13th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD 2009), Bangkok, Thailand, 2009

Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data

Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek

Ludwig-Maximilians-Universität München
Oettingenstr. 67, 80538 München, Germany
http://www.dbs.ifi.lmu.de
{kriegel,kroegerp,schube,zimek}@dbs.ifi.lmu.de

Abstract. We propose an original outlier detection schema that detects outliers in varying subspaces of a high dimensional feature space. In particular, for each object in the data set, we explore the axis-parallel subspace spanned by its neighbors and determine how much the object deviates from the neighbors in this subspace. In our experiments, we show that our novel subspace outlier detection is superior to existing full-dimensional approaches and scales well to high dimensional databases.

1 Introduction

Outlier detection aims at finding the "different mechanism" [1], i.e., detecting outliers that do not fit well to the mechanisms that generate most of the data objects. All existing approaches somehow rely on the full-dimensional Euclidean data space in order to examine the properties of each data object to detect outliers. However, today's applications are characterized by producing high dimensional data. In general, mining these high dimensional data sets is imprecated with the *curse of dimensionality*. For outlier detection, two specific aspects are most important. First, in high dimensional spaces Euclidean distances (and other L_p-norms) can no longer be used to differentiate between points. All points are more or less equi-distant to each other (see e.g. [2]). As a consequence, no particular outlier can be detected that deviates considerably from the majority of points. Second, we may have still concrete mechanisms that have generated the data but, usually, for each of these generating mechanisms only a subset of features may be relevant (this problem is known as *local feature relevance* [3]). In addition, these subsets of relevant features may be different for different generating mechanisms. As a consequence, outlier detection makes sense only when considering the subsets of relevant features of these generating mechanisms, i.e. subspaces of the original feature space. Figure 1(a) illustrates the general idea of finding outliers in subspaces. Point o is not an outlier in the full (two) dimensional space because it does not deviate considerably from its neighboring points (indicated by crosses). Since the density among o and its neighbors in the two dimensional feature space is rather uniform, o will also not be recognized as an outlier by any existing full dimensional outlier detection method. However, when projected on the axis A_1, point o is an outlier because it deviates

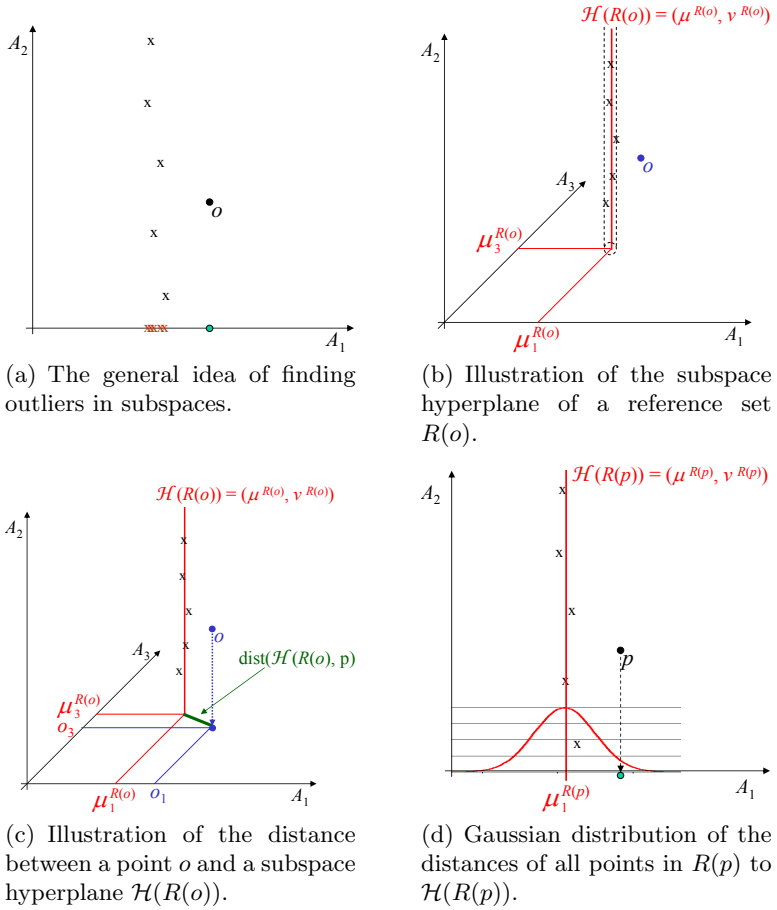
considerably from the neighboring points. Apparently, the points indicated by crosses have been generated by a mechanism where a low variance around a certain value in attribute A_1 is characteristic while the values of attribute A_2 are uniformly distributed and obviously not characteristic for the given mechanism. Finding outliers in subspaces is particularly interesting in high dimensional data where we can expect a rather uniform distribution in the full dimensional space but interesting distributions (including outliers) in subspaces. Since these subspaces of relevant features are usually not known beforehand (outlier detection is an unsupervised task), the search for outliers must be coupled with the search for the relevant subspaces. In this paper, we present a novel outlier detection schema that searches for outliers in subspaces of the original data. Our method is particularly useful for high dimensional data where outliers cannot be found in the entire feature space but in different subspaces of the original space. The remainder is organized as follows. We review related work in Section 2. Our novel subspace outlier model is described in Section 3. An experimental evaluation is presented in Section 4. Section 5 provides conclusions.

