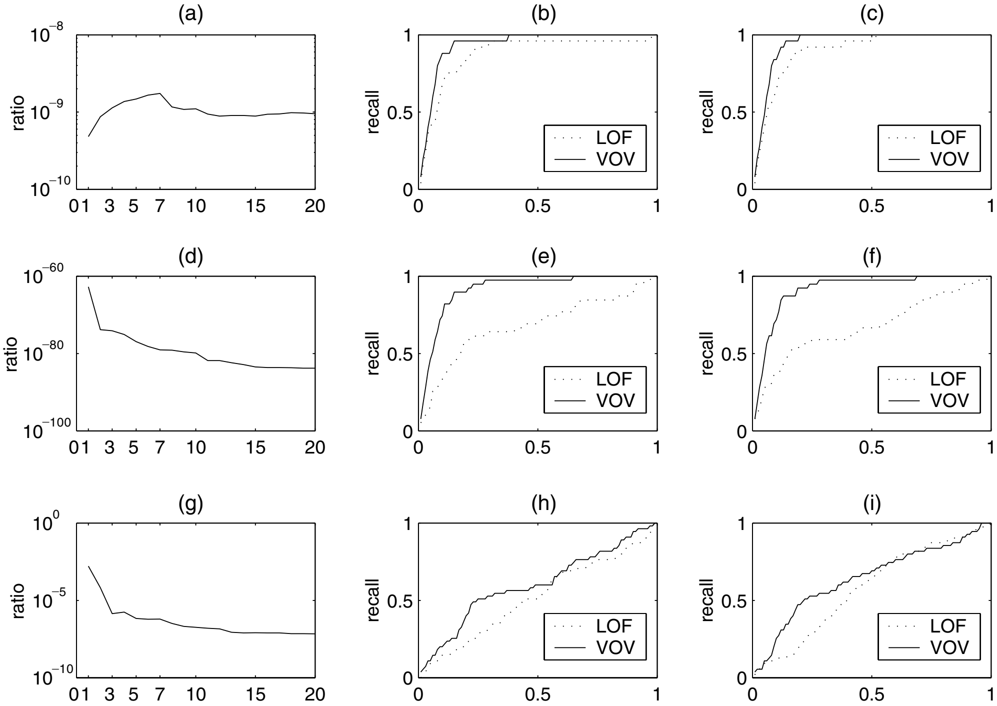

Fig. 5. (a) Ratio for ionosphere data, (b) LOF vs. VOV with $k = 3$ for ionosphere data, (c) LOF vs. VOV with $k = 7$ for ionosphere data, (d) ratio for cancer data. (e) LOF vs. VOV with $k = 3$ for cancer data, (f) LOF vs. VOV with $k = 7$ for cancer data, (g) ratio for diabetes data, (h) LOF vs. VOV with $k = 3$ for diabetes data, (i) LOF vs. VOV with $k = 7$ for diabetes data.

