Chapter 1

HIGH-DIMENSIONAL OUTLIER DETECTION: THE SUBSPACE METHOD

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

"In view of all that we have said in the foregoing sections, the many obstacles we appear to have surmounted, what casts the pall over our victory celebration? It is the curse of dimensionality, a malediction that has plagued the scientist from the earliest days."– Richard Bellman

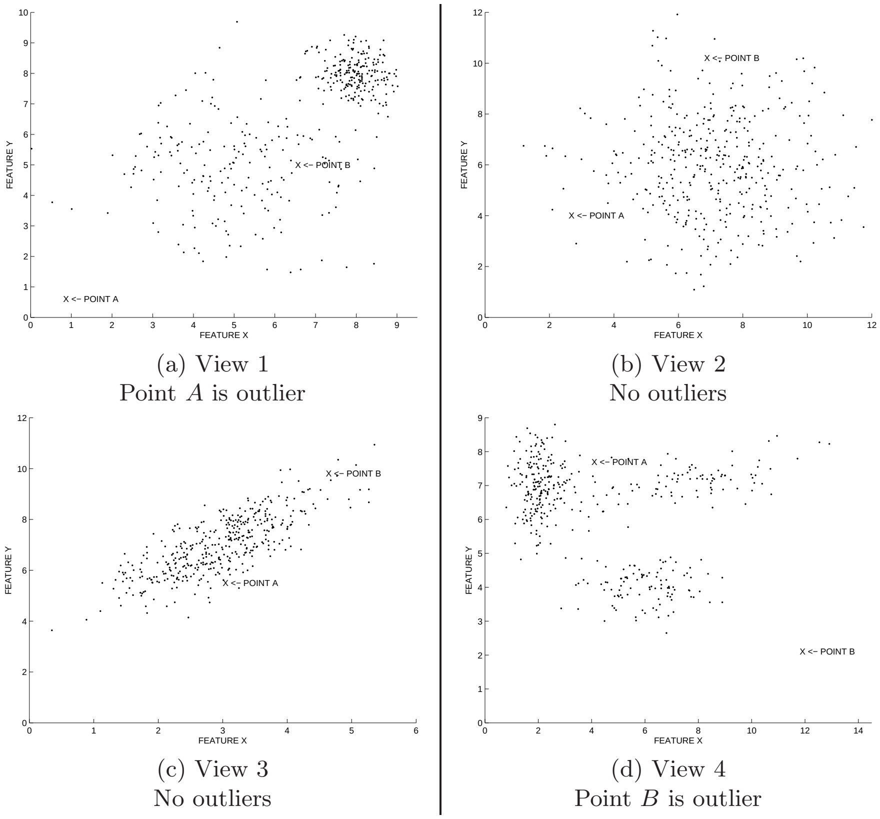
1. Introduction

Many real data sets are very high dimensional. In some scenarios, real data sets may contain hundreds or thousands of dimensions. With increasing dimensionality, many of the conventional outlier detection methods do not work very effectively. This is an artifact of the well known *curse of dimensionality*. In high-dimensional space, the data becomes sparse, and the true outliers become masked by the noise effects of multiple dimensions, when analyzed in *full dimensionality*.

A main cause of the dimensionality curse is the difficulty in defining locality for the high dimensional case. For example, proximity-based methods define locality with the use of distance functions. On the other hand, it has been shown in [65, 215], that all pairs of points are almost equidistant in high-dimensional space. This is referred to as *data sparsity*. Since outliers are defined as data points in sparse regions, this results in a poorly discriminative situation where all data points are situated in an almost equally sparse regions in full dimensionality. The challenges arising from the dimensionality curse are not specific to outlier detection. It is well known that many problems such as clustering and similarity search experience qualitative challenges with increasing dimensionality [5, 7, 95, 215]. In fact, it has been suggested that almost any algorithm which is based on the notion of proximity would degrade qualitatively in higher dimensional space, and would therefore need to re-defined in a more meaningful way [8]. The impact of the dimensionality curse on the outlier detection problem was first noted in [4].

In order to further explain the causes of the ineffectiveness of full dimensional outlier analysis algorithms, a motivating example will be presented. In Figure 1.1, four different 2-dimensional views of a hypothetical data set have been illustrated. Each of these views corresponds to a disjoint set of dimensions. It is evident that point A is exposed as an outlier in the first view of the data set, whereas point B is exposed as an outlier in the fourth view of the data set. However, neither of the data points A and B are exposed as outliers in the second and third views of the data set. These views are therefore *noisy* from the perspective of measuring the outlierness of A and B. In this case, three of the four views are quite non-informative and noisy for exposing any *particular* outlier A or B. In such cases, the outliers are lost in the random distributions within these views, when the distance measurements are performed in *full* dimensionality. This situation is often naturally magnified with increasing dimensionality. For data sets of very high dimensionality, it is possible that only a very small fraction of the views may be informative for the outlier analysis process.

What does the aforementioned pictorial illustration tell us about the issue of locally relevant dimensions? The physical interpretation of this situation is quite intuitive in practical scenarios. An object may have several measured quantities, and significantly abnormal behavior of this object may be reflected only in a small subset of these quantities. For example, in an airplane mechanical fault detection scenario, the results of thousands of different airframe tests on the same plane may mostly be normal, with some noisy variations, which are not significant. On the other hand, some deviations in a small subset of tests may be significant enough to be indicative of anomalous behavior. When the data from the tests are represented in full dimensionality, the anomalous data points will not appear significant in virtually all views of the data, except for a very small fraction of the dimensions. Therefore, aggregate proximity measures are unlikely to expose the outliers, since the noisy variations of the vast number of normal tests will mask the outliers. Furthermore, when different objects (instances of different airframes) are tested, then different tests (subsets of dimensions) may be relevant to finding the outliers, which emphasizes the *local* nature of the relevance.



(a) View 1
Point A is outlier

(b) View 2
No outliers

(c) View 3
No outliers

(d) View 4
Point B is outlier

Figure 1.1. The outlier behavior may be lost in a majority of randomly chosen subspaces in the high dimensional case.

What does this mean for full-dimensional analysis in such scenarios? When full-dimensional distances are used in order to measure deviations, the dilution effects of the vast number of "normally noisy" dimensions will make the detection of outliers difficult. In most cases, this will show up as concentration effects in the distances, from the noise in the other dimensions. This may make the computations more erroneous. Furthermore, the additive effects of the noise present in the large number of different dimensions will interfere with the detection of actual deviations. Simply speaking, *outliers are lost in low-dimensional subspaces, when full-dimensional analysis is used, because of the masking and dilution effects of the noise in full dimensional computations* [4].

Similar effects are also experienced for other distance-based methods such as clustering and similarity search. For these problems, it has been shown [5, 7, 215] that by examining the behavior of the data in subspaces, it is possible to design more meaningful clusters which are specific to the particular subspace in question. This broad observation is generally true of the outlier detection problem as well. Since the outliers may only be discovered in low dimensional subspaces of the data, it makes sense to explore the lower dimensional subspaces for deviations of interest. Such an approach filters out the additive noise effects of the large number of dimensions, and results in more robust outliers.

Such a problem is very challenging to address effectively. This is because the number of possible projections of high dimensional data is exponentially related to the dimensionality of the data. The problem of outlier detection is like finding a needle in a haystack, *even when we know* the relevant dimensions of interest. Being forced to determine the relevant subsets of dimensions *in addition to this challenge* is equivalent to suggesting that even the haystack of interest is hidden in an exponential number of possible haystacks. An important observation is that subspace analysis in the context of the outlier detection problem is generally more difficult than in the case for problems such as clustering, which are based on aggregate behavior. This is because outliers, by definition, are rare, and therefore statistical aggregates on individual dimensions in a given locality often provide *very weak* hints for the subspace exploration process as compared to aggregation-based methods such as clustering. When such weak hints result in the omission of relevant dimensions, the effects can be much more drastic than the inclusion of irrelevant dimensions, especially in the interesting cases when the number of locally relevant dimensions is a small fraction of the full data dimensionality. A common mistake is to assume that the complementarity relationship between clustering and outlier analysis can be extended to the problem of local subspace selection. In particular, blind adaptations of dimension selection methods from earlier subspace clustering methods, which are unaware of the nuances of subspace analysis principles across different problems, may sometimes miss important outliers. In this context, it is also crucial to recognize the difficulty in identifying relevant subspaces for outlier analysis, and use robust methods which combine the results from different subspaces.

An effective outlier detection method would need to search the data points and dimensions in *an integrated way*, so as to reveal the most relevant outliers. This is because different subsets of dimensions may be relevant to different outliers, as is evident from the example in Figure 1.1. The integration of point and subspace exploration leads to a further expansion in the number of possibilities which need to be examined for outlier analysis. This chapter will focus on subspace exploration methods, which attempt to find the relevant outliers by sifting through different subsets of dimensions in the data in an ordered way. This is accomplished simultaneously with a data-specific evaluation process, so that relevant data points are reported as outliers without having to explore all the subspaces in an exhaustive way. The idea is to determine the relevant subsets of dimensions in which the most important outliers are revealed as quickly as possible. This model is referred to as *projected outlier detection* [4]. Correspondingly, this chapter will present a number of algorithms, which achieve this goal.

Several classes of methods are commonly used in order to discover the relevant subspaces:

- **Rarity-based:** These methods attempt to discover the subspaces based on rarity of the underlying distribution. The major challenge here is computational, since the number of rare subspaces is far larger than the number of dense subspaces in high dimensionality.

- **Unbiased:** In these methods, the subspaces are sampled in an unbiased way, and scores are combined across different subspaces.

- **Aggregation-based:** In these methods, aggregate statistics such as cluster statistics, variance statistics, or non-uniformity statistics of local or global subsets of the data are used in order to determine the relevance of subspaces. Note that the difference from rarity-based statistics, is that instead of trying to determine the *number of data points* in a pre-specified local subspace, these methods typically analyze the statistical distributions of pre-specified local or global reference sets of points. Since such methods use statistics over local or global *subsets* of the data, it provides some *hints* for relevant subspaces for exploration. However, since such hints are weak, and are not guaranteed to be the correct ones, multiple subspace sampling is crucial.

This chapter is organized as follows. Evolutionary algorithms for outlier detection are discussed in section 2. These algorithms are based on a grid-based approach for defining outliers. Distance-based methods for subspace outlier detection are studied in section 3. Methods for using and combining multiple subspaces in order to determine relevant outliers are discussed in section 4. The problem of determining outliers in generalized subspaces is discussed in section 5. The limitations of subspace analysis are discussed in section 6. The conclusions and summary are presented in section 7.

2. Projected Outliers with Grids

A first approach to projected outlier detection was presented in [4]. Projected outliers are determined by finding *localized regions of the data in low dimensional space*, which have abnormally low density. Thus, the first step is to identify and mine those *localized* patterns which contain data points, but have abnormally low density. Thus, the goal is to determine interesting anomalies, rather than the noise in the data. Once such localized regions have been identified, then the outliers are defined as those records which have such patterns present in them. An interesting observation is that such lower dimensional projections can be determined even in data sets with missing attribute values. This is quite useful for many real applications, in which feature extraction is a difficult process and full feature descriptions often do not exist. For example, in the airframe fault detection scenario introduced earlier in this chapter, it is possible that only a subset of tests may have been applied, and therefore the values in only a subset of the dimensions may be available for outlier analysis.

2.1 Defining Abnormal Lower Dimensional Projections

In order to find such abnormal lower dimensional projections, it is important to provide a proper statistical definition of an abnormal lower dimensional projection. An abnormal lower dimensional projection is one in which the density of the data is exceptionally lower than average. In this context, the methods for extreme value analysis introduced in Chapter 2 are useful.

A grid-based approach is used in order to determine projections of interest. The first step is to perform a grid discretization of the data. Each attribute of the data is divided into ϕ ranges. These ranges are created on an equi-depth basis. Thus, each range contains a fraction $f = 1/\phi$ of the records. The reason for using equi-depth ranges as opposed to equi-width ranges is that different localities of the data have different densities. Therefore, such an approach partially adjusts for the local variations in data density during the initial phase. These ranges form the units of locality which are used in order to define low dimensional projections which have unreasonably sparse regions.

Consider a k-dimensional cube which is created by picking grid ranges from k different dimensions. The expected fraction of the records in that region is equal to f^k, if the attributes were statistically independent. Of course, the data is far from statistically independent and therefore the actual distribution of points in a cube would differ significantly from average behavior. Many of the local regions may contain very few data points, if any. It is precisely these abnormally sparse regions, which are useful for the purpose of outlier detection.

It is assumed that the total number of points in the database is denoted by N. Under the afore-mentioned independence assumption, the presence or absence of any point in a k-dimensional cube is a bernoulli random variable with probability f^k. Then, the expected fraction and standard deviation of the points in a a k-dimensional cube is given by $N \cdot f^k$ and $\sqrt{N \cdot f^k \cdot (1 - f^k)}$. Furthermore, if the number of data points N is large, then the central limit theorem can be used to *approximate* the number of points in a cube by a normal distribution. Let $n(\mathcal{D})$ be the number of points in a k-dimensional cube $\mathcal{D}$. The sparsity coefficient $S(\mathcal{D})$ of the data set $\mathcal{D}$ can be computed as follows:

$$S(\mathcal{D}) = \frac{n(D) - N \cdot f^k}{\sqrt{N \cdot f^k \cdot (1 - f^k)}}$$

Only sparsity coefficients which are negative are indicative of local projected regions, in which the presence of the points is significantly lower than expected. Since $n(\mathcal{D})$ is assumed to fit a normal distribution, the normal distribution tables can be used to quantify the probabilistic level of significance of its deviation. Of course, while the independence assumption is almost never completely true, it provides a good heuristic for determining the level of abnormality of the underlying data points in practice.

2.2 Evolutionary Algorithms for Outlier Detection

It is evident from the discussion in the introduction, that an exhaustive search of all the subspaces in the data for outliers is unlikely to be fruitful, because of high computational complexity. Therefore, an ordered search method is required, which prunes off most of the subspaces automatically during the exploration process. Since the search space is noisy and unstructured in this case, this is a natural candidate for the use of evolutionary algorithms.

The nature of this problem is such that there are no upward or downward-closed properties on the grid-based subspaces satisfying the sparsity condition.[1] Unlike problems such as frequent pattern mining [28] where one is looking for large aggregate patterns, the problem of finding subsets of dimensions which are sparsely populated has the flavor of finding a needle in haystack. Furthermore, it may often be the case that even though particular regions may be well populated on certain sets of dimensions, they may be very sparsely populated when such dimensions are combined together. For example, in a given data set, there may be a large number of individuals clustered at the age of 20 (low local variance), and a modest number of individuals with varying levels of diabetes (modest local variance). However, *very rare* individuals would satisfy both criteria, because the disease does not affect young individuals. From the perspective of outlier detection, a 20-year old with diabetes is a very interesting record. However, the interestingness of the pattern is not even hinted at by its lower dimensional projections, or the relative variances in these individual projections. Therefore, the best projections are often created by an unknown combination of dimensions, whose lower dimensional projections may contain very few hints for proper subspace exploration. One solution is to change the measure in order to force better closure or pruning properties; however this can worsen the quality of the solution substantially by forcing the choice of the measure to be driven by algorithmic considerations. In general, it is not possible to predict the behavior of the data when two sets of dimensions are combined. Therefore, a natural option is to develop search methods which can identify such hidden combinations of dimensions. In order to search the exponentially increasing space of possible projections, the work in [4] borrows ideas from a class of evolutionary search methods in order to reduce the size of the search space.