2 Related Work

Existing approaches for outlier detection can be classified as global or local outlier models. A global outlier approach is based on differences of properties compared over the complete data set and usually models outlierness as a binary property: for each object it is decided whether it is an outlier or not. A local outlier approach rather considers a selection of the data set and usually computes a degree of outlierness: for each object a value is computed that specifies "how much" this object is an outlier w.r.t. the selected part of the data. Here, we focus on this second family of approaches. The first approach to overcome the limitations of a global view on outlierness has been the density-based local outlier factor (LOF) [4]. The LOF compares the density of each object o of a data set $\mathcal{D}$ with the density of the k-nearest neighbors of o. A LOF value of approximately 1 indicates that the corresponding object is located within a cluster, i.e. a region of homogeneous density. The higher the difference of the density around o is compared to the density around the k-nearest neighbors of o, the higher is the LOF value that is assigned to o. The outlier score ABOD [5] claims to be tailored to meet the difficulties in high dimensional data because it is not primarily based on conventional distance measures but assesses the variance in angles between an outlier candidate and all other pairs of points. Nevertheless, the special problem of irrelevant attributes in high dimensional data is not addressed by ABOD.

3 Outlier Detection in Axis-Parallel Subspaces

The general idea of our novel subspace outlier model is to analyze for each point, how well it fits to the subspace that is spanned by a set of reference points. The subspace spanned by a set of points is simply an axis-parallel hyperplane of any dimensionality $l < d$, where d is the dimensionality of the entire feature space,



(a) The general idea of finding outliers in subspaces.

(b) Illustration of the subspace hyperplane of a reference set $R(o)$.

(c) Illustration of the distance between a point o and a subspace hyperplane $\mathcal{H}(R(o))$.

(d) Gaussian distribution of the distances of all points in $R(p)$ to $\mathcal{H}(R(p))$.

Fig. 1. Illustration of basic concepts.

such that all points of the reference set are close to this hyperplane. If a point deviates significantly from this reference hyperplane, it is considered to be an outlier in the subspace that is perpendicular to that hyperplane.

In the following, we assume that $\mathcal{D} \subseteq \mathbb{R}^d$ is a database of n points in a d-dimensional feature space and *dist* is a metric distance function on the points in $\mathcal{D}$, e.g. one of the L_p-norms or the cosine distance. For any point $p \in \mathbb{R}^d$, we denote the projection of p onto attribute i by p_i.

