Explaining outliers by subspace separability

Barbora Micenková
and Xuan-Hong Dang and Ira Assent
Aarhus University
Aabogade 34, 8200 Aarhus, Denmark
Email: {barbora,dang,ira}@cs.au.dk

Raymond T. Ng
University of British Columbia
2366 Main Mall, Vancouver, BC, Canada
Email: rng@cs.ubc.ca

***Abstract*—Outliers are extraordinary objects in a data collection. Depending on the domain, they may represent errors, fraudulent activities or rare events that are subject of our interest. Existing approaches focus on detection of outliers or degrees of outlierness (ranking), but do not provide a possible explanation of how these objects deviate from the rest of the data. Such explanations would help user to interpret or validate the detected outliers.**

The problem addressed in this paper is as follows: given an outlier detected by an existing algorithm, we propose a method that determines possible explanations for the outlier. These explanations are expressed in the form of subspaces in which the given outlier shows separability from the inliers. In this manner, our proposed method complements existing outlier detection algorithms by providing additional information about the outliers. Our method is designed to work with any existing outlier detection algorithm and it also includes a heuristic that gives a substantial speedup over the baseline strategy.

I. Introduction and Motivation

In certain data analysis tasks, extraordinary objects (outliers) might be more important than the prevalent patterns. Exceptional data e.g. among scientific measurements may represent errors, but they can also indicate interesting phenomena, such as new stars in astrophysical data or forgeries in a forensic database of handwritten signatures. Given a database and a record identified to be an outlier, the task of *outlier explanation* is to describe what distinguishes the outlier from the rest of the database.

Existing outlier detection techniques find outliers based on certain criteria (e.g. distribution, distance, or density) and they handle outlierness in two different manners—as a binary property by labeling the data outlier/non-outlier, or as a degree. Outlierness degree is a real-numbered value assigned to each data point, reflecting the extent to which the point deviates from other data points. Degrees enable outlier comparison and ranking. However, almost all existing algorithms stop at the point of providing outlier ranking and leave the user without any explanation of *why* some data points deviate and *how*. Degrees do not bear any information about the *form* of outlierness, it is merely a quantitative information while any *qualitative* information is missing which makes interpretation of detected outliers hard. In particular, interpretation can be problematic if the data consists of many attributes where visualization or manual browsing is almost impossible. Thus, in order for outlier detection to be more usable in practice, an additional explanatory component is valuable.

In this paper we propose such an explanatory component that can extract additional knowledge about outliers in large databases with many dimensions. Specifically, given an outlier detected by existing outlier detection algorithms, we propose a method that determines possible explanations for the outlier. These explanations are expressed in the form of subspaces (attribute subsets) in which the given outlier shows separability from the inliers. In this manner, our proposed method complements existing outlier detection algorithms by providing additional information about the detected outliers. The method is designed to work with any existing outlier detection algorithm and it also includes a heuristic that gives a substantial speedup over the baseline strategy of searching through an exponential number of subspaces. For each outlier, the method constructs a binary classifier to separate the outlier from the inliers. Through extensive experimentation, we show that existing feature selection methods for this classification task can be leveraged to yield a subset of features that corresponds to a good explanation of the detected outlier.

The paper is organized as follows. Section II describes related work. In Section III, we discuss possible forms of explanations and propose a definition of outlier explanation. Section IV covers our approach to find explanatory subspaces and Section V describes experiments on a synthetic benchmark and several real-world data sets.

II. Related work

Very little work has been done towards outlier explanations so far. Both the two following studies concern the problem of outlier explanation[1] but they deal with categorical data: Angiulli et al. [1] measure the *abnormality* of combinations of attribute values featured by a given outlier with respect to the entire data set (global properties) or its subset (local properties). Ertöz et al. [2] present a whole framework for network intrusion detection (MINDS) and they explain anomalies by *association rules*. In this way, explanations are assigned to *groups* of outliers. Both of these methods assume categorical data or require categorization of attributes whereas we propose a mechanism that works with numerical (both discrete and continuous) attributes. Recently, Smets et al. [3] have indicated that outlier identification alone, without further explanation, is not sufficient. In that work, a method based on the minimum description length (MDL) has been proposed to identify anomalies from a transaction/binary dataset. By the MDL approach, a data transaction is an outlier if it needs

[1] The same problem has been called *outlying property detection* or *outlier characterization*.

to be encoded by an abnormally large number of bits. This information can then be used to characterize the exceptional properties of the outlying transactions. Akoglu et al. [4] also provide additional information that can be used for outlier explanation. However, they use weighted graphs in their work, and specific structures (such as near-cliques, heavy vicinities or dominant heavy links) for abnormal graphs (i.e., outliers) must be defined in advance in order to enable the feature extraction process. Unlike all these studies, our paper aims to develop an explanation mechanism for outliers in numerical data.

Only few outlier detection studies considered providing some qualitative information to explain the *form* of outlierness [5]. In multi-dimensional data analysis, *distance-based* (DB) outliers [6] have been widely used. By definition, they are such data objects where a predefined fraction of the database lies greater than distance D from them. Knorr and Ng [7] explain such outliers by the principle of dominance. They differentiate *strongest* and *weak* outliers, where an outlier $\mathbf{o}$ is strongest in some subspace $\mathbf{S}$ of the database iff there is no other object in any subspace $\mathbf{T} \subset \mathbf{S}$ that *dominates* $\mathbf{o}$. By this principle, an outlier is compared to every other outlier in the database in different attribute subspaces. This framework, however, does not scale up well for high dimensional data.

Density-based approaches (e.g. Local Outlier Factor—LOF [8]) measure the density of an object compared to the density in its neighborhood. For high dimensional data, *angle-based* outlier factor (ABOF) [9] has been proposed to assess the variance in angles among the difference vectors from the analyzed object to other objects in the database. Both of these methods output outlier scores and thus enable outlier ranking. However, they do not provide additional information to explain the outliers. On the other hand, LOCI [10] provides an "outlier plot" which gives the user an idea on how data is distributed in the vicinity of the analyzed object. From the plot, one can assess whether the object is inside a cluster, a part of a micro-cluster or if it is an outstanding outlier. However, no information on deviation in subspaces is given.

A different approach to outlier detection is based on *subspace clustering*. OutRank [11] utilizes a result of subspace clustering and scores each object according to the size and dimensionality of clusters in which it occurs. Drawbacks of the method are sensitivity to the employed subspace clustering technique and its high computational complexity. SOD [12] selects a reference set consisting of shared nearest neighbors for each outlier, and then it derives a subspace where the reference set exhibits lower variance than a specified threshold. OUTRES [13] and HiCS [14] both search for statistically relevant subspaces where the outlier might exhibit a high deviation, and then they aggregate outlier scores in these subspaces. The search for suitable subspaces can be prohibitive for high dimensional data. In sum, all of these methods are designed to be effective outlier detection methods; however they do not directly provide subspace explanation to help the user to interpret the results.

III. Problem definition

In this section, we examine possible forms of descriptions that could explain outliers in an intuitive manner and we give a definition of an explanatory subspace.