Evolutionary Algorithms [223] are methods which imitate the process of organic evolution [125] in order to solve parameter optimization problems. In evolutionary methods, every solution to an optimization problem can be disguised as an individual in an evolutionary system. The

[1] An upward closed pattern is one in which all supersets of the pattern are also valid patterns. A downward closed set of patterns is one in which all subsets of the pattern are also members of the set.

measure of fitness of this "individual" is equal to the objective function value of the corresponding solution, and the other species which this individual has to compete with are a group of other solutions to the problems. Appropriate operations are defined in order to imitate the recombination and mutation processes as well, and the simulation is complete. Each feasible solution is encoded in the form of a string and is the chromosome representation of the solution. The process of conversion of feasible solutions of the problem into strings which the algorithm can use is referred to as its *encoding*. The measure of fitness of a string is evaluated by the *fitness function*. This is equivalent to the objective function value of the solution. The better the objective function value, the better the fitness value. As the process of evolution progresses, all the individuals in the population typically improve in fitness and also become more similar to each other. Dejong [134] defined convergence of a particular position in the string, as the the stage at which 95% of the population had the same value for that gene. The population is said to have converged when all positions in the string representation have converged.

The relevant localized subspace patterns can be easily represented as strings. Let us assume that the grid range for the ith dimension is denoted by m_i. Then, the value of m_i can take on any of the values 1 through ϕ, or it can take on the value $*$, which denotes a "don't care". Thus, there are a total of $\phi + 1$ values that the dimension m_i can take on. Thus, consider a 4-dimensional problem with $\phi = 10$. Then, one possible example of a solution to the problem is given by *3*9. In this case, the ranges for the second and fourth dimension are identified, whereas the first and third are left as "don't cares". The evolutionary algorithm uses the dimensionality of the projection k as an input parameter. Therefore, for a d-dimensional data set, the string of length d will contain k specified position and $(d - k)$ "don't care" positions. The fitness for the corresponding solution may be computed using the sparsity coefficient discussed earlier. The evolutionary search technique starts with a population of p random solutions and iteratively used the processes of selection, crossover and mutation in order to perform a combination of hill climbing, solution recombination and random search over the space of possible projections. The process is continued until the population converges to a global optimum according to the *Dejong convergence criterion*[134]. At each stage of the algorithm, the m best projection solutions (most negative sparsity coefficients) are kept track of. At the end of the algorithm, these solutions are reported as the best projections in the data. The following operators are defined for selection, crossover and mutation:

- **Selection:** The copies of a solution are replicated by ordering them by rank and biasing them in the population in the favor of higher ranked solutions. This is referred to as *rank selection*.

- **Crossover:** The crossover technique is key to the success of the algorithm, since it implicitly defines the subspace exploration process. One solution is to use a uniform two-point crossover in order to create the recombinant children strings. The two-point crossover mechanism works by determining a point in the string at random called the crossover point, and exchanging the segments to the right of this point. However, such a blind recombination process may create poor solutions too often. Therefore, an optimized crossover mechanism is defined. In this case, it is guaranteed that both children solutions correspond to a k-dimensional projection as the parents, and the children typically have high fitness values. This is achieved by examining a subset of the different possibilities for recombination and picking the best among them.

- **Mutation:** In this case, random positions in the string are flipped with a predefined mutation probability. Care must be taken to ensure that the dimensionality of the projection does not change after the flipping process.

At termination, the algorithm is followed by a postprocessing phase. In the postprocessing phase, all data points containing the abnormal projections are reported by the algorithm as the outliers. The approach also provides the relevant projections which provide the *causality* (or intensional knowledge) for the outlier behavior of a data point. Thus, this approach also has a high degree of interpretability in terms of providing the reasoning for *why* a data point should be considered an outlier.

3. Distance-based Subspace Outlier Detection

In these methods, distance-based models are used in lower dimensional subspaces of the data in order to determine the relevant outliers. There are two major variations to the common task.

- In one class of models, the outliers are determined by exploring relevant subspaces.

- In another class of methods, the relevant outlying subspaces for a given data point are determined. This is more useful for providing *intensional knowledge*, for illustrating *why* a specific data point is an outlier.

The second class of methods shares similarities with the approach used in [262] for finding intensional knowledge from distance-based outliers. Both classes of methods will be discussed in subsequent sections.

3.1 Subspace Outlier Degree

A distance-based method for finding outliers in lower dimensional projections of the data is proposed in [273]. In this approach, instead of trying to find local subspaces of abnormally low density over the whole data, a local analysis is provided specific to each data point. For each data point $\overline{X}$, a set of reference points $S(\overline{X})$ are determined, which represent the proximity of the current data point being examined.

Once this reference set $S(\overline{X})$ has been determined, the relevant subspace for $S(\overline{X})$ is determined as the set $Q(\overline{X})$ of dimensions in which the variance is small. The specific threshold is picked as a user-specified fraction of the average dimension-specific variance of the data points in $S(\overline{X})$. Thus, this approach analyzes the statistics of individual dimensions independently of one another during the crucial step of subspace selection, though this may sometimes not be helpful for picking the best subspace projections. The approach of analyzing the distance behavior of individual dimensions for picking the subspace set $Q(\overline{X})$ is a rather naive generalization derived from subspace clustering methods. Unlike data clustering, the effectiveness of subspace outlier methods is almost entirely dependent upon the identification of dimensions containing rare points rather than dimensions with specific kinds of aggregate statistics. In outlier analysis, aggregate data measures such as the dimension-specific variance tell us very little about the subspace behavior of the rare points, and which choices of subspaces are likely to be most relevant for identification of these very unusual points. In some cases such as the example of the young diabetes patient discussed earlier, the unusual behavior is manifested in combinations of dimensions rather than the variances of the individual dimensions. If the absolute variance of a particular dimension such as the diabetes level is not deemed to be sufficiently low, it will not selected in the projection.

In the interesting cases, where the number of relevant dimensions is limited, the negative effects of removing a single relevant dimension can be even more drastic than keeping many irrelevant dimensions. The particularly problematic factor here is that if a mistake is made in subspace selection, there is virtually no chance of recovering from the mistake, when a single subspace is picked for analysis. As we will discuss later, other more insightful techniques in [256, 337] mitigate these impacts by using multiple subspaces for outlier analysis.

The euclidian distance of $\overline{X}$ is computed to the mean of the reference set $S(\overline{X})$ *in the subspace defined by* $Q(\overline{X})$. This is denoted by $G(\overline{X})$. The value of $G(\overline{X})$ is affected by the number of dimensions in $Q(\overline{X})$. The *subspace outlier degree* $SOD(\overline{X})$ of a data point is defined by normalizing this distance $G(\overline{X})$ by the number of dimensions in $Q(\overline{X})$.

$$SOD(\overline{X}) = \frac{G(\overline{X})}{|Q(\overline{X})|}$$

It remains to explain how the reference set $S(\overline{X})$ is generated with the use of distances. This may sometimes turn out to be a challenge, since the concept of proximity is itself hard to define in full dimensional space. Therefore, there is a circularity in using full dimensional distances to pick the reference set. The work [273] uses a shared nearest neighbor approach in order to compute this locality.

This work tries to find the outliers in a *single* subspace of the data, on the basis of local analysis. In practice, the deviations may be hidden in unusual subspaces which are not evident from the 1-d variance statistics of the reference set. Therefore, if the wrong subspace is selected by aggregate analysis, it is quite likely that many outliers may be missed. Furthermore, since the different dimensions in the data may combine to provide unusual results, it is sometimes more helpful to evaluate the locality of a data point in a subspace by examining the data distribution in the entire subspace, rather than examining the different dimensions independently from one another.

3.2 Finding Distance-based Outlying Subspaces

Most of the methods for outlier detection attempt to search for relevant subspaces in order to find outliers. However, some recent methods [499–501] are designed for finding the outlying subspaces *for a given data point*. Thus, the causality in this case is the other way around, where subspaces are determined from points.

A system called *HOS-Miner* was presented in [499]. According to this work, the definition of the outlying subspace for a given data point $\overline{X}$ is as follows:

Definition 1.1 *For a given data point* $\overline{X}$, *determine the set of subspaces such that the sum of its* k*-nearest neighbor distances in that subspace is at least* δ.

This approach does not normalize the distances with the number of dimensions. Therefore, a subspace becomes more likely to be outlying with increasing dimensionality. This definition also exhibits closure properties in which any subspace of a non-outlying subspace is also not outlying. Similarly, every superset of an outlying subspace is also outlying. Clearly, only *minimal* subspaces which are outliers are interesting. The method in [499] uses both downward- and upward-closure properties to prune off subspaces which are either not relevant or not interesting. An X-Tree is used in order to perform the indexing for performing the k-nearest neighbor queries in different subspaces efficiently. It should be noted that while the closure properties result in better efficiency and algorithmic convenience, they do not necessarily imply greater effectiveness. As the earlier example with the young diabetes patient illustrated, true outliers are often hidden in subspaces of the data, which cannot be inferred from their lower or higher dimensional projections.

In order to further improve the efficiency of the learning process, the work in [499] uses a random sample of the data in order to learn about the subspaces before starting the subspace exploration process. This is achieved by estimating a quantity called the *Total Savings Factor (TSF)* of the outlying subspaces. These are used to regulate the search process for specific query points and prune the different subspaces in an ordered way. Furthermore, the TSF values of different subspaces are dynamically updated as the search proceeds. It has been shown in [499] that such an approach can be used in order to determine the outlying subspaces of specific data points efficiently. Numerous methods for using different kinds of pruning properties and genetic algorithms for finding outlying subspaces are presented in [500, 501].

4. Combining Outliers from Multiple Subspaces

One of the major challenges of subspace analysis is that a given data point may show very different behavior in terms of its outlier degree in different subspaces. This also corresponds to the fact that the *outlier scores* from different subspaces may all be very different. These need to be combined into a unified outlier score. This principle is generally related to that of ensemble-analysis, which was discussed in Chapter 1. A variety of methods have been proposed for examining different subspaces for outlier ranking.

4.1 Random Subspace Sampling

The simplest method for combining outliers from multiple subspaces is the use of random subspace sampling. In the work in [289], an approach called *feature bagging* is used, which is analogous to the ensemble technique often used in data classification. This approach also falls in the class of *independent ensembles* introduced in Chapter 1.

The broad approach is to repeatedly apply the following two steps:

- Randomly select between $(d/2)$ and d features from the underlying data set in iteration t in order to create a data set D_t in the tth iteration.
- Apply the outlier detection algorithm O_t on the data set D_t in order to create score vectors S_t.

In principle, the outlier detection algorithm O_t used for the tth iteration could be different. However, the work in [289] uses the LOF algorithm for all the iterations.

At the end of the process, the outlier scores from the different algorithms need to be combined. There are two distinct methods which are used in order to combine the different subspaces:

- *Breadth-first Approach:* In this approach, the ranking of the algorithms is used for combination purposes. The top-ranked outliers over all the different executions are ranked first, followed by the second-ranked outliers (with repetitions removed), and so on. Minor variations could exist because of tie-breaking between the outliers within a particular rank.
- *Cumulative Sum Approach:* The outlier scores over the different algorithm executions are summed up. The top ranked outliers are reported on this basis.

It was shown in [289] by synthetic data analysis, that combining methods are important when some of the features are noisy. In such cases, full-dimensional algorithms are unable to distinguish the true outliers from the normal data, because of the additional noise. Improvements over the base LOF-approach were also observed with the use of real-data analysis. At first sight, it would seem that random subspace sampling [289] does not attempt to optimize the discovery of subspaces to finding rare instances at all. Nevertheless, it does have the paradoxical merit that it is relatively efficient to sample subspaces, and therefore a large number of subspaces can be sampled in order to improve robustness. The robustness resulting from multiple subspace sampling is clearly a very desirable quality, as long as the combination function at the end recognizes the differential behavior of different subspace samples for a given data point. In a sense, this approach implicitly recognizes the difficulty of detecting relevant and rare subspaces for the outlier detection problem, and therefore approaches the problem by sampling as many subspaces as possible in order to reveal the rare behavior. From a conceptual perspective, this approach is similar to that of harnessing the power of many weak learners to create a single strong learner in classification problems. The approach has been shown to show consistent performance improvement over full dimensional methods for many real data sets in [289]. This approach may also be referred to as the *feature bagging method* or *random subspace ensemble method*. This approach is likely to have significant potential for improving subspace analysis, by experimenting with different choices of combination functions.

The work in [310] designs the concept of *isolation forest*, which derives its motivation from another ensemble technique known as *random forests*, which are commonly used in classification. In this case, the data is recursively partitioned by axis-parallel cuts along randomly selected attributes, so as to isolate different kinds of instances from one another. In such cases, the tree branches containing outliers are noticeably less deep, because these data points are quite different from the normal data. Thus, data points which have noticeably shorter paths in the branches of different trees are more likely to be outliers. The different branches correspond to different local subspace regions of the data, depending on how the attributes are selected for splitting purposes. The smaller path methods correspond to lower dimensionality of the subspaces in which the outliers have been isolated. The final combination step is performed by using the path lengths of the data points in the different samples. One major challenge of using such an approach is that when the dimensionality of the data increases, an incorrect choice of attribute for splitting at the higher levels of the tree is more likely to mislead the detection approach. Nevertheless, the approach is efficient in determining each subspace sample, and the use of multiple subspace samples is a desirable quality of the approach.

4.2 Selecting High Contrast Subspaces

The subspace ensemble method [289] discussed in the last section randomly samples subspaces. If many dimensions are noisy, at least a few of them are likely to be included in each subspace sample. This implies that a larger number of subspace samples will be required in order to obtain more robust results. Therefore, it is natural to ask whether it is possible to perform a pre-processing in which a smaller number of *high-contrast* subspaces are selected.

In the work proposed in [256], the outliers are found only in these high-contrast subspaces, and the corresponding scores are combined together. Thus, this approach decouples the subspace search as a a generalized pre-processing approach from the outlier ranking of the individual data points. The approach discussed in [256] is quite interesting because of its pre-processing approach to finding relevant subspaces in order to reduce the irrelevant subspace exploration. While the high contrast subspaces are obtained using aggregation-based methods, the aggregation behavior is only used as hints in order to identify multiple subspaces for greater robustness. The assumption here is that rare events are *statistically more likely* to occur in subspaces where there is significant non-uniformity and contrast. The final outlier score combines the results over different subspaces. The insight in the work of [256] is to combine subspace selection and multiple subspaces analysis in order to determine the relevant outlier scores. Therefore, the risk of not picking the correct subspace is reduced. This approach has been shown to work well in [256] over the random subspace sampling method.

The conditional probability for an an attribute value along any particular dimension $P(x_1|x_2 \ldots x_d)$ is the same as its unconditional probability $P(x_1)$ for the case of uncorrelated data. High-contrast subspaces are likely to violate this assumption because of non-uniformity in data distribution. In our earlier example of the young diabetes patients, this corresponds to the unexpected rarity of the *combination* of youth and the disease. The idea is that subspaces with such unexpected non-uniformity are more *likely* to contain outliers, though it is treated only as a weak hint for pre-selection of one of multiple subspaces.

A variety of tests based on the student's t-distribution can be used in order to measure the deviation of this sample from the basic hypothesis of independence. This provides a measure of the non-uniformity of the subspace, and therefore provides a way to measure the quality of the subspaces in terms of their propensity to contain outliers. A bottom-up *Apriori* style [29] approach was proposed in order to determine the relevant projections. In this approach the subspaces are continuously extended to higher dimensions for testing. Details of the approach are available in [256].