Intuitively, the subspace hyperplane of a set of points S (the reference set) captures the subspace in which the variance of the points in S is high, whereas in the perpendicular subspace, the variance of the points in S is low. The *variance* $VAR^S \in \mathbb{R}$ of S is the average squared distance of the points in S to the mean value μ^S, i.e., $VAR^S = \frac{\sum_{p \in S} dist(p,\mu^S)^2}{Card(S)}$, where $Card(S)$ denotes the cardinality

of the set S. Analogously, the *variance along an attribute* i, denoted by $var_i^S \in \mathbb{R}$ of S is defined as $var_i^S = \frac{\sum_{p \in S} \left(dist\left(p_i, \mu_i^S\right)\right)^2}{Card(S)}$.

Let $R(p) \subseteq \mathcal{D}$ be a set of reference points for $p \in \mathcal{D}$, called *reference set* w.r.t. which the outlierness of p should be evaluated. The *subspace defining vector* $v^{R(p)} \in \mathbb{R}^d$ of a reference set $R(p)$ specifies the relevant attributes of the subspace defined by the set $R(p)$, i.e. the attributes where the points in $R(p)$ exhibit a low variance. Thereby, we differentiate between high and low variance as follows. In all d attributes, the points have a total variance of $VAR^{R(p)}$. Thus, the expected variance along the i-th attribute is $1/d \cdot VAR^{R(p)}$. We evaluate the variance of the points along the i-th attribute as *low* if $var_i^{R(p)}$ is smaller than the expected variance by a predefined coefficient α. For each attribute in which $R(p)$ exhibits a low variance, the corresponding value of the subspace defining vector $v^{R(p)}$ is set to 1, for the remaining attributes to 0. Formally,

$$v_i^{R(p)} = \begin{cases} 1 & \text{if} \quad var_i^{R(p)} < \alpha \frac{VAR^{R(p)}}{d} \\ 0 & \text{else.} \end{cases} \tag{1}$$

The *subspace hyperplane* $\mathcal{H}(R(p))$ of $R(p)$ is defined by a tuple of the mean value $\mu^{R(p)}$ of $R(p)$ and the subspace defining vector $v^{R(p)}$ of $R(p)$, i.e. $\mathcal{H}(R(p)) = (\mu^{R(p)}, v^{R(p)})$. Figure 1(b) illustrates a subspace hyperplane for a sample reference set $R(o)$ (indicated by crosses) of a point o (indicated by a dot) in a three dimensional feature space. The points of $R(o)$ form a line in the three dimensional space. Thus, the subspace defining vector of $R(o)$ is defined as $v^{R(o)} = (1, 0, 1)^\intercal$, because attribute A_1 and A_3 are relevant and attribute A_2 is not, i.e. the variance along A_1 and A_3 is small whereas it is high along A_2. The subspace hyperplane of $R(o)$ is defined by the mean $\mu^{R(o)}$ of $R(o)$ and $v^{R(o)}$ and is visualized as the red solid line perpendicular to the plane spanned by A_1 and A_3. Now, we are able to measure how much p deviates from the subspace hyperplane $\mathcal{H}(R(p))$ spanned by its reference set $R(p)$. The deviation of any point o to a subspace hyperplane $\mathcal{H}(S)$ is thereby naturally defined as the Euclidean distance in the subspace which is perpendicular to the hyperplane. This can simply be computed using a weighted Euclidean distance between o and μ^S using the subspace defining vector v^S as weight vector, i.e.,

$$dist(o, \mathcal{H}(S)) = \sqrt{\sum_{i=1}^{d} v_i^S \cdot (o_i, \mu_i^S)^2}. \tag{2}$$

The idea of this distance between a sample 3D point o and the subspace hyperplane of its reference set $R(o)$ is illustrated in Figure 1(c). This distance value is a very intuitive measurement for the degree of outlierness of any $p \in \mathcal{D}$ w.r.t. the set of points in $R(p)$. A value near 0 indicates that the particular point p fits very well to the hyperplane $\mathcal{H}(R(p))$, i.e., is no outlier, whereas a considerably higher value indicates that p is an outlier. The final subspace outlier degree is defined as follows.