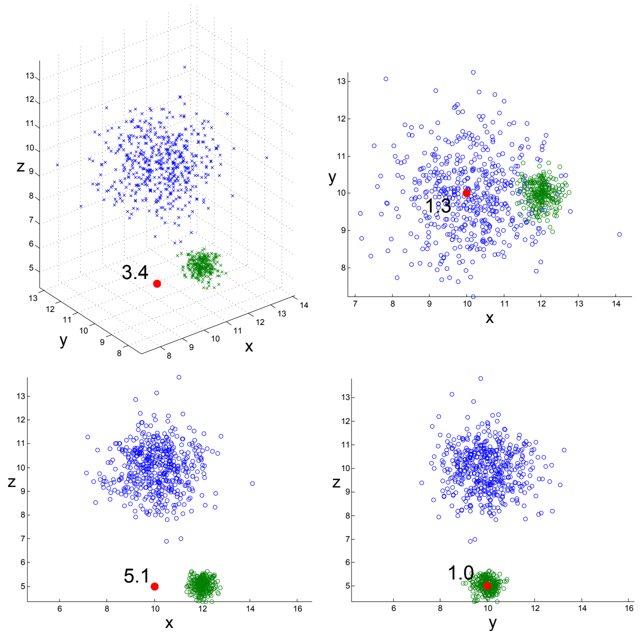


Fig. 1. A 3D space $\{x, y, z\}$ and all its 2D projections. $\{x, z\}$ is an explanatory subspace.

A. What is a meaningful explanation?

The form of a good explanation has to meet certain desirable properties. First, it must be easy and intuitive to process in order to avoid overloading the user with too much information. Furthermore, computation must be fast so that the technique can be used in an interactive manner. And lastly, every outlier should get a separate explanation such that it can possibly be further automatically processed.

We will briefly discuss some obvious possible explanation techniques: visualization, attribute weights and a transformed subspace. *Visualisation* is the closest analytical tool to human perception. However, even with the advances in information visualization [15], it is hard to visualize in full space high dimensional data. *Attribute weights* can reflect independent deviation in single attributes. Nevertheless, in real-world data sets, many outliers only become observable in certain combinations of attributes rather than in single dimensions. For the 3D data in Fig. 1 (upper left) and the marked outlier (big red point), attribute weights would not give any meaningful information since the object does not deviate in any single attribute. This seems to be possible to solve by comparing the outlier to its nearest neighbors and finding the weights say by local PCA. However, in real-world data sets, outliers often do not deviate from a compact cluster, and its nearest neighbors may be scattered (the outlier is "surrounded" by other data).

In order to overcome shortcomings of the presented approaches, we suggest a different kind of explanation: providing user with information about the combination of dimensions (an attribute subset) in which an outlier shows the greatest deviation. Thus, we need to find what we call an *explanatory subspace* projection (which may be different for different outliers). A simple illustration of the explanatory subspace projection is given in Fig. 1. The big red point was identified as an outlier, however, the user does not find anything unusual about it because the values of every single attribute are close to the mean values of the clusters. Intuitively, the point deviates in subspace $\{x, z\}$ (third from left), which is also the information that the user needs in order to understand *how* the outlier deviates. Even though the outlier is also observable in the 3D space, the third attribute, y, does not contribute to its deviation and it would be redundant to return it to the user. Thus, a good explanatory susbspace highlights the outlierness but at the same time it is minimal in number of attributes.

It should be emphasized that an explanatory subspace *cannot be derived* just by analyzing the vicinity of an outlier in the full space. In our example, if we compared the outlier to its nearest neighbors found in the full space, we would conclude that subspace $\{x\}$ is enough to explain the deviation. However, many other points in the database have approximately the same value of x and they were not identified as outliers (since there is a cluster in a higher dimensional space but the user does not know about it since no clustering has been done). The user would be confused by such an explanation and he/she would anyway have to look through all other attribute values manually which means that the provided explanation was useless. Thus, we are interested in the deviation of the point compared to its *projected* neighborhood in order to provide meaningful explanation.

B. *Explanatory subspaces*

We have a database $\mathbf{DB} \subseteq \mathbb{R}^d$ with n points in d dimensions and a point $\mathbf{p} = (x_1, \ldots, x_d) \in \mathbf{DB}$. Let $\mathbf{S} = \{s_1, \ldots, s_m\} \subseteq \mathbf{D} = \{1, \ldots, d\}$ denote an index set and $\mathbf{DB^S}$ a projection of the original database $\mathbf{DB^D}$ to the dimensions in $\mathbf{S}$. For ease of notation, we will further refer to the attribute subspace $\mathbf{DB^S}$ by its index set $\mathbf{S}$. Let $\mathbf{p_S} = (x_{s_1}, \ldots, x_{s_m})$ be a projection of $\mathbf{p}$ onto $\mathbf{S}$.

Definition 1: An **outlier scoring function** ω: $\mathbf{DB} \rightarrow \mathbb{R}$ is a measure of deviation that assigns each object $\mathbf{p}$ an outlierness score based on how $\mathbf{p}$ deviates from the other data. More deviating objects get higher scores.

Outlier scoring function allows for comparison of objects w.r.t. their outlierness and thus enables outlier ranking. Having such a function ω, we can apply it in all subspace projections and subsequently rank the subspaces according to how much the projected point deviates in it. In a subspace which is suitable to explain outliers, the point will exhibit high deviation and at the same time the subspace will be relatively low-dimensional. In principle, users could even be interested in multiple explanatory subspaces.

Definition 2 (Explanatory Subspace): An **explanatory subspace** $\mathbf{S_p^*}$ for an outlier $\mathbf{p}$ is an attribute subspace $\mathbf{S} \subseteq \mathbf{D}$ where the outlier score of the projected point, $\omega(\mathbf{p_S})$, is high and at the same time the dimensionality of the subspace, $|\mathbf{S}|$, is low.

A bottleneck of this approach is that we are comparing scores of a point projected onto subspaces of different dimensionality. As described in [16], outlier measures based on computations of distances suffer from the *curse of dimensionality* which leads to prioritizing outliers in either low or high-dimensional projections (*dimensionality bias*). There are several reasons for this. Most importantly, the expected values of L_p norms depend on dimensionality, and scores tend to concentrate in higher dimensions which leads to their poor contrast. Thus, outlier scores from different subspaces are generally uncomparable and there is still not a satisfactory solution to this problem. However, for a couple of outlier measures, normalization methods have been proposed [17] that make comparisons of scores from different subspaces possible.

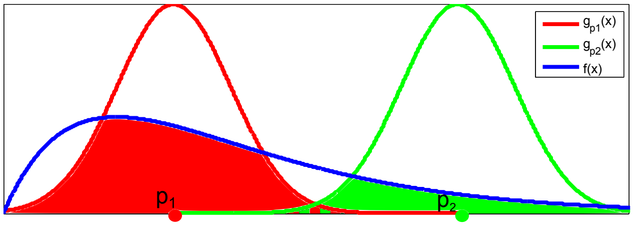


Fig. 2. Measuring outlierness by separability. $\mathbf{p}_1, \mathbf{p}_2$ are points from the distribution $f(x)$ and the normal distributions $g_{p_1}(x)$ and $g_{p_2}(x)$ were artificially generated.