4.3 Local Selection of Subspace Projections

The work in [337] uses *local* statistical selection of relevant subspace projections in order to determine outliers. In other words, the selection of the subspace projections is optimized to specific data points, and therefore the locality of a given data point matters in the selection process. For each data point $\overline{X}$, a set of subspaces is identified, which are considered *high contrast* subspaces from the perspective of outlier detection. However, this exploration process uses the high contrast behavior as statistical *hints* in order to explore *multiple* subspaces for robustness, since a single subspace may often miss the true projection.

```
Algorithm $OUTRES$(Data Point: $\overline{X}$
            Subspace: $S$);
begin
  for each attribute $i$ not in $S$
  if $S_i = S \cup \{i\}$ passes non-uniformity test then
  begin
    Compute $OS(S_i, \overline{X})$;
    $O(\overline{X}) = OS(S_i, \overline{X}) \cdot O(\overline{X})$;
    $OUTRES(\overline{X}, S_i)$;
  end
end
```

Figure 1.2. The OUTRES Algorithm

The $OUTRES$ method [337] examines the density of lower dimensional subspaces in order to identify relevant projections. The basic hypothesis, is that for a given data point $\overline{X}$ it is desirable to determine subspaces in which the data is sufficiently non-uniformly distributed in its locality. In order to characterize the distribution of the locality of a data point, the work in [337] computes the density of the locality of data point $\overline{X}$ in subspaces S as follows:

$$den(S, \overline{X}) = |\mathcal{N}(\overline{X}, S)| = |\{\overline{Y} : dist(\overline{X}, \overline{Y} \leq \epsilon\}|$$

This is the simplest possible definition of the density, though other more sophisticated methods such as kernel density estimation [409] are used in $OUTRES$ in order to obtain more refined results. Kernel density estimation is also discussed in Chapter 4. A major challenge here is in comparing the subspaces of varying dimensionality. This is because the density of the underlying subspaces reduces with increasing dimensionality. It has been shown in [337], that it is possible to obtain comparable density estimates across different subspaces of different dimensionalities, by selecting the bandwidth of the density estimation process according to the dimensionality of the subspace.

Furthermore, the work in [337] uses statistical techniques in order to meaningfully compare different subspaces. For example, if the data is uniformly distributed, then the number of data points lying within a distance ϵ of the data point should be regulated by the fractional volume of the data in that subspace. Specifically, the fractional parameter defines a binomial distribution characterizing the number of points in that volume, if that data were to be uniformly distributed. Of course, one is really interested in subspaces which deviate significantly from this behavior. The (local) relevance of the subspace for a particular data point $\overline{X}$ is computed using statistical testing. The two hypothesis are as follows:

- Hypothesis H_0: The local subspace neighborhood $\mathcal{N}(\overline{X}, S)$ is uniformly distributed.
- Hypothesis H_1: The local subspace neighborhood $\mathcal{N}(\overline{X}, S)$ is not uniformly distributed.

The Kolmogorov-Smirnoff goodness of fit test [424] is used to determine which of the afore-mentioned hypothesis are true. It is important to note that this process provides an idea of the *usefulness* of a subspace, and is used in order to enable a *filtering condition* for removing irrelevant subspaces from the process of computing the outlier score of a specific data point. A subspace is defined as relevant, if it passes the hypothesis condition H_1. In other words, outlier scores are computed using a combination of subspaces which *must* satisfy this relevance criterion.

In order to combine the scores which are obtained from multiple *relevant* subspaces, the work in [337] uses the product of the outlier scores obtained from different subspaces. Thus, if $S_1 \ldots S_k$ be the different abnormal subspaces found for data point $\overline{X}$, and if $O(S_i, \overline{X})$ be the outlier score from subspace S_i, then the overall outlier score $OS(\overline{X})$ is defined as follows:

$$OS(\overline{X}) = \prod_i O(S_i, \overline{X})$$

It is evident that *low scores* represent a greater tendency to be an outlier. The advantage of using the product over the sum, is that the latter is dominated by the high scores, as a result of which a few subspaces containing normal behavior will dominate the sum. On the other hand, in the case of the product, the outlier behavior in a small number of subspaces will be greatly magnified. This is particularly appropriate for the problem of outlier detection. So far, it has not been discussed, how the actual subspaces $S_1 \ldots S_k$ are determined. This will be achieved with a careful subspace exploration.

In order to actually define the outlier score, subspaces are considered significant for particular objects only if their density is at least two standard deviations less than the mean value. This is essentially a filter condition for that subspace to be considered deviant. Thus, the deviation $dev(\overline{X}, S_i)$ of the data point $\overline{X}$ in subspace S_i is defined as the ratio of the deviation of the density of the object from the mean density, divided by two standard deviations.

$$dev(S_i, \overline{X}) = \frac{\mu - den(S_i, \overline{X})}{2 \cdot \sigma}$$

The outlier score of a data point in a subspace is the ratio of the density of the point in the space to its deviation, if it satisfies the filter condition of the density being at least two standard deviations less than the mean. Otherwise the outlier score is considered to be 1, and it does not affect the overall outlier score in the product function defined earlier for combining different subspaces. Thus, for the points satisfying the filter condition, the outlier score $OS(S_i, \overline{X})$ is defined as follows:

$$O(S_i, \overline{X}) = \frac{den(S_i, \overline{X})}{dev(S_i, \overline{X})}$$

An observation in [337] is that subspaces which are either very low dimensional (eg. 1-d subspaces) or very high dimensional are not very informative from an outlier detection perspective. A recursive exploration of the subspaces is performed, where an additional attribute is included in the subspace for statistical testing. Therefore, the work in [337] uses recursive processing in which the subspaces are built in recursive fashion. When an attribute is added to the current subspace S_i, the non-uniformity test is utilized to determine whether or not that subspace should be used. Otherwise, this subspace is discarded.

The overall algorithm uses a recursive subspace exploration procedure in order to measure the outlierness of any particular object. Note that the entire recursive algorithm uses the data point $\overline{X}$ as input, and therefore the procedure needs to be applied separately *for each data point.* For any given subspace, an attribute is incrementally added. Then, the non-uniformity test is applied to determine if it is relevant. If it is not relevant, then the subspace is discarded. Otherwise, the outlier score $O(S_i, \overline{X})$ in that subspace is computed for the data point, it is multiplied with the current value of $OS(\overline{X})$. Since the outlier scores of subspaces, which do not meet the filter condition are set to 1, they do not affect the density computation in this multiplicative approach. The procedure is then recursively called in order to explore the next subspace. Thus, such a procedure potentially explores an exponential number of subspaces, though the real number is likely to be much smaller in practice. This is because of the non-uniformity test, which prunes off large parts of the recursion tree during the exploration. The overall algorithm for subspace exploration for a given data point $\overline{X}$ is illustrated in Figure 1.2.

5. Generalized Subspaces

A significant amount of success has been achieved for finding outliers in axis-parallel subspaces in recent work. While these methods are effective for finding outliers in cases where the outliers naturally deviate in

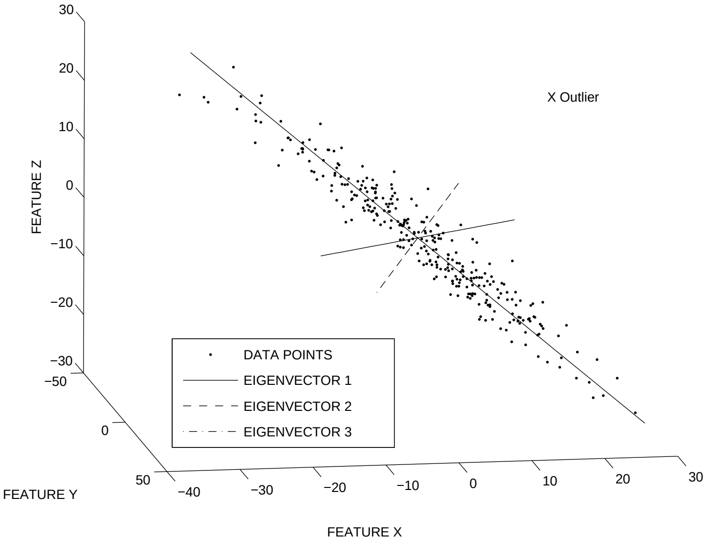


Figure 1.3. Global PCA can discover outliers in cases, where the entire data is aligned along lower dimensional manifolds.

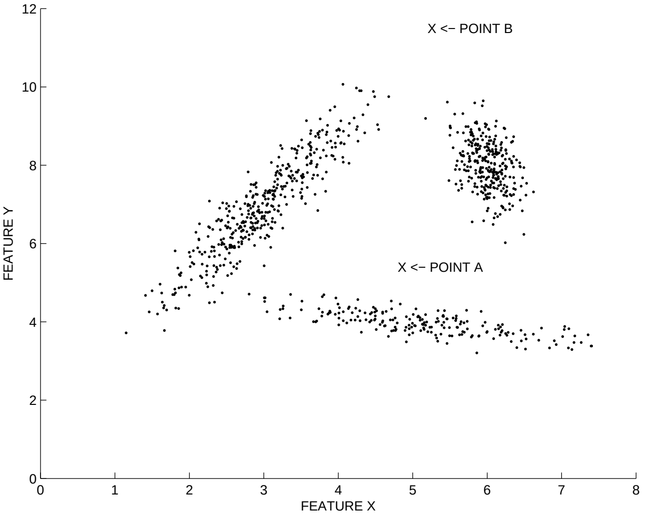


Figure 1.4. Outliers are best discovered by determining deviations from local PCA-based clusters. Neither axis-parallel subspace outliers nor global-PCA can capture such clusters.

specific subspaces from the clusters, they are not very useful for finding clusters in cases where the points are aligned along lower-dimensional manifolds of the data. For example, in the case of Figure 1.4, no 1-dimensional subspace analysis from the 2-dimensional data can find the outliers. On the other hand, it is possible to find *localized* 1-dimensional correlated subspaces so that most of the data aligns along these localized 1-dimensional subspaces, and the remaining deviants can be classified as outliers.

These algorithms are generalizations of the following two classes of algorithms:

- The PCA-based linear models discussed in Chapter 3 find the *global* regions of correlation in the data. For example, in the case of Figure 1.3, the outliers can be effectively identified by determining these global directions of correlation. However, no such *global* directions of correlation exist in the case of Figure 1.4.

- The axis-parallel subspace outliers discussed earlier in this chapter can find deviants, when the data is naturally aligned along low dimensional axis-parallel subspace clusters. However, this is not the case in Figure 1.4, where the data is aligned along arbitrary directions of correlation.

This problem can be partially addressed with the use of generalized projected clustering methods, where the clusters are determined in arbitrarily aligned subspaces of the data [7]. The method discussed in [7] has a built-in mechanism in order to determine the outliers *in addition to* the clusters. Such outliers are naturally data points which do not align with the clusters. However, the approach is not particularly optimized for finding the outliers, because the primary purpose of the method is to determine the clusters. The outliers are discovered as a side-product of the clustering algorithm, rather than as the primary goal. Therefore, the approach may discover the weaker outliers, which correspond to the noise in the data. Similarly, the approach in [132] is focussed on determining the noise in the data for improving mixture modeling of probabilistic PCA algorithms. In order to determine the outliers which are optimized to the locality of a particular data point, it is critical to determine localized subspaces which are optimized to the data point $\overline{X}$, which is being evaluated for its outlier score. The determination of such subspaces is non-trivial, since it often cannot be inferred from locally aggregate properties of the data, for detecting the behavior of *rare* instances.

Another method was recently proposed in [274] for finding outliers in generalized subspaces of the data. The main difference from earlier generalized subspace clustering methods is that local reference sets are used for local correlation analysis. For a given data point $\overline{X}$, this method finds the full-dimensional k-nearest neighbors of $\overline{X}$. This provides a reference set S with mean vector $\overline{\mu}$. The PCA approach of Chapter 3 is applied to the covariance matrix $\Sigma(S)$ of the *local* reference set S in order to determine the key eigenvectors $\overline{e_1} \ldots \overline{e_d}$, in increasing order of variance, with corresponding eigenvalues $\lambda_1 \leq \lambda_2 \ldots \leq \lambda_d$. The discussion in the Appendix performs these same steps [406] except that they are performed on a *global* basis, rather than on a local reference set S. Even if all d dimensions are included, it is possible to create a normalized outlier score of a data point $\overline{X}$, to the centroid $\overline{\mu}$ of the data with the use of local eigenvalue scaling, as discussed in Chapter 3:

$$Score(\overline{X}) = \sum_{j=1}^{d} \frac{|(\overline{X} - \overline{\mu}) \cdot \overline{e_j}|^2}{\lambda_j} \tag{1.1}$$

As discussed in earlier work [406], can be approximately modeled as a χ^2 distribution with d degrees of freedom for each data point, and the outlier scores of the different data points can be reasonably compared to one another. Such an approach is used in [406] in the context of global data analysis. The survey paper of Chandola et al. [107] provides a simpler exposition. The work in [274] uses a similar approach with the use of a local reference set, selected with the use of full dimensional k-nearest neighbor distances.

Eigenvectors with large values of λ_i will usually not contribute much to the score, though as discussed below, this may not always be the case. Such directions are pruned from the score. The δ eigenvectors[2] with the smallest eigenvalues are picked for the computations above. Correspondingly, the pruned score is defined on the basis of the first $\delta \leq d$ eigenvectors only with the smallest eigenvalues.

$$Score(\overline{X}, \delta) = \sum_{j=1}^{\delta} \frac{|(\overline{X} - \overline{\mu}) \cdot \overline{e_j}|^2}{\lambda_j} \tag{1.2}$$

How should the value of δ be determined for a particular data point $\overline{X}$? The score is a χ^2-distribution with δ-degrees of freedom. It was observed in [274] that the value of δ can be parameterized, by treating the χ^2 distribution as a special case of the Γ distribution.

$$Score(\overline{X}, \delta) \sim \Gamma(\delta/2, 2)$$

[2] The work in [274] uses δ as the number of *longest* eigenvectors, which is only a notational difference, but is noted here to avoid confusion.

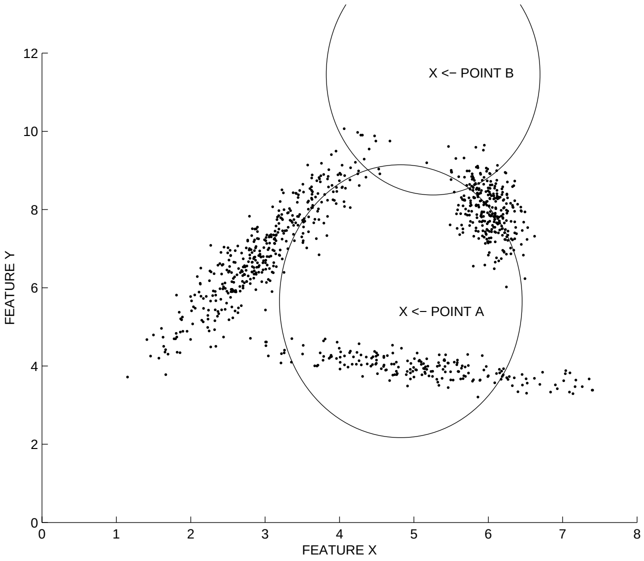


Figure 1.5. Local reference set may sometimes contain points from multiple generating mechanisms

The optimal value of δ is picked specifically for each data point, by picking the value of δ in order to determine the maximal unlikely deviation based on this model. This is done by using the cumulative density function of the aforementioned distribution. While this value can be directly used as an outlier score, it was also shown in [274], how this score may be converted into a more intuitive probability value.