Definition 1 (subspace outlier degree). *Let $R(p)$ denote a set of reference objects for object $p \in \mathcal{D}$. The* subspace outlier degree *(SOD) of p w.r.t. $R(p)$, denoted by $SOD_{R(p)}(p)$, is defined as*

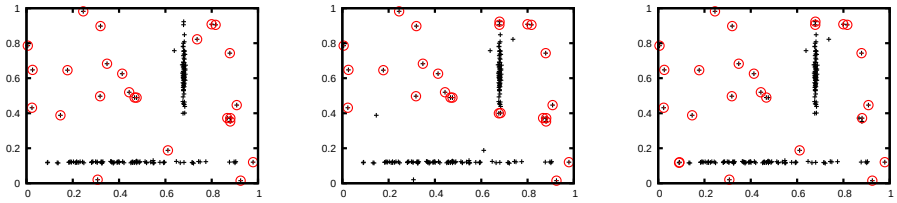
$$SOD_{R(p)}(p) := \frac{dist(o, \mathcal{H}(R(p)))}{\|v^{R(p)}\|_1},$$

i.e., the distance between point p and its reference set $R(p)$ according to Equation 2, normalized by the number of relevant dimensions as given e.g. by the number of entries $v_i^{R(p)} = 1$ in the weighting vector $v^{R(p)}$ as defined in Equation 1.

In contrast to most of the existing approaches, our model also gives an explanation *why* a point p is an outlier. Given an outlier p, we can obtain the subspace in which p is an outlier by simply inverting the subspace defining vector $v^{R(p)}$. This yields the subspace that is perpendicular to the subspace hyperplane of $R(p)$. In addition, we can derive the mean value of the points in $R(p)$ in that subspace. Thus, our model implicitly provides not only a quantitative outlier model but also a qualitative outlier model by specifying for each outlier the features that are relevant for the outlierness.

We now discuss how to choose a meaningful reference set for a given point $p \in \mathcal{D}$ to compute the outlierness of p. Existing local (full dimensional) outlier detection models usually examine the local neighborhood of p, e.g. the k-nearest neighbors or the ε-neighborhood based on Euclidean distance. However, due to the curse of dimensionality, distances cannot be used to differentiate points clearly in high dimensional feature spaces. As a consequence, the concept of "local neighborhood" is rather meaningless in high dimensional data (see e.g. [2]). An SNN approach usually measures the similarity of points based on the number of common nearest neighbors. An explanation for the robustness of SNN is that even though all points are almost equidistant to a given point p, a nearest neighbor *ranking* of the data objects is usually still meaningful. Two points p and q that have been generated by the same generating mechanism will most likely be neighbors or have similar neighbors in the subspace that is relevant for the common generating mechanism. Adding irrelevant attributes will blur these neighborhood relations by means of the absolute distances. However, most points of the common generating mechanisms will still be among the nearest neighbors of p and q. Thus, the number of shared neighbors of p and q will be large if both points originate from the same generating mechanism. Formally, let $N_k(p) \subseteq \mathcal{D}$ be the k-nearest neighbors of $p \in \mathcal{D}$ w.r.t. the distance function *dist*. The *shared nearest neighbor similarity* between two points $p, q \in \mathcal{D}$ is defined as $sim_{SNN}(p, q) = Card\,(N_k(p) \cap N_k(q))$. Now, the reference set $R(p)$ of p is the set of l-nearest neighbors of p using sim_{SNN}, i.e.,a subset of $\mathcal{D}$ that contains l points according to the following condition: $\forall o \in R(p), \forall \hat{o} \in \mathcal{D} \setminus R(p) : sim_{SNN}(\hat{o}, p) \leq sim_{SNN}(o, p)$.