Another inherent problem of Def. 2 is that finding explanatory subspaces generally involves an exhaustive search through an exponential number of subspaces. In each subspace, a scoring function needs to be calculated which requires computing distances to all other points in the database. Overall, such a brute-force search method requires $O(n2^d)$ which is clearly computationally infeasible. Therefore, in the next section, we will present:

1) a subspace selection principle which can handle the dimensionality bias,
2) an outlier scoring function that is possible to approximate in each subspace without re-calculating all distances,
3) a fast heuristic search method to find explanatory subspaces.

IV. Approach

We have made clear that a meaningful explanation can be provided as a relatively low-dimensional subspace in which the point shows high deviation. In other words, we search for the *most relevant combination* of attributes. We present a technique that is based on the observation that outlierness of a point is related to its *separability* from the rest of the data and that in classification, good separability of classes is a precondition to successful learning. Therefore, we will first convert the problem into classification and then utilize some classical feature selection methods in order to find explanatory subspaces.

A. *Measure of separability*

In this section, we will define a measure which quantifies *separability* of a point from the rest of the data and show how it is related to its outlierness. The merit of having such a measure will become obvious later when we show how to approximate its computation across subspaces. A brief example of how separability is related to outlierness is given in Fig. 2. Two points of interest, p_1, p_2, are picked from univariate data with an arbitrary distribution $f(x)$. Then, artificial points drawn from normal distribution $g_{p_1}(x)$ and $g_{p_2}(x)$ are generated around both p_1 and p_2. According to Fig. 2, p_2 clearly deviates from the rest of $f(x)$ and thus is a true outlier, and at the same time the distributions $g_{p_2}(x)$ and $f(x)$ are quite well separable. On the other hand, p_1 lies in the middle of $f(x)$ and is not an outlier, and the distributions $g_{p_1}(x)$ and $f(x)$ are not separable. Thus, separability is related to the outlierness of the points and it can be quantified by computing the area of intersection of the two respective distributions.

Definition 3 (Separability): Let $f(\mathbf{x})$ be probability density function of $\mathbf{x} \in \mathbf{DB}$ in d dimensions and $g_{\mathbf{p}}(\mathbf{x})$ be probability density function of normal distribution $\mathcal{N}_d(\mathbf{p}, \Sigma)$ where Σ is a $d \times d$ scalar matrix $\Sigma = \lambda^2 I$. Then, **separability** of an object $\mathbf{p} \in \mathbf{DB}$ from the rest of the data is

$$\operatorname{sep}(\mathbf{p}) = \int_{-\infty}^{\infty} \min\bigl(f(\mathbf{x}), g_{\mathbf{p}}(\mathbf{x})\bigr) \mathrm{d}\mathbf{x}.$$

Note that λ is a parameter of the function which controls the width of the multivariate normal distribution. Separability is inversely proportional to outlier scores from Def. 1 because the overlap of the distributions is smaller for more striking outliers. Practically, separability is difficult to measure because $f(x)$ is unknown. However, it is possible to approximate it if we frame the problem as classification and let $f(x)$ and $g(x)$ be two different class distributions. In order to do that, we need to show that separability corresponds to the average probability of error at binary classification.

Theorem 1 (Separability as error at classification): Separability can be measured as a classification error.

Proof: In order to measure $\operatorname{sep}(\mathbf{p})$, we let all $\mathbf{x} \in \mathbf{DB}$ represent class C_1 and randomly generate class C_2 by drawing points from multivariate normal distribution $\mathcal{N}_d(\mathbf{p}, \Sigma)$. Then, $f(\mathbf{x})$ and $g_{\mathbf{p}}(\mathbf{x})$ correspond to class-conditional probability density functions $p(\mathbf{x}|C_1)$ and $p(\mathbf{x}|C_2)$, respectively. We can assume prior probabilities $P(C_1) = P(C_2) = 0.5$ because we can control how many points are generated in class C_2. According to Bayesian decision theory [18], the expected probability of error is defined as

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error} \mid \mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \int_{-\infty}^{\infty} \min\bigl[P(C_1 \mid \mathbf{x}), P(C_2 \mid \mathbf{x})\bigr] p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \tag{1}$$

and by Bayes' theorem

$$P(C_j \mid \mathbf{x}) \, p(\mathbf{x}) = p(\mathbf{x} \mid C_j) P(C_j). \tag{2}$$

The error can thus be expressed as

$$\begin{aligned} P(\text{error}) &= \int_{-\infty}^{\infty} 0.5 \cdot \min\bigl[p(\mathbf{x} \mid C_1), p(\mathbf{x} \mid C_2)\bigr] \, \mathrm{d}\mathbf{x} \\ &= 0.5 \cdot \int_{-\infty}^{\infty} \min\bigl[f(\mathbf{x}), g_{\mathbf{p}}(\mathbf{x})\bigr] \, \mathrm{d}\mathbf{x} = 0.5 \cdot \operatorname{sep}(\mathbf{p}). \end{aligned} \tag{3}$$

We have shown that separability from Def. 3 is proportional to the average error at classification. ■

Since an error is greater for grossly overlapping distributions while outlierness of a point should be less in that case, we will prefer to use accuracy at classification instead of an error to quantify outlierness. Moreover, if we wanted to calculate the expected error, we would need to estimate $f(x)$ by putting some assumptions (e.g. normality assumption) on the distribution of the data which is not desirable. Therefore, we will choose other (non-parametric) supervised learning techniques to measure outlierness.

B. Outlierness as accuracy at classification

Let us define a new outlier scoring function based on separability—a measure of accuracy at classification. First, we have to set up the classification problem and define the classes. Since $\mathbf{DB} \subseteq \mathbb{R}^d$, let $\mathbb{R}^d$ be our input space. Classification is the task of learning a target function f that maps each point $\mathbf{x}$ from the input space to a predefined class label attribute y. We have two classses in our setting and therefore $y = \{C_{in}, C_{out}\}$. The classifier learns from a classification set of the form $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ and builds a *model*. The model, in general, can be used to predict class labels of any data points from the input space and its *accuracy* is reported as the proportion of correctly classified points.

Definition 4 (Classification Set): Given a point $\mathbf{p} \in \mathbf{DB}$, $|\mathbf{DB}| = n$, let us define an inlier class as $\mathcal{I}_p = \{\mathbf{x} \,|\, \mathbf{x} \in \mathbf{DB}, \mathbf{x} \neq \mathbf{p}\}$ and an outlier class as $\mathcal{O}_p = \{\mathbf{p}\} \cup \{m_i\}_1^{n-2}, m_i \sim \mathcal{N}_d(\mathbf{p}, \Sigma)$ where Σ is a $d \times d$ scalar matrix $\Sigma = \lambda^2 I$. Then a **classification set** $\mathcal{T}_p$ for a point $\mathbf{p}$ is

$$\mathcal{T}_p = \{(x, C_{out}) \,|\, x \in \mathcal{I}_p\} \cup \{(y, C_{in}) \,|\, y \in \mathcal{O}_p\}.$$

The following definition of an outlier ranking function is based on the proof from Sec. IV-A and the fact that low expected error implies high classification accuracy [18].