This approach has several issues:

- A *single subspace* has been used by this approach for finding the outliers with the use of the local reference set S. If the local reference set S is not accurately determined, then this will not provide the proper directions of local correlation. The use of a single subspace is risky, especially with the use of weak aggregation-based hints, because it is often possible to unintentionally remove relevant subspaces. This can have drastic effects. The use of multiple subspaces may be much more relevant in such scenarios, such as the methods proposed in [289, 256, 337, 341].

- There is an inherent circularity in identifying the reference set with the use of full dimensional k-nearest neighbor distances, especially if the distances are not meaningfully defined in full dimensionality. The choice of points in the reference set and the choice of the subspace clearly impact each other in a circular way. This is a classical "chicken and egg" problem in subspace analysis, which was first pointed out in [5]. The analysis in such cases needs to be *simultaneous* rather than *sequential.* As is well known, the most robust techniques for handling circularity in virtually all problem domains (eg. the EM algorithm and many projected clustering methods) use iterative methods, so that the point-specific and dimension-specific aspects of the problem are able to interact with one another. This is however, not the case in [274], where a sequential analysis is used.

In particular, it may happen that many locally irrelevant features may be used during the determination of the local reference set, when full dimensional distances are used. This set could therefore contain data points from multiple generating mechanisms, as illustrated in Figure 1.5. When the number of irrelevant features is unknown, a specific number of points in the reference set will not be able to avoid this problem. The use of a smaller reference set size can reduce the chance of this happening to some extent, but can never guarantee it, especially when many irrelevant features are used. On the other hand, reducing the reference set size can also result in a correlation hyperplane, whose eigenvalue statistics overfit an artificially small set of reference points.

- An interesting question arises, as to whether it is necessary to select a particular set of dimensions in a hard way, since the eigenvalues in the denominator of Equation 1.1 already provide a soft weighting to the importance (or relevance) of the different dimensions. For example, if for a large value of λ_i, a data point shows even larger deviations along that direction, such an outlier would either be missed by dimension pre-selection, or would include other less relevant dimensions. An example is the outlier B in Figure 1.5, which is aligned along the longer eigenvector, and therefore the longest eigenvector is the *most informative* about its outlier behavior. In particular, the method of picking the δ smallest eigenvectors implicitly assumes that the relevance of the attributes are ordered by eigenvalue magnitude. While this may generally be true for aggregation-based clustering algorithms, it is very often not true in outlier analysis because of the unusual nature of outliers. The possibility of outliers aligning along long eigenvectors is not uncommon at all, since two highly correlated attributes may often show highly deviant behavior of a similarly correlated nature. This example also shows, how *brittle* the rare nature of outlier analysis is to aggregation-based measures. This is because of the varying causes of rarity, which cannot be fully captured in aggregation statistics. This is relevant to our discussion in the introduction section, that straightforward generalizations of subspace selection methods from clustering (based on aggregates), are often not appropriate or optimized for (the rare nature of) outlier analysis. One advantage of using all the dimensions is that it reduces to a local Mahalanobis distance with the same dimensionality, and allows better comparability in the scores across different outliers. In such cases, intuitive probability values may be derived more simply from the $\chi^2(d)$ distribution.

The high dimensional case is an extremely difficult one, and it is understandable that no given method will be able to solve these problems perfectly. It should also be pointed out that the iterative EM algorithm discussed in Chapter 2 will be able to discover the local directions of correlation along with outliers which have low fit value to the model. These may sometimes include weak outliers, which are not always interesting. Given that direct discovery of optimal subspaces in a given locality is much more difficult in outlier analysis, a possible line of work would be to use a two-phase approach of first finding the weak outliers, and then determining the strong ones among them by more detailed analysis. For example, it may be possible to use this pre-filtered set of weak outliers for intensive ensemble-based subspace exploration. Combining pre-filtered data points with pre-filtered high-contrast subspaces may provide an interesting direction of future exploration. A significant scope still exists for further improvement of the techniques designed in this area.

6. Discussion of Subspace Analysis

While subspace outlier analysis seems to be the only meaningful method for high dimensional outlier detection, the approach faces a number of challenges, a lot of which are computational in nature. In the high-dimensional case, a small number of deviant subspaces may remain hidden out of a large number of possibilities. This can create unprecedented challenges for outlier analysis. The combinatorial nature of the problem necessitates the design of more efficient algorithms which can perform an ordered exploration of these spaces. In spite of the recent advances in the literature, the design of efficient algorithms for the high dimensional subspace exploration scenario remains a challenge. This is of course an inherent property of high-dimensional data, in which the curse of dimensionality impacts the results both from a qualitative and efficiency perspective.

The second challenge arises from the fact that a subspace exploration technique reports a number of different possibilities for the projections. In such cases, it remains a challenge to combine the results from these deviant subspaces, and rank the resulting outliers effectively. This is of course an opportunity as well, since the results from multiple subspaces may provide more robust outliers. Therefore, significant advancements are required in *ensemble analysis* for outlier detection.

It has been claimed in [514] as an apparently new insight, that the major reason for difficulty in high dimensional outlier analysis is not the concentration of distances, but the masking effects of the locally noisy and irrelevant nature of some of the dimensions, and that the literature has failed to discuss the impact of locally relevant dimensions. This is an incorrect assertion, since both the aspects of local feature selection (relevance) and distance concentration have been studied extensively in the literature. While it is true that noisy and irrelevant attributes mask the outliers, the observation is certainly not new, and the two factors of distance concentration and local feature relevance are closely related. The original work in [4] (and virtually every other subsequent work [289, 256, 337] on this topic) provides a pictorial illustration and a fairly detailed discussion of how (locally) irrelevant attributes mask outliers in different feature-specific views of the data. As stated in [4]: "*... by using full dimensional distance measures it would be difficult to determine outliers effectively because of the averaging behavior of the noisy and irrelevant dimensions. Furthermore, it is impossible to prune off specific features a-priori, since different points may show different kinds of abnormal patterns, each of which use different features or views.*" The ineffectiveness of *global* feature selection in high dimensional data in fact forms the motivating reason for subspace analysis, which can be considered a *local* feature selection method, or a *local* dimensionality reduction method [7, 95]. These connections of local subspace analysis to the ineffectiveness of global feature selection in high dimensional data were explicitly discussed in detail in the motivational discussion of one of the earliest works on subspace analysis [5]. At this point, these results are well known and established[3] wisdom. While it is possible to reduce the distance concentration effects by carefully calibrating the fraction of informative dimensions, such cases are (usually) not interesting for subspace analysis.

[3] Some of the earliest methods even refer to these classes of techniques as local dimensionality reduction [95] in order to emphasize the enhanced and differential local feature selection effect, which arises as a result of different generating mechanisms.

Distance concentration and (too many) irrelevant attributes are closely related. The interesting cases for subspace analysis (typically) show some levels of both properties. Even limited levels of distance concentration impact the effectiveness of full dimensional distance-based algorithms, and this impact is therefore important to examine in outlier analysis. It should be noted that noisy and irrelevant attributes are more likely to lead to concentration of distances. For example, for the case of uniformly distributed data, where all attributes are noisy, the concentration effect is extreme, and an outlier deviating along *a relatively small number of dimensions* will be hard to discover by full dimensional methods. In such cases, from a full dimensional distance-based or density-based perspective, all data points have almost equally good outlier scores, and this can be equivalently understood in terms of *either* locally irrelevant features or distance concentration effects. Of course, real data sets are not uniformly distributed, but *both* irrelevant features and concentration effects are present to varying degrees in different data sets. The general assumption for subspace analysis is that the addition of more dimensions often does not add *proportionally* more information for a particular outlier. The challenging outliers are often defined by the behavior of a small number of dimensions, and when the point-specific information does not increase substantially with data dimensionality, even modest concentration effects will have a negative impact on full dimensional algorithms. The more the number of irrelevant attributes, the more erroneous the computations for full-dimensional distance-based methods. An extreme example at the other end of the spectrum is where an outlier shows informative and deviant behavior in every dimension, and therefore outlier characteristics grow *stronger* with increasing dimensionality. However, in this rather uninteresting case, since the outlier shows *both* many relevant features *and* also typically does not conform to the distance concentration behavior of the remaining data, a trivial full dimensional distance-based algorithm would find it easily in most cases. In general, cases where the informative dimensions also increase significantly with data dimensionality, are not as interesting for subspace analysis because the full dimensional masking behavior becomes less prominent in this easier case. Subspace analysis does not exclude the possibility that the more obvious deviants may also be found by full dimensional analysis.

Outliers, by their very rare nature, may often be hidden in small combinations of dimensions in a high dimensional data set. Subspace analysis is interesting for such scenarios. On the other hand, when more dimensions do add (significantly) more information, then this becomes an easy case for analysis, which no longer remains interesting. In the former case, the vast majority of noisy dimensions make all data points appear as outliers from a density-based or data sparsity perspective.

To summarize, subspace outlier analysis is one of the most challenging problems because of the rare and unusual nature of outliers. In order to design meaningful algorithms, the following principles need to be kept in mind.

- Aggregation-based methods for subspace analysis only provide very weak hints for outlier analysis as compared to clustering algorithms. A direct exploration of rare regions is possible, though it is computationally challenging because of combinatorial explosion [4]. As a result, it becomes necessary to use heuristic methods.

- Aggregation-based methods may be usable, if caution is utilized in recognizing the fact that a given subspace derived from such methods may not always include the relevant dimensions. Exclusion of relevant dimensions has more drastic effects than inclusion of many irrelevant dimensions. Where possible, subspace ensembles should be used in order to combine the weak hints derived from the different subspaces, if aggregation-based measures are used.

- The individual component of an ensemble should be designed with efficiency considerations. This is because the ability to execute the individual component more number of times within a fixed time frame, eventually provides more robustness.

7. Conclusions and Summary

Subspace methods for outlier detection are used in cases, where the outlier tendency of a data point is diluted by the noise effects of a large number of locally non-informative dimensions. In such cases, the outlier analysis process can be sharpened significantly by searching for subspaces in which the data points deviate significantly from the normal behavior. The earliest work on subspace outlier detection used evolutionary search methods in order to determine abnormal lower dimensional projections of the data. A number of subsequent methods have also been designed for determining multiple relevant subspaces for a candidate outlier, and then combining the results from different subspaces in order to create a more robust ensemble-based ranking. It is also possible to determine the outliers in arbitrarily oriented subspaces of the data. Such methods are able to exploit the local correlations in the data in order to determine relevant outliers.

Outlier analysis is the most difficult problem among all classes of subspace analysis problems. This difficulty arises out of the rare nature of outliers, which makes direct statistical analysis more difficult. Since subspace analysis and local feature selection are related, it is noteworthy that even for global feature selection, there are few known methods for outlier analysis, as compared to clustering and classification algorithms. The reason is simple: enough statistical evidence is often not available for the analysis of rare characteristics. Robust statistics is all about *more* data, and outliers are all about *less* data and statistical non-conformity with most of the data! Regions and subspaces containing statistical conformity tell us very little about the complementary regions of non-conformity in the particular case of high-dimensional subspace analysis, since the *potential* domain of the latter is much larger than the former. In particular, a local subspace region of the greatest aggregate conformity does not necessarily reveal anything about the rare point with the greatest statistical non-conformity.

While it is doubtful that the more difficult variations of the problem will ever be fully solved, or will work completely in all situations, it may be possible to design methods which work in many important scenarios. There are many merits in being able to design such methods, because of the numerous insights they can provide in terms of identifying the causes of abnormality. The main challenge is that outlier analysis is so brittle, that it is often impossible to make confident assertions about inferences drawn from aggregate data analysis. The issue of efficiency seems to be closely related to that of effectiveness in high dimensional outlier analysis. This is because the search process for outliers is likely to require exploration of multiple local subspaces of the data in order to ensure robustness. With increasing advances in the computational power of modern computers, there is as yet hope that this area will become increasingly tractable for analysis.

8. Bibliographic Survey

In the context of high-dimensional data, there are two distinct lines of research, one of which investigates the *efficiency* of high dimensional outlier detection [46, 185, 467], and the other investigates the more fundamental issue of the *effectiveness* of high dimensional outlier detection [4, 273]. Unfortunately, the distinction between these two lines of work is sometimes blurred in the literature, even though these are clearly different lines of work with very different motivations. It should be noted that the methods discussed in [46, 185, 467] are all *full dimensional methods*, because outliers are defined on the basis of their full dimensional deviation. While the method of [467] uses projections for indexing, this is used only as an approximation to improve the efficiency of the outlier detection process.

In the high-dimensional case, the efficiency of (full dimensional) outlier detection also becomes a concern, because most outlier detection methods require repeated similarity search in high dimensions in order to determine the nearest neighbors. The efficiency of these methods degrades because of two factors: (i) the computations now use a larger number of dimensions, and (ii) the effectiveness of pruning methods and indexing methods degrades with increasing dimensionality. The solution to these issues still remains unresolved in the vast similarity search literature. Therefore, it is unlikely that *significantly* more efficient similarity computations could be achieved in the context of high dimensional outlier detection, though some success has been claimed for improving the efficiency of high dimensional outlier detection in methods proposed in [46, 185, 467]. On the whole, it is unclear how these methods would compare to the vast array of techniques available in the similarity search literature for indexing high dimensional data. This chapter does *not* investigate the efficiency issue at all, because the efficiency of a *full dimensional* outlier detection technique is not important, if it does not even provide meaningful outliers. Therefore, the focus of the chapter is on methods which *re-define* the outlier detection problem in the context of lower dimensional projections. It is also noted that an angle-based outlier detection for high-dimensional data has been proposed in [269], though this method has been discussed in the chapter on extreme value analysis (Chapter 2), since this method is not a subspace exploration technique. It is also designed to find specific kinds of outliers which lie at the boundaries of the multivariate data, and is much closer in principle to other multivariate extreme value analysis methods such as depth-based and deviation-based methods.

The problem of subspace outlier detection was first proposed in [4]. In this paper, an evolutionary algorithm was proposed to discover the lower dimensional subspaces in which the outliers may exist. The method for distance-based outlier detection with subspace outlier degree was proposed in [273]. Another distance-based method for subspace outlier detection was proposed in [346]. Some methods have also been proposed for outlier analysis by randomly sampling subspaces and combining the scores from different subspaces [289, 310]. In particular, the work in [289] attempts to combine the results from these different subspaces in order to provide a more robust evaluation of the outliers. These are essentially *ensemble-based* methods, which attempt to improve detection robustness by bagging the results from analyzing different sets of features. The major challenge of these methods is that random sampling may not work very well, when the outliers are hidden in specific subspaces of the data. The work in [256] can be considered a generalization of the broad approach in [289], where only high contrast subspaces are selected for the problem of outlier detection.