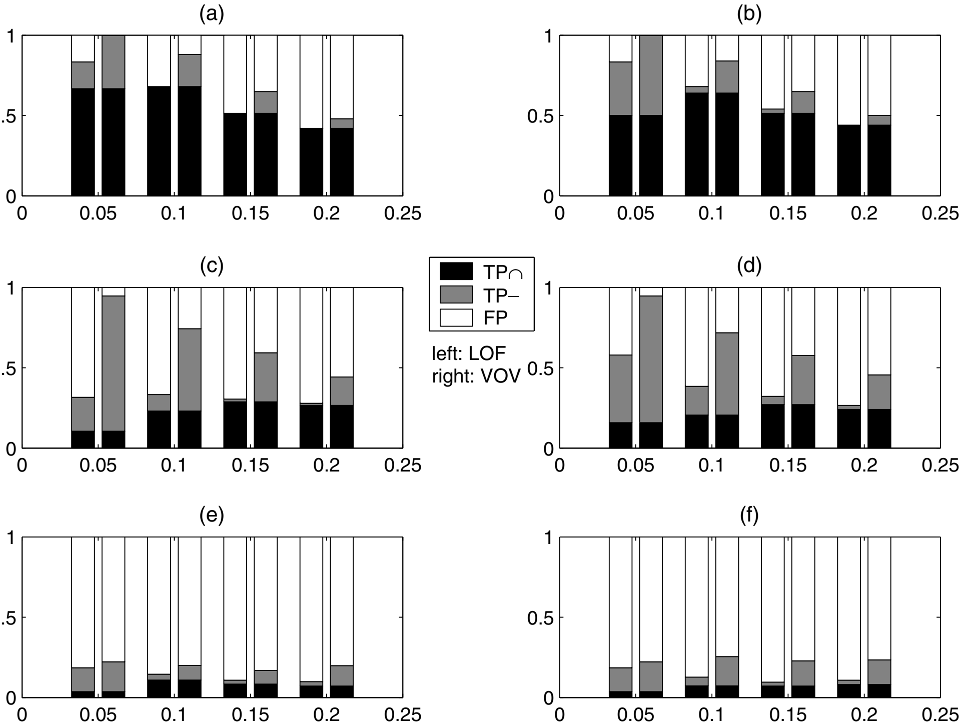

Fig. 6. Comparison of makeup of prediction by LOF (left bar) and VOV (right bar). TP∩ denotes true positive intersection. TP− denotes true positive difference. FP denotes false positive. Ionosphere data: (a) $k = 3$, (b) $k = 7$; Cancer data: (c) $k = 3$, (d) $k = 7$; Diabetes data: (e) $k = 3$, (f) $k = 7$.

by computing recall $|T(p) \cap O|/|O|$ and precision $|T(p) \cap O|/|T(p)|$. In this case, a larger recall also means a larger precision and we illustrate recall in Fig. 5(b,c), (e,f) and (h,i). In particular, we concentrate on two aspects. One aspect is recall at small $p$s, for it is the common practice that we select some top predicted outliers for further investigation. Furthermore, the smaller $p$ is, the more important the corresponding recall. The other aspect is the minimum of $p$ at which VOV and LOF achieve full recall. From these two aspects, we can see that VOV is consistently and significantly better than LOF on ionosphere and cancer data, which implies these two datasets coincide with our definition of outliers. As for diabetes data, our assumption is probably no longer valid; however, VOV is still much better than LOF on the recall at small $p$s. VOV is consistently better than LOF at $k = 3$ and is slightly overtaken by LOF at $p \in [0.6, 0.8]$ with $k = 7$.

To further analyze the prediction set, we divide $T(p)$ into three subsets: intersection of true positive $(T(p) \cap O)$ between LOF and VOV, difference of true positive, and false positive $(T(p) - O)$. Roughly speaking, true positive intersection includes those cluster-based outliers that both LOF and VOV are able to detect. True positive difference of VOV can be interpreted by those regularity-based outliers that LOF fails to detect. The fraction of these three subsets at four $p$s is shown in Fig. 6. We can see that at all $p$s LOF fails to capture some true outliers discovered by VOV. As $p$ increases to 0.15, however, almost all true outliers predicted by LOF are covered by VOV.

## 5. Conclusion

In this paper, we first illustrated that there are at least two patterns: high density clustering and low density regularity. Therefore, there are two kinds of outliers w.r.t. them. Under assumption of csr inside clusters, we proposed a technique to identify them, based on the volume of the sphere centered at each data point with radius equal to its *k*th nearest neighbor distance. Also based on the sample variance of this random variable, we developed a VOV approach to detecting outliers. Experimental results show VOV can simultaneously detect outliers w.r.t. both types of patterns and is better than LOF on three real datasets from the UCI repository, especially at small *k*s.

## References

Aggarwal, C.C., Yu, P.S., 2001. Outlier detection for high dimensional data. In: Proc. 2001 ACM SIGMOD Internat. Conf. on Management of Data. pp. 37–46.

Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., 1998. Automatic subspace clustering of high dimensional data for data mining applications. In: Proc. 1998 ACM SIGMOD Internat. Conf. on Management of Data. pp. 94–105.

Barnett, V., Lewis, T., 1994. Outliers in Statistical Data, third ed. John Wiley and Sons.

Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. LOF: Identifying density-based local outliers. In: Proc. 2000 ACM SIGMOD Internat. Conf. on Management of Data. pp. 93–104.

Cressie, N.A., 1993. Statistics for Spatial Data. Wiley.

Diggle, P.J., 1985. A kernel method for smoothing point process data. Appl. Statist. 34, 138–147.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining. pp. 226–231.

Hawkins, D.M., 1980. Identification of Outliers. Chapman and Hall.

Knorr, E.M., Ng, R.T., Tucakov, V., 2000. Distance-based outliers: Algorithms and applications. Very Large Data Bases J. 8 (3), 237–253.

Murphy, P.M., Aha, D.W., 1994. UCI Repository of Machine Learning Databases. Department of Information and Computer Science, University of California at Irvine, http://www.ics.uci.edu/mlearn/MLRepository.html.

Ng, R., Han, J., 2002. CLARANS: A method for clustering objects for spatial data mining. IEEE Trans. Knowledge Data Eng. 14 (5), 1003–1016.

Rasson, J.P., 1993. The non-stationary Poisson process: A fully non-parametric model for a new approach of supervised classification. Bull. Internat. Statist. Instit. 49 (2), 335.

Rasson, J.P., Granville, V., 1995. Multivariate discriminant analysis and maximum penalized likelihood density estimation. J. Roy. Statist. Soc. B (57), 501–517.

Ross, S., 1998. A First Course in Probability, fifth ed. Prentice Hall.