Definition 5 (Outlierness as Accuracy): Let us define an outlierness of a point $\mathbf{p}$, $\omega(\mathbf{p})$, as an accuracy at classification of the classification set $\mathcal{T}_p$.

We emphasize that the classification set will be different for every point $\mathbf{p}$ and that a parameter λ needs to be set in order to establish the classification set. We will pay attention to setting the parameter for outlier explanation in Sec. IV-E. Furthermore, we prefer the classes to be approximately balanced to avoid problems connected to imbalanced classification [19].

C. Explanatory subspaces by separability

In order to find explanatory subspaces, we need to be able to assess the separability in the form of outlierness at accuracy in every subspace:

Definition 6: The **outlierness** of a projected point $\mathbf{p_S}$, $\omega(\mathbf{p_S})$ is measured as accuracy at classification of the classification set $\mathcal{T}_p^S$ created in the full space of the database and then projected onto $\mathbf{S}$.

Intuitively, the projection where classes are well separable (and therefore accuracy at classification is high) will be a good subspace for explanation. An enourmous benefit of our measure is that we can find such projections very fast thanks to well-tuned pruning techniques that have been widely studied in machine learning under the name *feature selection*. Feature selection has been used to overcome the curse of dimensionality and to improve learning performance by removing irrelevant and redundant attributes [20]. In our scenario, we will apply it in a novel way—to find explanatory subspaces:

Property 1: An explanatory subspace for an outlier $\mathbf{p}$, $\mathbf{S}^*_{\mathbf{p}}$, can be detected by applying a **feature selection** method in order to find the set of attributes that are the most relevant for separation of the classification set $\mathcal{T}_p$.

A great number of ready-made techniques for feature selection can be found in literature [20]. A short overview is given in the next section.

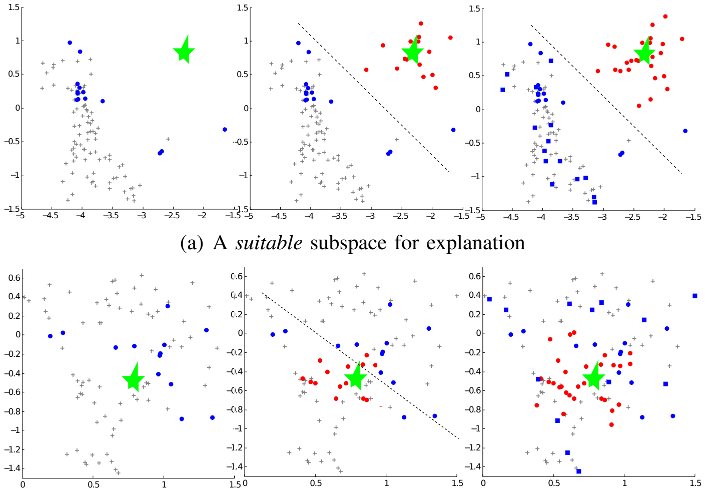
D. Feature selection

Feature selection methods include feature ranking and feature subset selection. *Ranking* of individual features is performed independently of the context of others which is not suitable for our setting since we want to find a combination of attributes that enhances outlierness. *Subset selection* methods either use a classifier as a black box to evaluate feature subsets (so called *wrappers*) or they perform feature selection as a part of the training process and are specific to a given classifier (*embedded methods*). In our experiments, we adopted two different feature selection techniques—forward selection by SVM [21] and lasso [22]. Forward selection is a wrapper algorithm that greedily incorporates attributes into larger subsets, being navigated by accuracy of support vector machine at classification of the training data. Lasso is an embedded method that combines subset selection and ridge regression. It can efficiently be implemented by least angle regression (LARS) [23]. In general, any efficient subset feature selection method can be used but it is absolutely crucial to avoid overfitting since in such case, classes of any distributions could be separated.

E. Sampling the classes for fast approximation

Now we describe in detail how the classification problem for feature selection is set up. Clearly, classifying 1 outlier against all inliers is not meaningful and therefore the outlier class needs to be oversampled. However, a drawback of the measure given in Def. 4 and 5 is that complexity of the computations depends on the size of the database. In this way, finding an outlier explanation can be slow for a very large data set. Therefore, we present a scheme where we *sample* the inlier class instead of including all data into it. Oversampling and subsampling techniques have been widely used in imbalanced classification where positive examples are very scarce [24, 19], and therefore we will build on these techniques.

Subsampling the inlier class: Uniform random subsampling of the data does not perform well because we may lose information about the vicinity of a point which is in fact essential for measuring its separability. However, nearest neighbors of different subspace projections $\mathbf{p_S}$ of the same point are *different*, and we want to avoid quering them in every single subspace. Thus, we give a heuristic approach which emphasizes the neighborhood of an outlier found in the full space and subsamples the rest of the data. We will further use the notions of *k-distance* and *reference set*, adapted from [8]:



(a) A *suitable* subspace for explanation

(b) An *unsuitable* subspace for explanation

Fig. 3. Sampling the classification set.

Definition 7: Let **k-distance** of an object $\mathbf{p} \in \mathbf{DB}$, denoted by k-distance($\mathbf{p}$), be the distance $d(\mathbf{p}, \mathbf{p}')$ between $\mathbf{p}$ and its k-th nearest neighbor $\mathbf{p}'$.

Definition 8: Let the **reference set** of a point $\mathbf{p}$, denoted by $R_k(\mathbf{p})$, be the set of points $\mathbf{x}$ whose distance from $\mathbf{p}$ is less or equal to the k-distance:

$$R_k(\mathbf{p}) = \{\mathbf{x} \in \mathbf{DB} \setminus \{\mathbf{p}\} \mid d(\mathbf{p}, \mathbf{x}) \leq k\text{-distance}(\mathbf{p})\}.$$

Now we are ready to redefine the inlier class $\mathcal{I}_p$ from Def. 4.

Definition 9 (Sampled inlier class): Given a point $\mathbf{p}$, a sampled **inlier class** $\mathcal{I}'_p$ is constructed as a union of the reference set $R_k(\mathbf{p})$ and a set of randomly drawn points from the rest of the database:

$$\mathcal{I}'_p = R_k(\mathbf{p}) \cup \{\mathbf{q}_i\}_1^r,$$

where $\mathbf{q}_i \in \mathbf{DB} \setminus R_k(\mathbf{p}) \setminus \{\mathbf{p}\}$, and $r = |R_k(\mathbf{p})|$.

Note that r, the cardinality of $R_k(\mathbf{o})$, is often equal to k, but it can be greater in case of ties. In Fig. 3, the principle of sampling the classes is illustrated. An outlier (the same in both figures) is marked by a green star and two different 2D subspace projections of the same 10-dimensional data are shown. The subspace in Fig. 3(a) is suitable for explanation while the one in Fig. 3(b) is not. Grey crosses denote the projected data points, blue dots the nearest neighbors from the full space and blue boxes are randomly selected points. Red dots are artificially generated data. Leftmost figures display the projected reference set. Middle figures show how the reference set itself is separable from the outlier class, and the figures on the right depict the final classification setting where the inlier class is sampled according to Def. 9. Note that for the good explanation subspace, separability from the reference set is the same as separability from the sampled inlier class. However, this does not hold for the other subspace (b) where the reference set is projected far from the outlier which clearly does not mean that the outlier is well seperable from the rest of the data.