The reverse problem of finding outlying subspaces *from* specific points was studied in [499–501]. In these methods, a variety of pruning and evolutionary methods were proposed in order to speed up the search process for outlying subspaces. The work in [47] also defines the exceptional properties of outlying objects both with respect to the entire population (global properties), and also with respect to particular sub-populations to which it belongs (local properties). Both these methods provide different but meaningful insights about the underlying data. A genetic algorithm for finding the outlying subspaces in high dimensional data is provided in [500]. In order to speed up the fitness function evaluation, methods are proposed to speed up the computation of the k-nearest neighbor distance with the use of bounding strategies. A broader framework for finding outlying subspaces in high dimensional data is provided in [501]. A method which uses two-way search for finding outlying subspaces is proposed in [482]. In this method, full dimensional methods are first used to determine the outliers. Subsequently, the key outlying subspaces from these outlier points are detected and reported. A method for using rules in order to explain the context of outlier objects is proposed in [340].

A number of ranking methods for subspace outlier exploration have been proposed in [337–339]. In these methods, outliers are determined in multiple subspaces of the data. Different subspaces may either provide information about different outliers, or about the same outliers. Therefore, the goal is to combine the information from these different subspaces in a robust way in order to report the final set of outliers. The *OUTRES* algorithm proposed in [337] uses recursive subspace exploration in order to determine all the subspaces relevant to a particular data point. The outlier scores from these different subspaces are combined in order to provide a final value. A tool-kit for ranking subspace outliers has been presented in [338]. A more recent method for using multiple views of the data for subspace outlier detection is proposed in [341]. Methods for subspace outlier detection in multimedia databases were proposed in [51].

Most of the methods for subspace outlier detection perform the exploration in axis-parallel subspaces of the data. This is based on the complementary assumption that the dense regions or clusters are hidden in axis-parallel subspaces of the data. However, it has been shown in recent work that the dense regions may often be located in arbitrarily oriented subspaces of the data [7]. While it has been shown in earlier work that the removal of noise (or weak outliers) improves the effectiveness of generalized subspaces clustering algorithms [7], specific techniques are also required in order to determine outliers in a way which is optimized to the data correlations. Another work in [274] provides an arbitrarily oriented solution for the generalized outlier analysis problem, which extends the correlation-analysis approach proposed in [7] to a method based on local reference sets rather than clusters.

Recently, the problem of outlier detection has also been studied in the context of dynamic data and data streams. The SPOT method was proposed in [498], which is able to determine projected outliers from high dimensional data streams. Thus approach employs a window-based time model and decaying cell summaries to capture statistics from the data stream. A set of top sparse subspaces are obtained by a variety of supervised and unsupervised learning processes. These are used in order detect the projected outliers. A multi-objective genetic algorithm is employed for finding outlying subspaces from training data.

The problem of high dimensional outlier detection has also been extended to other application-specific scenarios such as astronomical data [213], uncertain data [23], transaction data [210] and supervised data [513]. In the uncertain scenario, high dimensional data is especially challenging, because the noise in the uncertain scenario greatly increases the sparsity of the underlying data. Furthermore, the level of uncertainty in the different attributes is available. This helps decide the importance of different attributes for outlier detection purposes. Subspace methods for outlier detection in uncertain data are proposed in [23]. Supervised methods for high-dimensional outlier detection are proposed in [513]. In this case, a small number of examples are presented to user of the outliers. These are then used in order to learn the critical projections which are relevant to the outlierness of an object. The learned information is then leveraged in order to determine the relevant outliers in the underlying data.

Appendix: Distinction between Outlier Detection and Multivariate Extreme Value Analysis

There are a few high-dimensional extreme value analysis methods which are commonly confused with general outlier detection. This appendix (adapted from Chapters 1 and 2 of the book) clears the confusion between these methods. Extreme value analysis can be considered a very specialized (and much simpler) form of outlier analysis in which the entire data is assumed to be generated from a single cluster.

The most basic form of outlier detection is extreme value analysis of 1-dimensional data. These are very specific kinds of outliers, in which it is assumed that the values which are either too large or too small are outliers. Such special kinds of outliers are also important in many application-specific scenarios.

The key is to determine the *statistical tails of the underlying distribution.* The nature of the tails may vary considerably depending upon the underlying data distribution. The normal distribution is the easiest to analyze, because most statistical tests (such as the Z-value test) can be interpreted directly in terms of probabilities of significance. Nevertheless, even for arbitrary distributions, such tests provide a good heuristic idea of the outlier scores of data points, even when they cannot be interpreted statistically. The problem of determining the tails of distributions has been widely studied in the statistics literature. Details of such methods will be discussed in Chapter 2.

Extreme value statistics [364] is distinct from the traditional definition of outliers. The traditional definition of outliers, as provided by Hawkins, defines such objects by their *generative probabilities* rather than the extremity in their values. For example, in the data set $\{1, 2, 2, 50, 98, 98, 99\}$ of 1-dimensional values, the values 1 and 99, could very mildly, be considered extreme values. On the other hand, the value 50 is the average of the data set, and is most definitely not an extreme value. However, most probabilistic and density-based models would classify the value 50 as the strongest outlier in the data, on the basis of Hawkins' definition of generative probabilities. Confusions between extreme value analysis and outlier analysis are common, especially in the context of multivariate data. This is quite often the case, since many extreme value models also use probabilistic models in order to quantify the probability that a data point is an extreme value.

While extreme value analysis is naturally designed for univariate (one-dimensional) data, it is also possible to generalize it to multivariate data, by determining the points at the multidimensional *outskirts* of the data. It is important to understand that such outlier detection methods are tailored to determining *specific kinds of* outliers even in the multivariate case. For example, the point A in Figure A.1 will not be declared as an extreme value by such methods, since it does not lie on the outer boundary of the data, even though it is quite clearly an outlier in Figure A.1. On the other hand, the point B in Figure A.1 can be considered an extreme value, because it lies on the outskirts of the multidimensional data.

Extreme value modeling plays an important role in most outlier detection algorithms as a final step. *This is because most outlier modeling algorithms quantify the deviations of the data points from the normal patterns in the form of a numerical score.* Extreme value analysis is usually required as a final step on these modeled deviations, since they are now represented as univariate values in which extreme values correspond to outliers. In many multi-criteria outlier detection algorithms, a vector of outlier scores may be obtained (such as extreme values of temperature and pressure

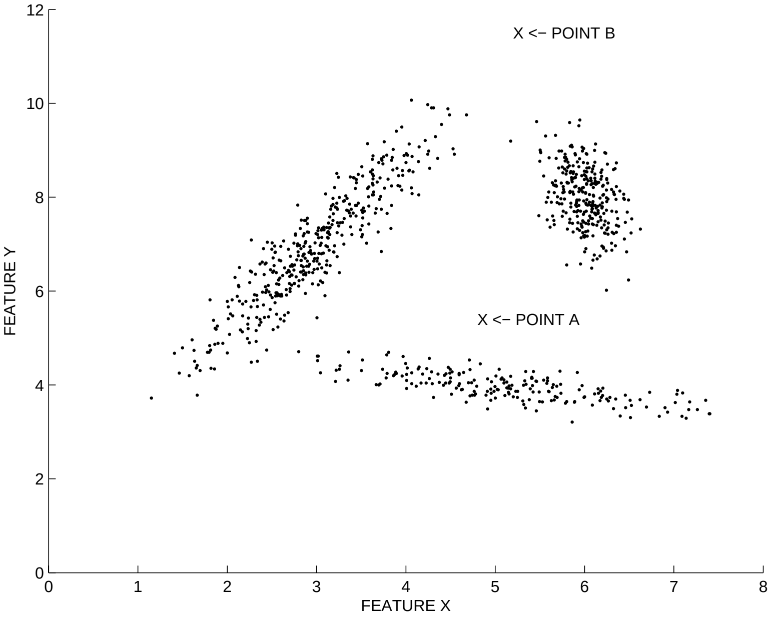


Figure A.1. Difference between outliers and extreme values

in a meteorological application). In such cases, multivariate extreme value methods can help *unify* these multiple outlier scores into a single value, and also generate a binary label output. Therefore, even though the original data may not be in a form where extreme value analysis is directly helpful, it remains an integral part of the outlier detection process. Furthermore, many variables are often tracked as statistical aggregates, in which extreme value analysis provides useful insights about outliers.

Extreme value analysis can also be extended to multivariate data with the use of distance-, or depth-based methods [243, 288, 388]. However, these methods are applicable only to certain kinds of specialized scenarios, where outliers are known to be present at the boundaries of the data. Many forms of post-processing on multi-criterion outlier scores may use such methods. On the other hand, such methods have often not found much utility in the literature for *generic* outlier analysis, because of their inability to discover outlier in the sparse *interior* regions of a data set. It is important to note that multivariate extreme value analysis is a *much simpler* problem than general outlier detection. This section will present some of the multivariate extreme value analysis methods, which are commonly confused with outlier analysis.

0.1 Angle-based Outlier Detection

This method was originally proposed as a general outlier analysis method, though this book has reclassified it to an extreme multivariate analysis method. The idea in angle-based methods is that data points at the boundaries of the data are likely to enclose the entire data within a smaller angle, whereas points in the interior are likely to have data points around them at different angles. For example, consider the

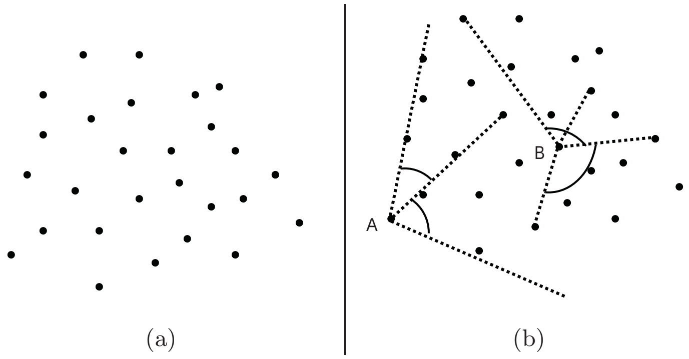


Figure A.2. Angle-based outlier detection

two data points A and B in Figure A.2, in which point A is an outlier, and point B lies in the interior of the data. It is clear that all data points lie within a limited angle centered at A. On the other hand, this is not the case for data point B, which lies within the interior of the data. In this case, the angles between different pairs of points can vary widely. In fact, the more isolated a data point is from the remaining points, the smaller the underlying angle. Thus, data points with a smaller angle spectrum are outliers, whereas those with a larger angle spectrum are not outliers.

Consider three data points $\overline{X}$, $\overline{Y}$, and $\overline{Z}$. Then, the angle between the vectors $\overline{Y}-\overline{X}$ and the $\overline{Z}-\overline{X}$, will not vary much for different values of $\overline{Y}$ and $\overline{Z}$, when $\overline{X}$ is an outlier. Furthermore, the angle is inverse weighted by the distance between the points. The corresponding angle (weighted cosine) is defined as follows:

$$WCos(\overline{Y}-\overline{X},\overline{Z}-\overline{X}) = \frac{<(\overline{Y}-\overline{X}),(\overline{Z}-\overline{X})>}{||\overline{Y}-\overline{X}||_2^2 \cdot ||\overline{Z}-\overline{X}||_2^2}$$

Here $||\cdot||_2$ represents the L_2-norm, and $<\cdot>$ represents the scalar product. Note that this is a weighted cosine, since the denominator contains the squares of the L_2-norms. The inverse weighting by the distance further reduces the weighted angles for outlier points, which also has an impact on the spectrum of angles. Then, the *variance in the spectrum* of this angle is measured by varying the data points $\overline{Y}$ and $\overline{Z}$, while keeping the value of $\overline{X}$ fixed. Correspondingly, the *angle-based outlier factor (ABOF)* of the data point $\overline{X} \in \mathcal{D}$ is defined as follows:

$$ABOF(\overline{X}) = Var_{\{Y,Z \in \mathcal{D}\}} WCos(\overline{Y}-\overline{X},\overline{Z}-\overline{X})$$

Data points which are outliers will have a smaller spectrum of angles, and will therefore have lower values of the angle-based outlier factor $ABOF(\overline{X})$.

The angle-based outlier factor of the different data points may be computed in a number of ways. The naive approach is to pick all possible triples of data points and compute the $O(N^3)$ angles between the different vectors. The $ABOF$ values can be explicitly computed from these values. However, such an approach can be impractical for very large data sets. A number of efficiency-based optimizations have therefore been proposed.

In order to speed up the approach, a natural possibility is to use sampling in order to approximate this value of the angle-based outlier factor. A sample of k data points can be used in order to approximate the ABOF of a data point $\overline{X}$. One possibility is to use an unbiased sample. However, since the angle-based outlier factor is inverse weighted by distances, it follows that the nearest neighbors of a data point have the largest contribution to the angle-based outlier factor. Therefore the k-nearest neighbors of $\overline{X}$ can be used to approximate the outlier factor much more effectively than a unbiased sample of the all the data points. It has also been shown in [269] that many data points can be filtered out on the basis of approximate computation, since their approximate values of the ABOF are too high, and they cannot possibly be outliers. The exact values of the ABOF are computed only for a small set of points, and the points with the lowest values of the ABOF are reported as outliers. We refer the reader to [269] for the details of these efficiency optimizations. An approximation algorithm [363] for the problem has also been proposed in later work.

Because of the inverse weighting by distances, angle-based outlier analysis methods can be considered a hybrid between distance-based and angle-based methods. As discussed earlier with the use of the illustrative example, the latter factor is primarily optimized to finding multivariate extreme values in the data. The precise impact of each of these factors[4] does not seem to be easily quantifiable in a statistically robust way. In most data sets such as in Figure A.1, outliers lie not just on the boundaries of the data, but also in the interior of the data. Unlike extreme values, outliers are defined by generative probabilities. While the distance factor can provide some impact for the outliers in the interior, the work is primarily focussed on the advantage of angular measures, and it is stated in [269] that the degree of impact of distance factors is minor compared to the angular factors. This implies that outliers on the boundaries of the data will be highly favored in terms of the overall score, because of the lower spectrum of angles. Therefore, the angle-based method treats outliers with similar generative probabilities in the interior and the boundaries of the data in a differential way, which is not statistically desirable for general outlier analysis. Specifically, the outliers at the boundaries of the data are more likely to be favored in terms of the outlier score. Such methods can effectively find outliers for the case illustrated in Figure A.3, but the outlier A illustrated in Figure A.1 will be favored less. Therefore, while this approach was originally presented as a general outlier analysis method, it has been classified in the section on multivariate extreme value analysis methods in this book.

It has been claimed in [269] that the approach is more suitable for high dimensional data because of its use of angles, as opposed to distances. However, it has been shown in earlier work [380], that angle-based measures are not immune to the dimensionality curse, because of concentration effects in the cosine measure. Such concentration effects would also impact the spectrum of the angles, even when they are combined with distances. The variation in the angle spectrum in Figure A.2 is easy to show visually in 2-dimensional data, but the sparsity effects will also impact the spectrum of angles in higher dimensions. In high-dimensional space, all the cosines will converge to a constant value of 0.5, as all triangles become increasingly equilateral. Note that this is an even stronger form of convergence than distance values. The cosine between two data points can be directly expressed in terms of the Euclidean distance, which

[4] When a random variable is scaled by a factor of a, its variance is scaled by a factor of a^2. However, the scaling here is not by a constant factor.

is itself known to work poorly in high-dimensions.

$$\text{Cosine}(\overline{X}, \overline{Y}) = \frac{||X - 0||^2 + ||Y - 0||^2 - ||X - Y||^2}{2 \cdot ||X - 0|| \cdot ||Y - 0||} \tag{A.1}$$

Therefore, the use of the spectrum of angles simply pushes the challenges of high dimensions to a different part of the analysis. A clear explanation of why the spectrum of angles should be more robust to high dimensionality than distances has not[5] been provided in [269]. More importantly, such methods do not address the issue of locally irrelevant attributes [4], which are the primary impediment to effective outlier analysis methods with increasing dimensionality. Another important point to note is that multivariate extreme value analysis is much simpler than general outlier analysis in high dimensionality, because the parts of the data to explore are approximately known, and therefore the analysis is global rather than local. The evidence over different dimensions can be accumulated with the use of a very simple classical distance-distribution method [288, 406]. The approach, discussed in the next section, is also suitable for high-dimensional extreme value analysis, because it implicitly weights globally relevant and irrelevant directions in the data in a different way, and is statistically sound, in terms of probabilistic interpretability of the extreme values.