The SOD algorithm relies on two input parameters. First, k specifies the number of nearest neighbors that are considered to compute the shared nearest neighbor similarity. This is not really a critical parameter as long as it is chosen high enough to grasp enough points from the same generating mechanism.



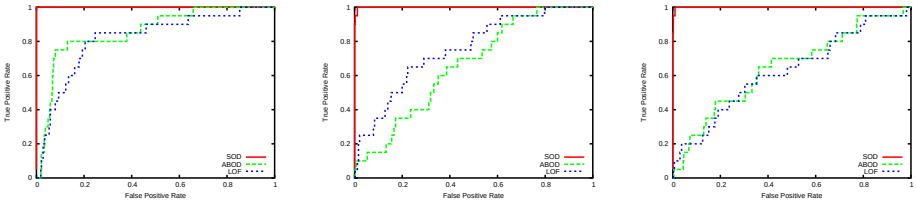
(a) Results of SOD. (b) Results of LOF. (c) Results of ABOD.

Fig. 2. Comparison of the 25-top ranked outliers in a sample 2D data set.

Second, l specifies the size of the reference sets. This parameter should also not be chosen too small for the same reason. Obviously, l should be chosen smaller or equal than k. Let us note that we have a third parameter α that specifies a threshold to decide about the significance of an attribute. If the variance of the reference set along an attribute is smaller than α times of the expected variance, then this attribute is considered relevant. In our experiments, setting $\alpha = 0.8$ yields consistently good results so we recommend to choose it accordingly. To compute the SOD, first the set of k-nearest neighbors of each of the n points of the database needs to be computed which requires in summary $O(d \cdot n^2)$ in the worst-case. This can be reduced to $O(d \cdot n \log n)$ if an index structure is applied to support the NN queries. Then, for each point p, the reference set of p consisting of the l nearest neighbors of p w.r.t. the SNN similarity needs to be computed which takes $O(k \cdot n)$, the mean and the variance of this reference set needs to be computed which takes $O(d \cdot l)$, and finally, the SOD can be computed. In summary, since $k \ll n$ and $l \ll n$, the runtime complexity of the latter steps and the overall complexity is in $O(d \cdot n^2)$ which is comparable to most existing outlier detection algorithms.

4 Experiments

We report the results of an experimental comparison of SOD with the full-dimensional distance-based LOF outlier model as one of the best-known outlier models and the full-dimensional angle-based ABOD outlier model as the most recent approach claiming to be specifically applicable to high dimensional data. All competitors are implemented within the ELKI-framework [6]. We first applied the competing outlier models to several synthetic data sets. Here, we focus on a toy 2D data set to illustrate the difference between a full dimensional approach like LOF or ABOD and the idea of a subspace outlier model followed by SOD. The results are visualized in Figure 2. Most points of the data set are produced by one of two generating mechanisms, for each mechanism only one attribute is relevant whereas the other is not. This results in one cluster of 80 points scattered along a line parallel to the y-axis and one cluster of 50 points scattered along a line parallel to the x-axis. In addition, 25 points have been generated randomly as outliers. The Figures display the 25 top-ranked outliers by



(a) 37 irr. attr. (b) 67 irr. attr. (c) 97 irr. attr.

Fig. 3. ROC curves for data sets of varying number of irrelevant attributes.