Oversampling the outlier class: An outlier class is generated in the same way as in Def. 4 but it is desirable to have the same size as the inlier class so that the classification problem is balanced. Since $k, r \ll |DB|$ and k is a user parameter, scalability with respect to the database size is ensured.

Setting the covariance matrix $\Sigma = \lambda^2 I$ of the normal distribution directly affects separability. However, since we are only interested in relative separability in different subspaces, setting the parameter is not crucial, which we will also show in the experiments. The wider the distribution is, the worse class separability in *all* subspaces. Thus, as long as we do not set the parameter α too low (and then the accuracy of the classifier is 100% in all subspaces) or too high (and the accuracy is 50% in all subspaces), it does not change the relative order of subspaces. To ensure that λ is always set reasonably and that it is tailored to data of different densities, we propose to make it proportional to the k-distance of $\mathbf{p}$ and normalize it by the dimensionality d of the database:

Definition 10 (Covariance matrix of outlier class): Let the **covariance matrix** of the normal distribution of the outlier class be $\Sigma = \lambda^2 I$, where $\lambda = \alpha \cdot \frac{1}{\sqrt{d}} \cdot k\text{-distance}(\mathbf{p})$.

Thus, instead of λ, we introduce a user parameter α which is better interpretable. $\sqrt{d}$ is the upper bound on distance between any two arbitrary data points $\mathbf{p}, \mathbf{q} \in \mathbf{DB}$ where $p_i, q_i \in [0, 1]$. Normalization by $\sqrt{d}$ is merely needed to ensure that α has the same effect independently of the dimensionality of data since k-distance increases with dimensionality.

V. Evaluation

In order to precisely evaluate our method, we need data with a specific kind of ground-truth: labeled outliers with annotated attribute subsets that represent an ideal explanation. To the best of our knowledge, there is no such explanatory ground-truth for real data publicly available. Later in this section we therefore show how to perform a valuable evaluation on real data in the absence of such a ground-truth. For synthetic data sets where we have outliers and annotated subspaces, we use *Jaccard index* and *precision* to assess the accuracy of the computed explanation. Retrieving a subspace that is similar but not exactly the same as in the ground-truth is therefore counted as a partial success. Let us denote the true subspace that explains an outlier as T, and a retrieved subspace as P. Jaccard index is defined as a fraction of the size of the intersection and the size of the union of two sets: $Jaccard(T, P) = \frac{|T \cap P|}{|T \cup P|}$ while precision is a fraction of the size of the intersection and the retrieved subspace: $precision(T, P) = \frac{|T \cap P|}{|P|}$.

For real-world data with labeled outliers but no annotated subspaces, we use an existing outlier scoring technique, run it exhaustively on all possible subspaces projections of the data and normalize scores to overcome the dimensionality bias. Then, we create a *ranking of subspaces* for each outlier, and select the top subspace as a reference to be compared with the results of our proposed approach. We chose a traditional technique LOF [8] for comparisons.

LOF (Local Outlier Factor) is a density-based outlier detection algorithm that computes an outlierness score (outlier factor) of a point $\mathbf{p}$ by comparing the density of $\mathbf{p}$ to the density of its k-nearest neighbors. It can be proved [8] that for uniform distribution, an outlier factor of a point is expected to be 1 regardless the dimensionality. However, it has been recently shown [16] that for other distributions the variance of scores decreases with dimensionality and therefore normalization is still needed in order to compare scores from different subspaces. Thus, we normalize scores as suggested in [17]. We will refer to this approach as to a *reference exhaustive algorithm* but the reader should be aware that a choice of a different scoring function than LOF could lead to slightly different reference results. It should also be noted that traversing all subspaces is very inefficient and therefore we had to limit this approach to experiments on data with maximum of 10 dimensions.

VI. Experiments

We validate the proposed algorithm in a series of experiments on a synthetic benchmark and three real-world data sets. We show that the proposed algorithm gives very good results for data sets with medium number of dimensions (up to 75) while being orders of magnitude faster than the reference exhaustive algorithm. We further show that our algorithm has a superior performance to all the baseline approaches while being comparably efficient. In Sec. VI-B, we demonstrate robustness with respect to the parameter settings.

The proposed explanation algorithm will be applied in two variants, differing in the feature selection component: 1) forward feature selection by SVM [21] (FS SVM) and 2) LARS-lasso [23]. For data sets without a ground-truth, we refer to the reference exhaustive algorithm— subspace ranking by LOF (Sec. V). For baseline comparisons, we picked the following three approaches:

1) SOD [12],
2) random subspace selection,
3) selection of top attributes ranked by LOF scores obtained in single dimensions (LOF top-random)

The first baseline, SOD, is an outlier detection method that identifies a subspace for each outlier. For the second baseline, we randomly select 1 to 5 attributes. The result of the third baseline is not a subspace but a ranking and so we randomly select 1 to 5 attributes from the top of the ranking.

We implemented our method in Java and used the RapidMiner software [25] for feature selection, and MATLAB [26] to produce figures. SOD is implemented in the ELKI framework [27]. We used Intel® i5-560M CPU with 4GB RAM for all experiments.

A. Data sets

For experiments on **synthetic** data, we use the collection of benchmarking data sets published by Keller et al. [14]. They contain subspace clusters and outliers generated in subspaces of 2 to 5 dimensions. These subspaces are used as a ground-truth for explanations. The outliers were generated in such a way that they are not observable in any lower dimensional subspace projection which makes explanations more challenging. The collection contains data sets of 10, 20, 30, 40, 50 and 75 dimensions, each consisting of 1000 data points and 19 to 111 outliers.

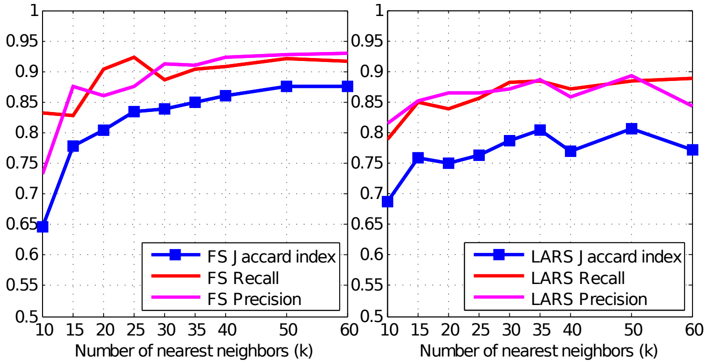


Fig. 4. Sensitivity of parameter k which controls subsampling of the inlier class. HiCS data set with 10 dims and 1000 data points. The average Jaccard index, precision and recall were measured. Forward feature selection method evaluated by SVM on the left, LARS-lasso method on the right.