0.2 Distance Distribution-based Methods

A *distribution-dependent* approach is to model the entire data set to be normally distributed about its mean in the form of a multivariate Gaussian distribution. Let $\overline{\mu}$ be the d-dimensional mean vector of a d-dimensional data set, and Σ be its $d \times d$ co-variance matrix. In this case, the (i, j)th entry of the covariance matrix is equal to the covariance between the dimensions i and j. Then, the probability distribution $f(\overline{X})$ for a d-dimensional data point $\overline{X}$ can be defined as follows:

$$f(\overline{X}) = \frac{1}{\sqrt{|\Sigma|} \cdot (2 \cdot \pi)^{(d/2)}} \cdot \exp(-\frac{1}{2} \cdot (\overline{X} - \overline{\mu}) \cdot \Sigma^{-1} \cdot (\overline{X} - \overline{\mu})^T)$$

The value of $|\Sigma|$ denotes the determinant of the covariance matrix. We note that the term in the exponent is (half) the *Mahalanobis distance* between the data point $\overline{X}$ and the mean $\overline{\mu}$ of the data. The computation of the Mahalanobis distance requires the inversion of the covariance matrix Σ. The value in the exponent of the normal distribution above is used as the outlier score.

The Mahalanobis distance is similar to the euclidian distance, except that it normalizes the data on the basis of the inter-attribute correlations. For example, if the axis system of the data were to be rotated to the principal directions (shown in Figure A.3), then the data would have no inter-attribute correlations. It is actually possible to determine such directions of correlations generally in d-dimensional data sets. The Mahalanobis distance is simply equal to the Euclidean distance in such a transformed (axes-rotated) data set *after* dividing each of the transformed coordinate values by

[5] The use of the cosine function in some high-dimensional domains such as text has been cited as an example in a later work [270]. In domains with small and varying non-zero attributes, the cosine is preferred because of important normalization properties, and not because of greater dimensionality resistance. The cosine function is not immune to the dimensionality curse even for the unique structure of text [380]. An increasing fraction of non-zero attributes, towards more general distributions, directly impacts the data hubness.

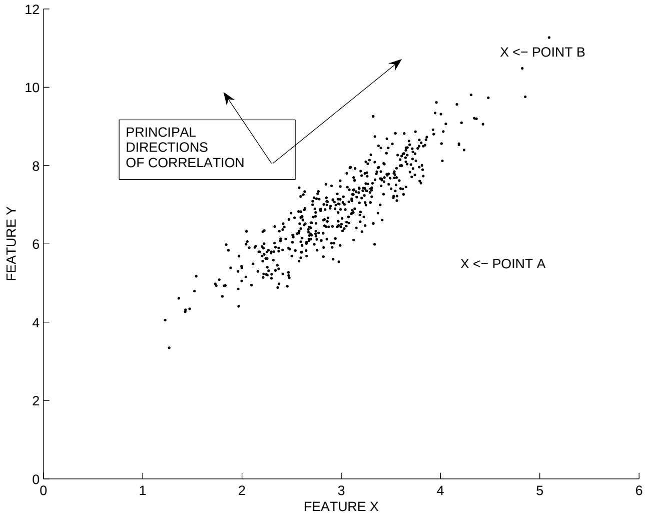


Figure A.3. Extreme value analysis in multivariate data with Mahalanobis distance

the standard-deviation of that direction. While principal component analysis can also be used in order to compute the value in the exponent of the normal distribution above, a simpler way to do it is by evaluating the term in the exponent of the modeled normal distribution. More will be discussed about this issue in Chapter 3.

This approach recognizes the fact that the different directions of correlation have different variance, and the data should be treated in a statistically normalized way along these directions. For example, in the case of Figure A.3, the data point A can be more reasonably considered an outlier than data point B, on the basis of the natural correlations in the data. On the other hand, the data point A is closer to the centroid of the data (than data point B) on the basis of *euclidian distance*, but not on the basis of the Mahalanobis distance. Interestingly, data point A also seems to have a much higher spectrum of angles than data point B, at least from an average sampling perspective. This implies that, at least on the basis of the primary criterion of angles, the angle-based method would likely favor data point B. This is because it is unable to account for the relative relevance of the different directions, an issue which becomes more prominent with increasing dimensionality. The Mahalanobis method is robust to increasing dimensionality, because it uses the covariance matrix in order to summarize the high dimensional deviations in a statistically effective way.

We further note that each of the distances along the principal correlation directions can be modeled as a one-dimensional standard normal distribution, which is approximately independent from the other orthogonal directions of correlation. As discussed earlier in this chapter, the sum of the squares of d variables drawn independently from a standard normal distributions, will result in a variable drawn from a χ^2 distribution with d degrees of freedom. Therefore, the cumulative probability distribution tables of the χ^2 distribution can be used in order to determine the outliers with the appropriate level of significance.

This simple approach is effective for the example of Figure A.3, because the entire data set is distributed in one large cluster about the mean. For cases in which the data may have many different clusters with different orientations, such an extreme value approach may not be effective. An example of such a data set is illustrated in Figure A.1. For such cases, more general distribution-based modeling algorithms are needed.

References

[1] N. Abe, B. Zadrozny, and J. Langford. Outlier Detection by Active Learning, *ACM KDD Conference*, 2006.

[2] E. Achtert, A. Hettab, H.-P. Kriegel, E. Schubert, and A. Zimek. Spatial Outlier Detection: Data, Algorithms, Visualizations. *SSTD Conference*, 2011.

[3] N. R. Adam, V. P. Janeja, and V. Atluri. Neighborhood based detection of anomalies in high dimensional spatio-temporal sensor datasets. *ACM SAC Conference*, 2004.

[4] C. C. Aggarwal, and P. S. Yu. Outlier Detection in High Dimensional Data, *ACM SIGMOD Conference*, 2001.

[5] C. C. Aggarwal, C. Procopiuc, J. Wolf, P. Yu, and J. Park. Fast Algorithms for Projected Clustering, *ACM SIGMOD Conference*, 1999.

[6] C. Aggarwal, J. Han, J. Wang, and P. Yu. A Framework for Projected Clustering of High Dimensional Data Streams. In *VLDB Conference*, 2004.

[7] C. C. Aggarwal, and P. S. Yu. Finding Generalized Projected Clusters in High Dimensional Spaces, *ACM SIGMOD Conference*, 2000.

[8] C. C. Aggarwal. Re-designing Distance Functions and Distance-based Applications for High Dimensional Data, *ACM SIGMOD Record*, 2001.

[9] C. C. Aggarwal. On Abnormality Detection in Spuriously Populated Data Streams, *SIAM Conference on Data Mining*, 2005.

[10] C. C. Aggarwal. Data Streams: Models and Algorithms, *Springer*, 2007.

[11] C. C. Aggarwal. Social Network Data Analytics, *Springer*, 2011.

[12] C. C. Aggarwal. On Effective Classification of Strings with Wavelets, *ACM KDD Conference*, 2002.

[13] C. C. Aggarwal, N. Ta, J. Wang, J. Feng, and M. J. Zaki. Xproj: A Framework for Projected Structural Clustering of XML Documents. *ACM KDD Conference*, 2007.

[14] C. C. Aggarwal, and P. S. Yu. On String Classification in Data Streams, *ACM KDD Conference*, 2007.

[15] C. C. Aggarwal, Y. Zhao, and P. S. Yu. Outlier Detection in Graph Streams, *ICDE Conference*, 2011.

[16] C. C. Aggarwal. A Framework for Diagnosing Changes in Evolving Data Streams, *ACM SIGMOD Conference*, 2003.

[17] C. C. Aggarwal, and P. S. Yu. Online Analysis of Community Evolution in Data Streams, *SDM Conference*, 2005.

[18] C. C. Aggarwal. On the Effects of Dimensionality Reduction on High Dimensional Similarity Search, *ACM PODS Conference*, 2001.

[19] C. C. Aggarwal. Managing and Mining Sensor Data, *Springer*, 2013.

[20] C. C. Aggarwal, and C. K. Reddy. Data Clustering: Algorithms and Applications, *CRC Press*, 2013.

[21] C. Aggarwal, and C. Zhai. Managing and Mining Text Data, *Springer*, 2012.

[22] C. C. Aggarwal, A. Hinneburg, and D. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space, *ICDT Conference*, 2001.

[23] C. C. Aggarwal, and P. S. Yu. Outlier Detection with Uncertain Data, *SDM Conference*, 2008.

[24] C. C. Aggarwal, Y. Xie, and P. S. Yu. On Dynamic Data-Driven Selection of Sensor Streams, *ACM KDD Conference*, 2011.

[25] C. C. Aggarwal, J. Han. J. Wang, and P. Yu. A Framework for Clustering Evolving Data Streams, *VLDB Conference*, 2003.

[26] C. C. Aggarwal, and P. Yu. On Clustering Massive Text and Categorical Data Streams, *Knowledge and Information Systems*, 24(2), pp. 171–196, 2010.

[27] C. C. Aggarwal, and K. Subbian. Event Detection in Social Streams, *SDM Conference*, 2012.

[28] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules Between Sets of Items in Large Databases. *ACM SIGMOD Conference*, 1993.

[29] R. Agrawal, and R. Srikant. Fast algorithms for finding Association Rules in Large Databases, *VLDB Conference*, 1994.

[30] M. Agyemang, K. Barker, and R. Alhajj. A Comprehensive Survey of Numeric and Symbolic Outlier Mining Techniques, *Intelligent Data Analysis*, 10(6). pp. 521–538, 2006.

[31] A. Ahmad and L. Dey. A Method to Compute Distance between two Categorical Values of Same Attribute in Unsupervised Learning for Categorical Data Set. *Pattern Recognition Letters*, 28(1), pp. 110–118, 2007.

[32] R. Ahuja, J. Orlin, and T. Magnanti. Network Flows: Theory, Algorithms and Applications, *Prentice Hall*, 1993.

[33] L. Akoglu, M. McGlohon, and C. Faloutsos. Oddball: Spotting Anomalies in Weighted Graphs, *PAKDD Conference*, 2010.

[34] L. Akoglu, H. Tong, J. Vreeken, and C. Faloutsos. Fast and Reliable Anomaly Detection in Categorical Data, *CIKM Conference*, 2012.

[35] E. Aleskerov, B. Freisleben, and B. Rao. CARDWATCH: A Neural Network based Database Mining System for Credit Card Fraud Detection. *IEEE Computational Intelligence for Financial Engineering*, pp. 220–226, 1997.

[36] T. Al-Khateeb, M. Masud, L. Khan, C. Aggarwal, J. Han, and B. Thuraisingham. Recurring and Novel Class Detection using Class-based Ensemble, *ICDM Conference*, 2012.

[37] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. *ACM SIGIR Conference*, 1998.

[38] J. Allan, V. Lavrenko, and H. Jin. First story detection in TDT is hard. *ACM CIKM Conference*, 2000.

[39] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic Detecting and Tracking Pilot Study Final Report, *CMU Technical Report, Paper 341*, 1998.

[40] F. Alonso, J. Caraca-Valente, A. Gonzalez, and C. Montes. Combining Expert Knowledge and Data Mining in a Medical Diagnosis Domain. *Expert Systems with Applications*, 23(4), pp. 367–375, 2002.

[41] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden Narkov Support Vector Machines. *ICML Conference*, 2003.

[42] M. R. Anderberg. Cluster Analysis for Applications, *Academic Press*, New York, 1973.

[43] D. Anderson, T. Lunt, H. Javitz, A. Tamaru, and A. Valdes. Detecting Unusual Program Nehavior using the Statistical Components of NIDES, *Techical Report, SRI–CSL–95–06, Computer Science Laboratory*, SRI International, 1995.

[44] D. Anderson, T. Frivold, A. Tamaru, and A. Valdes. Next-generation Intrusion Detection Expert System (nides), Software Users Manual, Beta-update Release. *Technical Report SRI–CSL–95–07, Computer Science Laboratory*, SRI International, 1994.

[45] S. Ando. Clustering Needles in a Haystack: An Information Theoretic Analysis of Minority and Outlier Detection. *ICDM Conference*, 2007.

[46] F. Angiulli, and C. Pizzuti. Fast Outlier Detection in High Dimensional Spaces. *European Conferece on Principles of Knowledge Discovery and Data Mining*, 2002.

[47] F. Angiulli, F. Fassetti, amd L. Palopoli. Finding Outlying Properties of Exceptional Objects, *ACM Transactions on Database Systems*, 34(1), 2009.

[48] F. Angiulli and F. Fassetti. Detecting Distance-based Outliers in Streams of Data, *ACM CIKM Conference*, 2007.

[49] A. Arning, R. Agrawal, and P. Raghavan. A Linear Method for Deviation Detection in Large Databases. *ACM KDD Conference*, 1996.

[50] I. Assent, P. Kranen, C. Beldauf, and T. Seidl. AnyOut: Anytime Outlier Detection in Streaming Data, *DASFAA Conference*, 2012.

[51] I. Assent, R. Krieger, E. Muller, and T. Seidl. Subspace Outlier Mining in Large Multimedia Databases, *Parallel Universes and Local Patterns*, 2007.

[52] A. Auer Jr. Correlation of Land Use and Cover with Meteorological Anomalies. *Journal of Applied Meteorology*, 17(5) pp. 636–643, 1978.

[53] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. *KDD Conference*, 2006.

[54] D. Ballard. Generalizing the Hough Transform to Detect Arbitrary Shapes, *Pattern Recognition*, 11(2), pp. 111–122, 1981.

[55] D. Barbara, Y. Li, J. Couto, J.-L. Lin, and S. Jajodia. Bootstrapping a Data Mining Intrusion Detection System. *Symposium on Applied Computing*, 2003.

[56] D. Barbara, J. Couto, S. Jajodia, and N. Wu. ADAM: A Testbed for Exploring the Use of Data Mining in Intrusion Detection. *ACM SIGMOD Record*, 30(4), pp. 15–24, 2001.

[57] D. Barbara, J. Couto, S. Jajodia, and N. Wu. Detecting Movel Network Intrusions using Bayes Estimators. *SIAM Conference on Data Mining*, 2001.

[58] V. Barnett and T. Lewis. Outliers in Statistical Data, *Wiley*, 1994.

[59] R. Baragona and F. Battaglia. Outlier Detection in Multivariate Time Series by Independent Component Analysis. *Neural Computation*, 19(1), pp. 1962–1984, 2007.

[60] S. Bay, and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. *ACM KDD Conference*, 2003.

[61] S. Bay, K. Saito, N. Ueda, and P. Langley. A Framework for Discovering Anomalous Regimes in Multivariate Time-series Data with Local Models. *Technical report, Center for the Study of Language and Information*, Stanford University, 2004.

[62] R. Beckman, and R. Cook. Outliers, *Technometrics*, 25(2), pp. 119–149, 1983.