each approach (marked by red circles). It can be observed that SOD has no problems in finding the outliers that deviate from the generating mechanisms. On the other hand, LOF and ABOD have two potential types of errors. First, points of the generating mechanisms are wrongly reported as outliers (false alarms) because the density is not high enough in the surrounding area. Second, points that are not generated by one of the generating mechanisms are missed (false drops) because their surrounding area is dense enough. We further conducted several experiments on higher dimensional synthetic data sets. Here, we defined a Gaussian distribution of 430 points in three dimensions with $\mu = 0.33$ and $\sigma = 0.08$. Additionally, 20 outliers are placed in a range of 0.455 to 1.077 as minimal and maximal distance from the cluster center, respectively, whereas the outmost cluster point has a distance of 0.347 from μ. These values are given w.r.t. the three relevant dimensions only. Then we added 7, 17, ..., 97 irrelevant attributes with values uniformly distributed in the range $[0, 1]$, resulting in 9 additional data sets of dimensionality 10, 20, ..., 100, respectively. In all experiments, SOD produced better results in terms of accuracy compared to LOF and ABOD. Figure 3 presents example ROC curves catching the performance of all three approaches for these data sets. While LOF and ABOD are very competitive in lower dimensional data sets, their performance considerably deteriorates with higher dimensionality while SOD remains very stable at optimal values. Only starting at 80 dimensions, 77 of which are irrelevant attributes, SOD starts to retrieve a false positive as the 19th outlier. Even at 100 dimensions, it only retrieves one false positive as 18th outlier.

We applied SOD and its competitors to a data set of career statistics of current and former NBA players[1] including 15 important parameters like points per game, rebounds per game, assists per game until the end of the 2007/2008 season. The data are normalized in order to avoid a bias due to different scaling of the attributes. The eight players with top SOD and ABOD values are displayed in Table 1. Both ABOD and SOD give some insightful results on this data set. They also agree on many of the top outliers. Eddy Curry — top outlier for both algorithms — for example is a significant outlier because of his 100% quote on three point field goals (2 of 2). We also ran LOF on this data set detecting mostly

[1] Obtained from `http://www.nba.com`

Table 1. Results on NBA data set.

(a) Top-8 outlier retrieved by SOD.

Rank	Name	SOD
1	Eddy Curry	0.0807
2	Dennis Rodman	0.0678
3	Amir Johnson	0.0560
4	Karl Malone	0.0473
5	Shawn Marion	0.0470
6	Michael Jordan	0.0457
7	Avery Johnson	0.0408
8	Andrei Kirilenko	0.0386

(b) Top-8 outlier retrieved by ABOD.

Rank	Name	ABOD
1	Eddy Curry	0.0021
2	Amir Johnson	0.0035
3	John Stockton	0.0043
4	Hakeem Olajuwon	0.0053
5	Dennis Rodman	0.0058
6	Karl Malone	0.0063
7	Shaquille O'Neal	0.0068
8	Andrei Kirilenko	0.0076

players with exceptional high values in particular parameters or particularly low numbers.This experiment with real world data emphasizes that SOD can provide insightful information. It does not show a clear advantage over ABOD; they mostly agree on the outlier results. Depending on the use case, one or the other result can be seen as more useful. Outliers detected by ABOD can be seen as a more global kind of outlier, whereas SOD is stronger at detecting local outliers and additionally accounting for local feature correlation.

5 Conclusions

In this paper, we introduced SOD, a completely new approach to model outliers in high dimensional data. SOD explores outliers in subspaces of the original feature space by combining the task of outlier detection and relevant subspace finding. Our experimental evaluation showed that SOD can find more interesting and more meaningful outliers in high dimensional data with higher accuracy than full dimensional outlier models by no additional computational costs.

References

1. Hawkins, D.: Identification of Outliers. Chapman and Hall, London (1980)
2. Hinneburg, A., Aggarwal, C.C., Keim, D.A.: What is the nearest neighbor in high dimensional spaces? In: Proc. VLDB. (2000)
3. Kriegel, H.P., Kröger, P., Zimek, A.: Clustering high dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. (to appear in: ACM Transactions on Knowledge Discovery from Data (TKDD))
4. Breunig, M.M., Kriegel, H.P., Ng, R., Sander, J.: LOF: Identifying density-based local outliers. In: Proc. SIGMOD. (2000)
5. Kriegel, H.P., Schubert, M., Zimek, A.: Angle-based outlier detection in high-dimensional data. In: Proc. KDD. (2008)
6. Achtert, E., Kriegel, H.P., Zimek, A.: ELKI: a software system for evaluation of subspace clustering algorithms. In: Proc. SSDBM. (2008)