For experiments on **real-world** data, we use 3 publicly available data sets: modified `vowel` data set from the UCI repository [28], `stamps` data set that we collected and for the first time publish in this work[2], and the well-known KDD Cup'99 data set for intrusion detection. `Vowel` is not primarily an outlier detection data set and therefore we had to adjust it for our purposes by randomly picking all data of one class to be normal instances and then one instance from every remaining class to represent outliers. In total, the data set has 10 dimensions and 100 instances from which 10 are outliers. `Stamps` is a forensic data set and its records are features extracted from scanned images of ink and photocopied stamps. The features are based on color distribution, texture and edge sharpness of the stamps. A detailed description of the data set can be found in [29]. Ink (genuine) stamps represent normal points and photocopied (forged) stamps form an outlier class. It has 9 dimensions and contains 340 instances from which 31 are outliers. `Stamps` is a difficult data set for outlier detection—maximum AUC (area under the ROC) of only 0.71 could be achieved by LOF method while AUC $= 0.95$ was achieved for `vowel`.

Further, we use the test data from KDD Cup'99 which consists of $311,029$ instances in 33 dimensions (we picked all numerical attributes except for `num_outbound_cmds` which was redundant). The instances correspond to connection records from a LAN at MIT Lincoln Labs. The data set has been criticised in the intrusion detection community for the fact that it may not reflect real-life conditions since the traffic was simulated [30], however, it has been widely used in other studies and it is one of few well-documented data sets for intrusion detection. Each connection is labeled either normal or one of 37 different attacks. Some attack types only have few instances in the data while others are represented by tens of thousands of records. We remove duplicate records and normalize all attributes. Since there are too many anomalies in the data (ca 80%), we randomly choose 3500 that we keep, such that all attack types are covered.

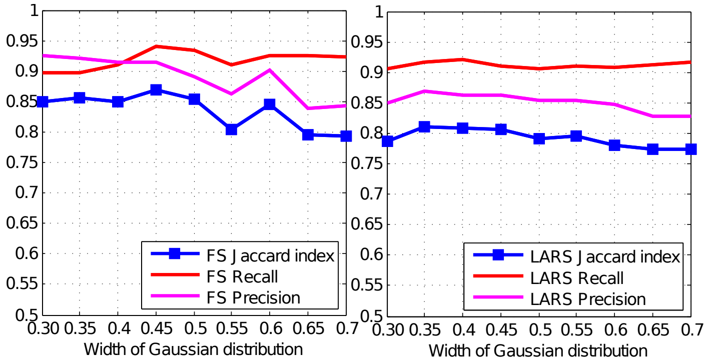


Fig. 5. Sensitivity of parameter α which controls the width of the Gaussian distribution of the outlier class. The same data and methods as in Fig. 4.

B. *Sensitivity of parameters*

Our method requires 2 parameters to be set: k determining k-nearest neighbors in the reference set which directly influences the subsampling ratio of the inlier class, and α controlling the width of the Gaussian distribution of artificial points generated around each outlier directly influencing the outlier class.

We evaluated our method on the 10-dimensional HiCS data set, varying the value of k from 10 to 60 nearest neighbors while fixing $\alpha = 0.35$. The results for both feature selection variants are shown in Fig. 4. For each value of k, they are averaged over 5 runs and over all outliers. For feature selection evaluated by SVM (the left plot), there is a clear trend while for the LARS-lasso method (the right plot) there are small fluctuations. Nevertheless, we can observe that the performance improves up to $k = 35$ and then it stabilizes. Better performance for greater values of k is expected since a bigger part of the data set is sampled. Overall, the experiments reveal that the method is not very sensitive to the number of selected nearest neighbors and that with our approach (Sec. IV-E), it is sufficient to sample a small fraction of the data. We will set $k = 35$ for all further computations.

In the second experiment with the same data set, we vary α from 0.3 to 0.7. Accuracy (see Fig. 5) is very stable for LARS-lasso (on the right) and slightly decreasing for FS SVM (on the left) which is in accordance with the theory (Sec.IV-E). Results are stable due to the fact that the separability of classes is compared among different subspaces (internally by the feature selection method) but no absolute values are used. We fix $\alpha = 0.35$ for all our experiments.

Besides, LARS-lasso method for feature selection requires setting a threshold t. Nevertheless, the threshold is not very sensitive and we obtained stable results for $t \in [0.25, 0.6]$ and set $t = 0.35$ for all our experiments. The forward feature selection comprises SVM. Since we want to avoid overfitting, we use linear C-SVM with regularization factor $C = 1$ in all experiments.

C. *Experiments on synthetic data sets*

In Fig. 6, we compare the results of our methods (solid lines) and 3 baseline approaches (dashed lines) on the HiCS

[2] The data sets can be downloaded from this link: http://cs.au.dk/~barbora/outDet.html

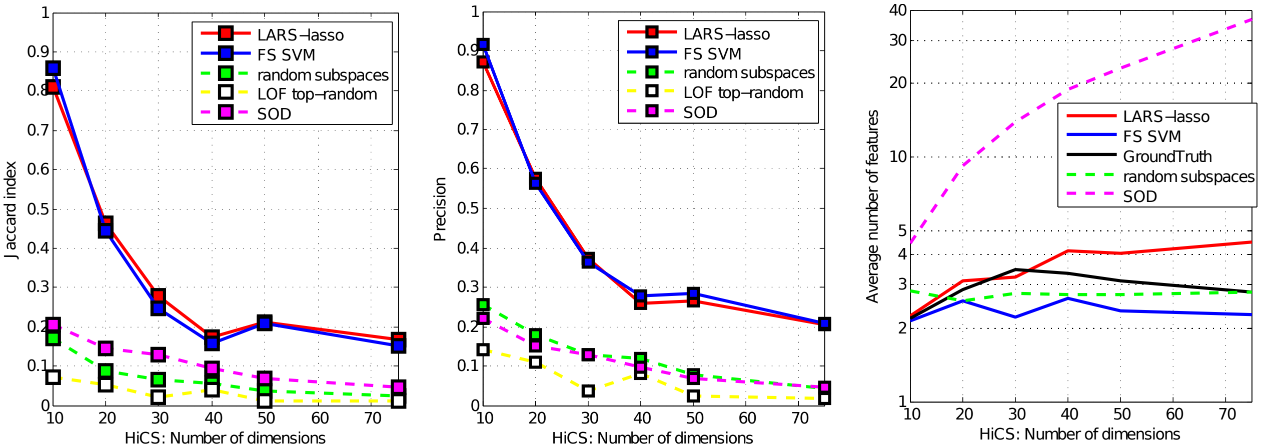


Fig. 6. Evaluation on 6 HiCS data sets of different dimensionality (10 - 75 dimensions). From left to right: Jaccard index, precision and average subspace dimensionality.

synthetic benchmark data sets. The values were obtained by averaging results for all outliers over 10 runs. SOD has 3 parameters and the best results could be achieved with the following settings: kNN=40, SNN=20 and α=0.9.