[63] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The R*-tree: An Efficient and Robust Access method for Points and Rectangles. *ACM SIGMOD Conference*, 1990.

[64] M. Berlingerio, F. Bonchi, B. Bringmann, and A. Gionis. Mining Graph Evolution Rules. *ECML/PKDD Conference*, 2009.

[65] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? *International Conference on Database Theory*, 1999.

[66] K. Bhaduri, B. Matthews, and C. Giannella. Algorithms for Speeding up Distance-based Outlier Detection. *ACM KDD Conference*, 2011.

[67] E. Blanzieri and A. Bryl. A Survey of Learning-based Techniques of Email Spam Filtering. *Artificial Intelligence Review*, 29(1), pp. 63–92, 2008.

[68] M. Bosc, F. Heitz, J.-P. Armspach, I. Namer, D. Gounot, and L. Rumbach. Automatic Change Detection in Multimodal Serial MRI: application to multiple sclerosis lesion evolution, *NeuroImage*, 20(2), 2003, Pages 643–656

[69] Y. Bilberman. A Context Similarity Measure. *ECML Conference*, 1994.

[70] P. Billingsley. Probability and Measure, Second Edition, *Wiley*, 1986.

[71] D. Birant, and A. Kut. Detecting Spatio-temporal Outliers in Large Databases, *Journal of Computing and Information Technology*, 14(4), pp. 291–297, 2006.

[72] D. Blei, and J. Lafferty. Dynamic topic models. *ICML Conference*, 2006.

[73] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation, *Journal of Machine Learning Research*, 3: pp. 993–1022, 2003.

[74] C. Bohm, K. Haegler, N. Muller, and C. Plant. Coco: Coding Cost for Parameter Free Outlier Detection, *ACM KDD Conference*, 2009.

[75] S. Boriah, V. Chandola, and V. Kumar. Similarity Measures for Categorical Data: A Comparative Evaluation, *SIAM Conference on Data Mining*, 2008.

[76] J. Branch, B. Szymanski, C. Giannella, R. Wolff, and H. Kargupta. In-Network Outlier Detection in Wireless Sensor Networks. *ICDCS Conference*, 2006.

[77] T. Brants, F. Chen, and A. Farahat. A System for New Event Detection. *ACM SIGIR Conference*, 2003.

[78] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying Density-based Local Outliers, *ACM SIGMOD Conference*, 2000.

[79] M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. OPTICS-OF: identifying local outliers. *PKDD Conference*, 1999.

[80] L. Brieman. Bagging Predictors, *Machine Learning*, 24: pp. 123–140, 1996.

[81] M. R. Brito, E. L. Chavez, A. J. Quiroz, and J. E. Yukich. Connectivity of the Mutual k-Nearest Neighbor Graph in Clustering and Outlier Detection. *Statistics and Probability Letters*, 35(1), pp. 33–42, 1997.

[82] R. G. Brown, and P. Hwang. Introduction to Random Signals and Applied Kalman Filtering, *John Wiley and Sons*, 1997.

[83] Y. Bu, L. Chen, A. W.-C. Fu, and D. Liu. Efficient Anomaly Monitoring over Moving Object Trajectory Streams. *ACM KDD Conference*, 2009.

[84] S. Budalakoti, A. Srivastava, R. Akella, and E. Turkov. Anomaly Detection in Large Sets of High-dimensional Symbol Sequences, *NASA Ames Research Center*, Technical Report NASA TM-2006-214553, 2006.

[85] S. Budalakoti, A. Srivastava, and M. Otey. Anomaly Detection and Diagnosis Algorithms for Discrete Symbol Sequences with Applications to Airline Safety, *IEEE International Conference on Systems, Man, and Cybernetics*, 37(6), 2007.

[86] S Burdakis, A Deligiannakis. Detecting Outliers in Sensor Networks using the Geometric Approach, *ICDE Conference*, 2012.

[87] T. Burnaby. On a Method for Character Weighting a Similarity Coefficient, Employing the Concept of Information. *Mathematical Geology*, 2(1), 25–38, 1970.

[88] S. Byers, and A. Raftery. Nearest-Neighbor Clutter Removal for Estimating Features in Spatial Point Processes, *JASIS*, 93, pp. 577–584, June 1998.

[89] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Visualization of Navigation Patterns on a Web Site using Model-based Clustering, *ACM SIGMOD Conference*, 2000.

[90] P. H. Calamai. Projected gradient methods for linearly constrained problems. *Mathematical Programming*, 39: pp, 93–116, 1987.

[91] C. Campbell, and K. P. Bennett. A Linear-Programming Approach to Novel Class Detection, *NIPS Conference*, 2000.

[92] M. J. Canty. Image Analysis, Classification and Change Detection in Remote Sensing: with Algorithms for ENVI/IDL, *CRC Press*, 2006.

[93] C. Caroni. Outlier Detection by Robust Principal Component Analysis. *Communications in Statistics – Simulation and Computation*, 29: pp. 129–151, 2000.

[94] L. E. Carr and R. L. Elsberry. Monsoonal interactions leading to sudden tropical cyclone track changes, *Monthly Weather Review*, 123(2), pp. 265–290, Feb. 1995.

[95] K. Chakrabarti, S. Mehrotra. Local Dimensionality Reduction: A New Approach to Indexing High Dimensional Spaces. *VLDB Conference Proceedings*, 2000.

[96] S. Chakrabarti, S. Sarawagi, and B. Dom. Mining Surprising Patterns using Temporal Description Length. *VLDB Conference*, 1998.

[97] V. Chandola, V. Mithal, and V. Kumar. A Comparative Evaluation of Anomaly Detection Techniques for Sequence Data, *International Conference on Data Mining*, 2008.

[98] A. Chaudhary, A. S. Szalay, and A. W. Moore. Very Fast Outlier Detection in Large Multidimensional Data Sets. *DMKD Workshop*, 2002.

[99] D. Chakrabarti, and C. Faloutsos. Evolutionary Clustering, *ACM KDD Conference*, 2006.

[100] D. Chakrabarti. AutoPart: Parameter-Free Graph Partitioning and Outlier Detection. *PKDD Conference*, 2004.

[101] C.-H. Chan and G. Pang. Fabric Defect Detection by Fourier Analysis, *IEEE Transactions on Industry Applications*, 36(5), 2000.

[102] N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: Special Issue on Learning from Imbalanced Data Sets, *ACM SIGKDD Explorations Newsletter*, 6(1), 1–6, 2004.

[103] N. V. Chawla, K. W. Bower, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research (JAIR)*, 16, pp. 321–356, 2002.

[104] N. Chawla, A. Lazarevic, L. Hall, and K. Bowyer. SMOTEBoost: Improving prediction of the minority class in boosting, *PKDD*, pp. 107–119, 2003.

[105] N. V. Chawla, D. A. Cieslak, L. O. Hall, and A. Joshi. Automatically Cuntering Imbalance and its Empirical Relationship to Cost. *Data Mining and Knowledge Discovery*, 17(2), pp. 225–252, 2008.

[106] P. K. Chan and S. J. Stolfo. Toward Scalable Learning with Non-uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. *KDD Conference*, pp. 164–168, 1998.

[107] V. Chandola, A. Banerjee, and V. Kumar. Anomaly Detection: A Survey. *ACM Computing Surveys*, 41(3), 2009.

[108] V. Chandola, A. Banerjee, and V. Kumar. Anomaly Detection for Discrete Sequences: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 24(5): pp. 823–839, 2012.

[109] I. Chang, G. C. Tiao, and C. Chen. Estimation of Time Series Parameters in the Presence of Outliers. *Technometrics*, 30(2), pp. 193–204, 1988.

[110] C. Chen and L.-M. Liu. Joint Estimation of Model Parameters and Outlier Effects in Time Series. *Journal of the American Statistical Association*, 88(421), pp. 284–297, March 1993.

[111] Y. Chen, and L. Tu. Density-based Clustering for Real Time Stream Sata, *ACM KDD Conference*, 2007.

[112] D. Chen, C.-T. Lu, Y. Chen, and D. Kou. On Detecting Spatial Outliers, *Geoinformatica*, 12: pp. 455–475, 2008.

[113] T. Cheng and Z. Li. A Hybrid Approach to Detect Spatialtemporal Outliers. *International Conference on Geoinformatics*, 2004.

[114] T. Cheng and Z. Li. A Multiscale Approach for Spatio-temporal Outlier Detection, *Transactions in GIS*, 10(2), pp. 253–263, March 2006.

[115] H. Cheng, P.-N. Tan, C. Potter, and S. Klooster. Detection and Characterization of Anomalies in Multivariate Time Series, *SIAM Conference on Data Mining*, 2009.

[116] Y. Chi, X. Song, D. Zhou, K. Hino, and B. Tseng. Evolutionary Spectral Clustering by Incorporating Temporal Smoothness. *ACM KDD Conference*, 2007.

[117] A. Chiu, and A. Fu. Enhancements on Local Outlier Detection. *Database Engineering and Applications Symposium*, 2003.

[118] M. Chow, R. Sharpe, and J. Hung. On the Application and Design of Artificial Neural Networks for Motor Fault Detection. *IEEE Transactions on Industrial Electronics*, 40(2), 1993.

[119] C. Chow, sand D. Yeung. Parzen-Window Network Intrusion Detectors. *International Conference on Pattern Recognition*, 4, 2002.

[120] W. Cohen. Fast Effective Rule Induction. *ICML Conference*, 1995.

[121] D. Cohn, R. Atlas, and N. Ladner. Improving Generalization with Active Learning, *Machine Learning*, 15, pp. 201–221.

[122] R. Cooley, B. Mobashar, and J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns, *Knowledge and Information Systems*, 1, pp, 5–32, 1999.

[123] D. Cook, and L. Holder. Graph-Based Data Mining. *IEEE Intelligent Systems*, 15(2), pp. 32–41, 2000.

[124] G. Cormack. Email Spam Filtering: A Systematic Review, *Foundations and Trends in Information Retrieval*, 1(4), pp. 335–455, 2007.

[125] C. Darwin. The Origin of the Species by Natural Selection, 1859. Manuscript now publicly hosted at: `http://www.literature.org/authors/darwin-charles/the-origin-of-species/`

[126] G. Das and H. Mannila. Context-based Similarity Measures for Categorical Databases. *PKDD Conference*, 2000.

[127] K. Das, J. Schneider, and D. Neill. Anomaly Pattern Detection in Categorical Data Sets, *ACM KDD Conference*, 2008.

[128] S. Das, B. Matthews, A. Srivastava, and N. Oza. Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study. *ACM KDD Conference*, 2010.

[129] D. Dasgupta and S. Forrest. Novelty Detection in Time Series using Ideas from Immunology, *International Conference on Intelligent Systems*, 1996.

[130] D. Dasgupta, and F. Nino. A comparison of negative and positive selection algorithms in novel pattern detection. *IEEE Conference on Systems, Man, and Cybernetics*, 1, pp. 125–130, 2000.

[131] T. Dasu, S. Krishnan, S. Venkatasubramanian, and K. Yi. An Information-Theoretic Approach to Detecting Change in Multidimensional Data Streams, *Symposium on the Interface of Computer Science, Statistics, and Applications*, 2006.

[132] N. Delannay, C. Archambeau, and M. Verleysen. Improving the Robustness to Outliers of Mixtures of Probabilistic PCAs. *PAKDD Conference*, 2008.

[133] S. T. Deerwester, S. T. Dumais, G. Furnas, and R. Harshman. Indexing by Latent Semantic Analysis, *JASIS*, 1990.

[134] K. A. De Jong. Analysis of the behaviour of a class of Genetic Adaptive Systems. *Ph.D. Dissertation, University of Michigan*, Ann Arbor, MI, 1975.

[135] A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, B, vol. 39(1), pp. 1–38, 1977.

[136] D. Denning. An Intrusion Detection Model. *IEEE Transactions of Software Engineering*, 13(2), pp. 222–232.

[137] R. Derrig. Insurance Fraud. *Journal of Risk and Insurance*, 69(3), pp. 271–287, 2002.

[138] M. Desforges, P. Jacob, and J. Cooper. Applications of Probability Density Estimation to the Detection of Abnormal Conditions in Engineering. *Proceedings of Institute of Mechanical Engineers*, Vol. 212, pp. 687–703, 1998.

[139] M. Deshpande and G. Karypis. Evaluation of Techniques for Classifying Biological Sequences. *PAKDD Conference*, 2002.

[140] M. Deshpande and G. Karypis. Selective Markov Models for Predicting Web Page Accesses, *ACM Transactions on Internet Technology*, 4(2), pp. 163–184, 2004.

[141] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Neasures. *PVLDB*, 1(2), pp. 1542–1552, 2008.

[142] S. Donoho. Early Detection of Insider Trading in Option Markets. *ACM KDD Conference*, 2004.

[143] C. Drummond and R. Holte. C4.5, Class Imbalance, and Cost Sensitivity: Why Undersampling beats Oversampling. *ICML Workshop on Learning from Imbalanced Data Sets*, 2003.

[144] C. Drummond and R. Holte. Explicitly Representing Expected Cost: An Alternative to ROC representation. *ACM KDD Conference*, pp. 198–207, 2001.

[145] P. Domingos. MetaCost: A General Framework for Making Classifiers Cost-Sensitive, *ACM KDD Conference*, 1999.

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

[147] H. Dutta, C. Giannella, K. Borne, and H. Kargupta. Distributed top-k Outlier Detection in Astronomy Catalogs using the Demac System. *SDM Conference*, 2007.

[148] W. Eberle and L. B. Holder. Mining for Structural Anomalies in Graph-based Data. *DMIN*, 2007.

[149] F. Y. Edgeworth. On Discordant Observations. *Philosophical Magazine*, 23(5), pp. 364–375, 1887.

[150] M. Elahi, K. Li, W. Nisar, X. Lv, and H. Wang. Efficient Clustering-Based Outlier Detection Algorithm for Dynamic Data Stream. *FSKD Conference*, 2008.

[151] C. Elkan. The Foundations of Cost-Sensitive Learning, *IJCAI*, 2001.

[152] C. Elkan, and K. Noto. Learning Classifiers from only Positive and Unlabeled Data, *ACM KDD Conference*, 2008.

[153] D. Endler. Intrusion detection: Applying Machine Learning to Solaris Audit Data, *Annual Computer Security Applications Conference*, 1998.

[154] E. Eskin. Anomaly Detection over Noisy Data using Learned Probability Distributions, *ICML Conference*, 2000.

[155] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A Geometric Framework for Unsupervised Anomaly Detection, In *Applications of Data Mining in Computer Security.* Kluwer, 2002.

[156] E. Eskin, W. Lee, and S. Stolfo, Modeling System Call for Intrusion Detection using Dynamic Window Sizes, *DISCEX*, 2001.

[157] H. Fan, O. Zaiane, A. Foss. and J. Wu. A Nonparametric Outlier Detection for Efficiently Discovering top-n Outliers from Engineering Data. *PAKDD Conference*, 2006.

[158] W. Fan, S. Stolfo, J. Zhang, and P. Chan. AdaCost: Misclassification Cost Sensitive Boosting, *ICML Conference*, 1999.

[159] T. Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers, *Technical Report HPL-2003-4*, Palo Alto, CA: HP Laboratories, 2003.

[160] T. Fawcett and F. Provost. Activity Monitoring: Noticing Interesting Changes in Behavior. *ACM KDD Conference*, 1999.