We can observe that the two variants of the proposed method, FS SVM and LARS-lasso, outperform all the three baselines for all 6 data sets. For the 10-dimensional data set, Jaccard index of LARS-lasso is 81% and even 86% for FS SVM (the leftmost plot). For random subspace selection it is merely 17%, for SOD 20% and the Jaccard index of top-ranked single attributes (LOF top-random) is only 7%. The last value is low due to the nature of the data set—all outliers are "hidden in subspaces" and cannot be identified by examining single attributes. For 75 dimensions, the performance of LARS-lasso drops to 17%, however, the problem becomes exponentially harder since we want to find a single subspace from more than $3.7 \cdot 10^{32}$ candidates—Jaccard index of random selection for the 75-dimensional data set is only 2%.

The plot of precision (middle) shows a similar trend. Average dimensionality of selected subpaces is plotted on the right (notice that the y axis is in log scale). Black solid line represents the ground-truth values—on average ca 3 attributes. Both FS SVM and LARS-lasso select a comparable number of attributes, however, SOD is very little selective and on average picks a subspace of size $d/2$ which is not desirable as explanations should be minimal. You might also notice a drop in the performance for the data set with 40 dimensions that might be related to the process of generation of the data.

D. Experiments on real-world data sets

For real-world data sets, we do not have a ground-truth with explanatory information and so we provide two different kinds of analysis. To verify that our method indeed finds explanatory subspaces, we compare our results to the result of a reference exhaustive algorithm described in Sec. V. For each outlier in the `vowel` and `stamps` data sets, we create a ranking of subspaces based on the normalized LOF score of the point in that particular susbpace. In order to approximately evaluate the result of our method (and all the baseline approaches), we report the position of the selected subspace in the subspace ranking. For data sets of different dimensionalities, the ranking has a different length and therefore we report a relative position. E.g. reporting that a selected subspace was in the 1% of the ranking for a 10-dimensional data set (1023 subspaces) means that the subspace was placed 1st to 10th. For a 9-dimensional data set (511 subspaces) it is 1st to 5th. We ran all the methods 20 times on each data set and reported the total percentages of all outliers occurring in the relative positions of the ranking.

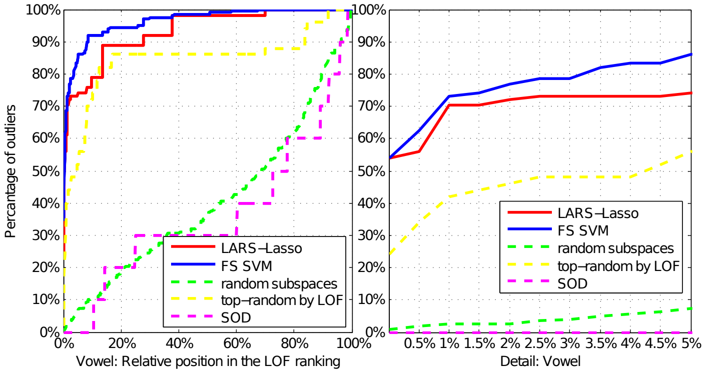
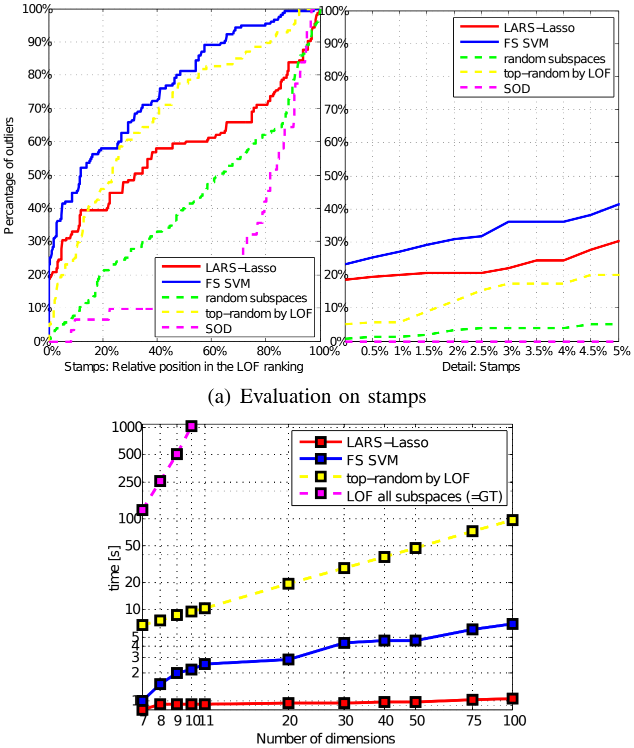


Fig. 7. Evaluation on the vowel data set (10 dimensions). Percentage of outliers plotted against their relative position in the LOF ranking. For more than 50% of all outliers, the subspaces found by both LARS-lasso and FS SVM were on the top of the ranking, and more than 70% were in the 1%.

Fig. 7 shows the results for the **vowel** data set. The whole ranking is depicted in the left plot and a detail of the first 5% (51 places) is on the right. FS SVM has slightly superior performance to LARS-lasso and substantially superior to the baselines. Both FS SVM and LARS-lasso were able to select the best subspace according to LOF ranking for 55% of all outliers! Over 70% is in the first percentile and 87% are in the first 5% according to FS SVM. SOD gives a similar result as random selection. Fig. 8 shows results for the **stamps** data set. The performance of all methods is by far worse which can be explained by the inherent complexity of the data set. Either some outliers cannot be distinguished from the normal data or LOF is not a suitable technique to distinguish them (e.g. due to gradually changing densities) and thus introduces errors to the evaluation process. FS SVM has a superior performance.



(a) Evaluation on stamps

(b) Scalability w.r.t. dimensionality

Fig. 8. In a), evaluation on stamps data (9 dimensions). Percentage of outliers plotted against their relative position in the LOF ranking. In b), runtimes on 1000-point data sets.

Top subspaces by LOF gain much better results than for the synthetic data sets since many outliers can be trivially revealed in single dimensions.

To validate that the proposed technique finds meaningful explanations, we perform another kind of analysis on the KDD Cup'99 data set. Since the data has 33 numerical attributes, we could not generate explanatory ground-truth in the same way as for `vowel` and `stamps` due to computational complexity of the reference exhaustive method. Thus, we derive explanations for all 3500 anomalies (representing attacks of 37 different types) by LARS-lasso. The practical merit of the method is shown by grouping the attacks by their type and comparing their derived explanations. An analysis of the data is given in Fig. 9. Assuming that attacks of the same type can often be explained by the same attributes, we constructed a 2D histogram of the occurrences of each attribute in explanations of attacks of a specific type. Each column is normalized by the number of attacks of that type. Thus, black color means that all attacks of a specific type (column) had a specific attribute (row) in their explanation while white color indicates that the attribute did not explain any attack of that type. Average dimensionality of the explanations was 1.8. To interpret the result well, deep knowledge of intrusion detection is needed. However, a simple example can be the `guess_passw` attack (*guess password*) in column 5 whose main indication is `num_failed_logins` (*number of failed logins*) in row 7. For some types of attacks, multiple attributes seem to be relevant, which might be caused by the fact that these attacks can show up in different parameters. Nevertheless, it is clearly visible from the histogram that our method returns similar explanations for attacks of the same type (note that identical records were removed from the data), which is meaningful and desired. For a method returning random results, the histogram would be gray. Considering efficiency, for this data set with 33 dimensions, deriving explanations of 3500 outliers took less than a minute without an index.

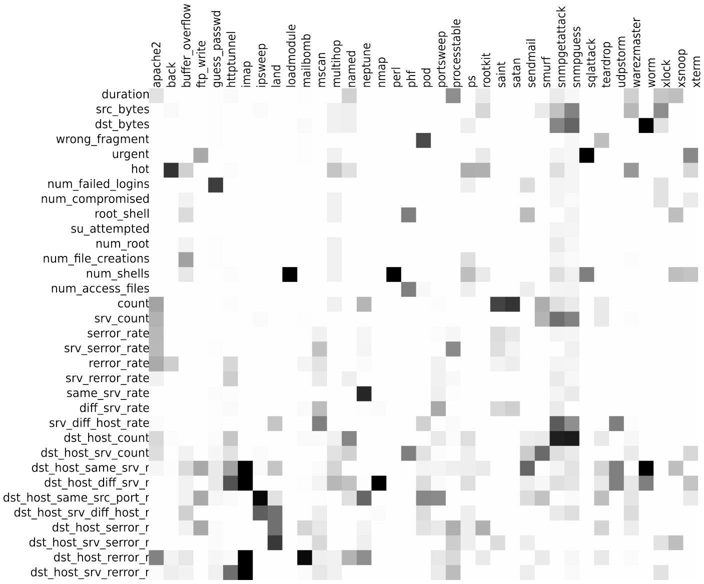


Fig. 9. Histogram of occurrences of attributes selected for explanation by our technique. The outliers are described by 33 attributes (rows) and divided into 37 attack categories (columns).

E. Scalability

Regarding runtimes, the proposed method scales very well with the *size* of the data set. k-nearest neighbors are only queried once in the full space of the data. Further computations are done with a small portion of the data—the number of sampled points is constant and equals approximately $4 \cdot k$ where k is a small number. Another benefit of the presented approach is that it makes use of well established methods such as SVM and lasso that have been optimized for fast computations.

In Fig. 8(b), runtimes with respect to the *dimensionality* of data are reported. Data sets with 1000 points, 10 outliers and 7 to 100 dimensions were used. LOF in all subspaces (the reference exhaustive technique) is only reported up to 10 dimensions—beyond that, computations were unfeasible.

VII. Conclusion

In this paper, we proposed an outlier explanation technique for numerical data which is independent of a detection mechanism. The technique derives a small attribute subset for each outlier where the point is well separable from the rest of the data. In this way it recommends a meaningful and easily interpretable outlier explanations which can be used for data exploration or validation of the outlier detection results. The importance of such a technique is even amplified by the fact that most of the state-of-the-art outlier detection algorithms do not provide descriptions of the form of the outlierness beyond a score. As a part of this work, we defined a new measure of outlierness by separability of sampled classes. The measure was used for extraction of explanations, however there is a potential to apply it for outlier ranking which remains a topic for future work.

References

[1] F. Angiulli, F. Fassetti, and L. Palopoli, "Detecting outlying properties of exceptional objects," *ACM Trans. Database Syst.*, vol. 34, no. 1, 2009.

[2] L. Ertöz, E. Eilertson, A. Lazarevic, A. Lazarevic, P. ning Tan, V. Kumar, P. Dokas, and J. Srivastava, "MINDS - minnesota intrusion detection system," 2004.

[3] K. Smets and J. Vreeken, "The odd one out: Identifying and characterising anomalies," in *In Proc. of the 11th SIAM Int. Conf. on Data Mining*, ser. SDM'11. SIAM/Omnipress, 2011.

[4] L. Akoglu, M. McGlohon, and C. Faloutsos, "OddBall: spotting anomalies in weighted graphs," in *In Proc. of the 14th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining*, ser. PAKDD'10. Springer-Verlag, 2010.

[5] C. C. Aggarwal, *Outlier Analysis*. Springer New York, 2013.

[6] E. M. Knorr and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets," in *In Proc. of the 24rd Int. Conf. on Very Large Data Bases*, ser. VLDB'98. Morgan Kaufmann Publishers Inc., 1998.

[7] ——, "Finding intensional knowledge of distance-based outliers," in *In Proc. of the 25th Int. Conf. on Very Large Data Bases*, ser. VLDB'99. Morgan Kaufmann Publishers Inc., 1999.

[8] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *In Proc. of the 26th Int. Conf. on Management of Data*, ser. SIGMOD'00. ACM, 2000.

[9] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in *In Proc. of the 14th Int. Conf. on Knowledge Discovery and Data Mining*, ser. KDD'08. ACM, 2008.

[10] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," ser. ICDE'03. IEEE Computer Society, 2003.

[11] E. Müller, I. Assent, U. Steinhausen, and T. Seidl, "OutRank: ranking outliers in high dimensional data," in *In Proc. of the 24th Int. Conf. on Data Engineering Workshop*, ser. ICDEW '08. IEEE Computer Society, 2008.

[12] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Outlier detection in axis-parallel subspaces of high dimensional data," in *In Proc. of the 13th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining*, ser. PAKDD'09. Springer-Verlag, 2009.

[13] E. Müller, M. Schiffer, and T. Seidl, "Statistical selection of relevant subspace projections for outlier ranking," in *Proc. of the 27th Int. Conf. on Data Engineering*, ser. ICDE '11. IEEE Computer Society, 2011.

[14] F. Keller, E. Müller, and K. Böhm, "HiCS: High contrast subspaces for density-based outlier ranking," in *Proc. of the 28th Int. Conf. on Data Engineering*, ser. ICDE '12. IEEE Computer Society, 2012.

[15] S. K. Card, J. D. Mackinlay, and B. Shneiderman, Eds., *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., 1999.

[16] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statistical Analysis and Data Mining*, vol. 5, no. 5, 2012.

[17] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Interpreting and unifying outlier scores," in *In Proc. of the 11th SIAM Int. Conf. on Data Mining*, ser. SDM'11. SIAM/Omnipress, 2011.

[18] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley, 2001.

[19] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *ICML*, D. H. Fisher, Ed. Morgan Kaufmann, 1997.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, 2003.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, 1995.

[22] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society (Series B)*, vol. 58, 1996.

[23] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, 2004.

[24] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, 2002.

[25] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid prototyping for complex data mining tasks," in *Proc. of the 12th Int. Conf. on Know. Discovery and Data Mining*, ser. KDD'06. ACM, 2006.

[26] MATLAB, *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc., 2010.

[27] E. Achtert, S. Goldhofer, H.-P. Kriegel, E. Schubert, and A. Zimek, "Evaluation of clusterings - metrics and visual support," in *Proc. of the 28th Int. Conf. on Data Engineering*, ser. ICDE'12. IEEE Computer Society, 2012.

[28] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[29] B. Micenková and J. van Beusekom, "Stamp verification for automated document authentication," in *5th Int. Workshop on Computational Forensics*. Springer, 2012.

[30] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, "Toward credible evaluation of anomaly-based intrusion-detection methods," *Trans. on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 5.