[161] D. Fetterly, M. Manasse, and M. Najork. Spam, Damn Spam, and Statistics: using Statistical Analysis to Locate Spam Web Pages, *WebDB*, 2004.

[162] A. Lung-Yut-Fong, C. Levy-Leduc, and O. Cappe. Distributed Detection/localization of Change-points in High-dimensional Network Traffic Data. *Corr*, abs/0909.5524, 2009.

[163] S. Forrest, C. Warrender, and B. Pearlmutter. Detecting Intrusions using System Calls: Alternate Data Models, *IEEE ISRSP*, 1999.

[164] S. Forrest, S. Hofmeyr, A. Somayaji, and T. A. Longstaff. A Sense of Self for Unix Processes, *ISRSP*, 1996.

[165] S. Forrest, P. D'Haeseleer, and P. Helman. An Immunological Approach to Change Detection: Algorithms, Analysis and Implications. *IEEE Symposium on Security and Privacy*, 1996.

[166] S. Forrest, F. Esponda, and P. Helman. A Formal Framework for Positive and Negative Detection Schemes. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, pp. 357–373, 2004.

[167] S. Forrest, A. Perelson, L. Allen, and R. Cherukuri. Self-Nonself Discrimination in a Computer. *IEEE Symposium on Security and Privacy*, 1994.

[168] A. Fox. Outliers in Time Series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(3), pp. 350–363, 1972.

[169] A. Frank, and A. Asuncion. UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2010. `http://archive.ics.uci.edu/ml`

[170] C. Franke and M. Gertz. ORDEN: Outlier Region Detection and Exploration in Sensor Networks. *ACM SIGMOD Conference*, 2009.

[171] A. Fu, O. Leung, E. Keogh, and J. Lin. Finding Time Series Discords based on Haar Transform. *Advanced Data Mining and Applications*, 2006.

[172] R. Fujumaki, T. Yairi, and K. Machida. An Approach to Spacecraft Anomaly Detection Problem using Kernel Feature Space. *ACM KDD Conference*, 2005.

[173] R. Fujamaki. Anomaly Detection Support Vector Machine and Its Application to Fault Diagnosis, *ICDM Conference*, 2008.

[174] P. Galeano, D. Pea, and R. S. Tsay. Outlier detection in Multivariate Time Series via Projection Pursuit.*Statistics and Econometrics Working Papers WS044221*, Universidad Carlos III, 2004.

[175] P. Gambaryan. A Mathematical Model of Taxonomy. *Izvest. Akad. Nauk Armen*, SSR, 17(12), pp. 47–53, 1964.

[176] V. Ganti, J. Gehrke, and R. Ramakrishnan. CACTUS: Clustering Categorical Data using Summaries. *ACM KDD Conference*, 1999.

[177] B. Gao, H.-Y. Ma, and Y.-H. Yang, HMMs (Hidden Markov Models) based on Anomaly Intrusion Detection Method, *International Conference on Machine Learning and Cybernetics*, 2002.

[178] H. Gao, X. Wang, J. Tang and H. Liu. Network Denoising in Social Media, *Technical Report, Arizona State University*, 2011.

[179] J. Gao and P.-N. Tan. Converting Outlier Scores from Outlier Detection Algorithms into Probability Estimates, *ICDM Conference*, 2006.

[180] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han. On Community Outliers and their Efficient Detection in Information Networks. *ACM KDD Conference*, pp. 813–822, 2010.

[181] Y. Ge, H. Xiong, Z.-H. Zhou, H. Ozdemir, J. Yu, and K. Lee. Top-Eye: Top-k Evolving Trajectory Outlier Detection. *CIKM Conference*, 2010.

[182] A. Ghosh, J. Wanken, and F. Charron. Detecting Anomalous and Unknown Intrusions against Programs, *Annual Computer Security Applications Conference*, 1998.

[183] A. Ghosh, A. Schwartzbard, and M. Schatz. Learning Program Behavior Profiles for Intrusion Detection, *USENIX Workshop on Intrusion Detection and Network Monitoring*, pp. 51–62, 1999.

[184] S. Ghosh and D. Reilly. Credit Card Fraud Detection with a Neural Network, *International Conference on System Sciences: Information Systems: Decision Support and Knowledge-Based Systems*, 3, pp. 621–630, 1994.

[185] A. Ghoting, S. Parthasarathy, and M. Otey. Fast Mining of Distance-based Outliers in High Dimensional Spaces. *SIAM Conference on Data Mining*, 2006.

[186] D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: an approach based on dynamical systems. *The VLDB Journal*, 8(3), pp. 222–236, 2000.

[187] D. W. Goodall. A new similarity index based on probability. *Biometrics*, 22(4), pp. 882–907, 1966.

[188] F. Grubbs. Procedures for Detecting Outlying Observations in Samples, *Technometrics*, 11(1), pp. 1–21, 1969.

[189] S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5), pp. 345–366, 2000.

[190] D. Gunopulos and G. Das. Time-series Similarity Measures, and Time Series Indexing, *ACM SIGMOD Conference*, 2001.

[191] S. Gunter, N. N. Schraudolph, and S. V. N. Vishwanathan. Fast Iterative Kernel Principal Component Analysis. *Journal of Machine Learning Research*, 8, pp 1893–1918, 2007.

[192] M. Gupta, C. Aggarwal, J. Han, and Y. Sun. Evolutionary Clustering and Analysis of Bibliographic Networks, *ASONAM Conference*, 2011.

[193] M. Gupta, C. Aggarwal, and J. Han. Finding Top-k Shortest Path Distance Changes in an Evolutionary Network, *SSDBM Conference*, 2011.

[194] M. Gupta, J. Gao, Y. Sun, and J. Han. Community Trend Outlier Detection Using Soft Temporal Pattern Mining. *ECML/PKDD Conference*, 2012.

[195] M. Gupta, J. Gao, Y. Sun, and J. Han. Integrating Community Matching and Outlier Detection for Mining Evolutionary Community Outliers. *KDD Conference*, 2012.

[196] D. Gusfield. Algorithms for Strings, Trees and Sequences, *Cambridge University Press*, 1997.

[197] D. Guthrie, L. Guthrie, and Y. Wilks. An Unsupervised Approach for the Detection of Outliers in Corpora, *LREC*, 2008.

[198] S. Guttormsson, R. Marks, M. El-Sharkawi, and I. Kerszenbaum. Elliptical Novelty Grouping for Online Short-turn Detection of Excited Running Rotors. *IEEE Transactions on Energy Conversion*, 14(1), pp. 16–22, 1999.

[199] R. Gwadera, M. Atallah, and W. Szpankowski. Markov Models for Identification of Significant Episodes, *SDM Conference*, 2005.

[200] R. Gwadera, M. Atallah, and W. Szpankowskii. Detection of Significant Sets of Episodes in Event Sequences, *IEEE ICDM Conference*, 2004.

[201] R. Gwadera, M. Atallah, and W. Szpankowski. Reliable Detection of Episodes in Event Sequences, *Knowledge and Information Systems*, 7(4), pp. 415–437, 2005.

[202] F. Hampel. A General Qualitative Definition of Robustness, *Annals of Mathematics and Statistics*, 43, pp. 1887–1896, 1971.

[203] J. Haslett, R. Brandley, P. Craig, A. Unwin, and G. Wills. Dynamic Graphics for Exploring Spatial Data With Application to Locating Global and Local Anomalies. *The American Statistician*, 45: pp. 234–242, 1991.

[204] V. Hautamaki, I. Karkkainen, and P. Franti. Outlier Detection using k-nearest neighbor graph. *International Conference on Pattern Recognition*, 2004.

[205] D. Hawkins. Identification of Outliers, *Chapman and Hall*, 1980.

[206] G. G. Hazel. Multivariate Gaussian MRF for Multispectral Scene Segmentation and Anomaly Detection, *GeoRS*, 38(3), pp. 1199–1211, 2000.

[207] J. He, and J. Carbonell. Nearest-Neighbor-Based Active Learning for Rare Category Detection. *CMU Computer Science Department*, Paper 281, 2007.
`http://repository.cmu.edu/compsci/281`

[208] Z. He, S. Deng, and X. Xu. Outlier Detection Integrating Semantic Knowledge. *Web Age Information Management (WAIM)*, 2002.

[209] Z. He, X. Xu, J. Huang, and S. Deng. FP-Outlier: Frequent Pattern-based Outlier Detection, *COMSIS*, 2(1), 2005.

[210] Z. He, X. Xu, and S. Deng. Discovering Cluster-based Local Outliers, *Pattern Recognition Letters*, Vol 24(9–10), pp. 1641–1650, 2003.

[211] Z. He, X. Xu, and S. Deng. An Optimization Model for Outlier Detection in Categorical Data. *International Conference on Intelligent Computing*, 2005.

[212] Z. He, S. Deng, X. Xu, and J. Huang. A Fast Greedy Algorithm for Outlier Mining. *PAKDD Conference*, 2006.

[213] M. Henrion, D. Hand, A. Gandy, and D. Mortlock. CASOS: A Subspace Method for Anomaly Detection in High Dimensional Astronomical Databases. *Statistical Analysis and Data Mining*, 2012. Online first: `http://onlinelibrary.wiley.com/doi/10.1002/sam.11167`

[214] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Statistical Outlier Detection using Direct Density Ratio Estimation. *Knowledge and information Systems*, 26(2), pp. 309–336, 2011.

[215] A. Hinneburg, C. Aggarwal, and D. Keim. What is the nearest neighbor in high-dimensional spaces?, *VLDB Conference*, 2000.

[216] H. O. Hirschfeld. A connection between correlation and contingency, *Proc. Cambridge Philosophical Society*, 31, pp. 520–524, 1935.

[217] S.-S. Ho. A Martingale Framework for Concept Change Detection in Time-Varying Data Streams, *ICML Conference*, 2005.

[218] V. Hodge and J. Austin. A Survey of Outlier Detection Methodologies, *Artifical Intelligence Review*, 22(2), pp. 85–126, 2004.

[219] J. Hodges. Efficiency in normal samples and tolerence of extreme values for some estimates of location, *Fifth Berkeley Symposium on Mathematics, Statistics and Probability*, 1, pp. 163–168, 1967.

[220] H. Hoffmann. Kernel PCA for Novelty Detection, *Pattern Recognition*, 40(3), pp. 863–874, 2007.

[221] T. Hofmann. Probabilistic Latent Semantic Indexing. *ACM SIGIR Conference*, 1999.

[222] S. Hofmeyr, S. Forrest, and A. Somayaji. Intrusion Detection using Sequences of System Calls, *Journal of Computer Security*, 6(3), pp. 151–180, 1998.

[223] J. H. Holland. Adaptation in Natural and Artificial Systems. *University of Michigan Press*, Ann Arbor, MI, 1975.

[224] G. Hollier, and J. Austin. Novelty Detection for Strain-gauge Degradation using Maximally Correlated Components. *European Symposium on Artificial Neural Networks*, 2002.

[225] J. Hollmen, and V. Tresp. Call-based Fraud Detection in Mobile Communication Networks using a Hierarchical Regime-switching Model. *NIPS Conference*, pp. 889–895, 1998.

[226] P. Horn, L. Feng, Y. Li, and A. J. Pesce. Effect of Outliers and Nonhealthy Individuals on Reference Interval Estimation. *Clinical Chemistry*, 47(12), pp. 2137–2145, 2001.

[227] L. Huang, M. I. Jordan, A. Joseph, M. Garofalakis, and N. Taft. In-network PCA and anomaly detection, *NIPS Conference*, 2006.

[228] J. W. Hunt and T. G. Szymanski. A Fast Algorithm for Computing Longest Common Subsequences, *Communications of the ACM*, 20(5), pp. 350–353, 1977.

[229] T. Ide, and H. Kashima. Eigenspace-based Anomaly Detection in Computer Systems. *ACM KDD Conference*, 2004.

[230] P. Indyk, R. Motwani, P. Raghavan, and S. Vempala. Locality-preserving Hashing in Multidimensional Spaces. *ACM STOC Conference*, 1997.

[231] D. Jackson, and Y. Chen. Robust Principal Component Analysis and Outlier Detection with Ecological Data, *Environmentrics*, 15, pp. 129–139, 2004.

[232] A. Jain, and R. Dubes. Algorithms for Clustering Data, *Prentice Hall*, 1988.

[233] H. Jagadish, N. Koudas, and S. Muthukrishnan. Mining Deviants in a Time-Series Database, *VLDB Conference*, 1999.

[234] V. Janeja and V. Atluri. Random Walks to Identify Anomalous Free-form Spatial Scan Windows. *IEEE Transactions on Knowledge and Data Engineering*, 20(10), 2008.

[235] P. Janeja and V. Atluri. Spatial Outlier Detection in Heterogeneous Neighborhoods. *Intelligent Data Analysis*, 13(1), 2008.

[236] H. Javitz, and A. Valdez. The SRI IDES Statistical Anomaly Detector. *IEEE Symposium on Security and Privacy*, 1991.

[237] Y. Jeong, M. Jeong, and O. Omitaomu, Weighted Dynamic Time Warping for Time Series Classification, *Pattern Recognition*, 44, pp. 2231–2240, 2010.

[238] B. Jiang, and J. Pei. Outlier Detection on Uncertain Data: Objects, Instances, and Inferences, *ICDE Conference*, 2011.

[239] M. F. Jiang, S. S. Tseng, and C. M. Su. Two-phase Clustering Process for Outliers Detection. *Pattern Recognition Letters*, 22, 6–7, pp. 691–700, 2001.

[240] R. Jiang, H. Fei, and J. Huan. Anomaly Localization for Network Data Streams with Graph Joint Sparse PCA. *ACM KDD Conference*, 2011.

[241] W. Jin, A. Tung, and J. Han. Mining Top-n Local Outliers in Large Databases. *ACM KDD Conference*, 2001.

[242] W. Jin, A. Tung, J. Han, and W. Wang. Ranking outliers using symmetric neighborhood relationship. *PAKDD Conference*, 2006.

[243] T. Johnson, I. Kwok, and R. Ng. Fast Computation of 2-dimensional Depth Contours. *ACM KDD Conference*, 1998.

[244] I. Jolliffe. Principal Component Analysis, *Springer*, 2002.

[245] M. Joshi, R. Agarwal, and V. Kumar. Mining Needles in a Haystack: Classifying Rare Classes via Two-Phase Rule Induction, *ACM SIGMOD Conference*, 2001.

[246] M. Joshi, V. Kumar, and R. Agarwal. Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements. *ICDM Conference*, pp. 257–264, 2001.

[247] M. Joshi, and R. Agarwal. PNRule: A Framework for Learning Classifier Models in Data Mining (A Case Study in Network Intrusion Detection), *SDM Conference*, 2001.

[248] M. Joshi, R. Agarwal, and V. Kumar. Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong? *ACM KDD Conference*, 2002.

[249] M. Joshi. On Evaluating Performance of Classifiers on Rare Classes, *ICDM Conference*, 2003.

[250] P. Juszczak and R. P. W. Duin. Uncertainty Sampling Methods for One-class Classifiers. *ICML Workshop on Learning from Imbalanced Data Sets*, 2003.

[251] J. Kang, S. Shekhar, C. Wennen, and P. Novak. Discovering Flow Anomalies: A SWEET Approach. *ICDM Conference*, 2008.

[252] G. Karakoulas and J. Shawe-Taylor. Optimising Classifiers for Imbalanced Training Sets, *NIPS*, 1998.

[253] D. R. Karger. Random sampling in cut, flow, and network design problems, *STOC*, pp. 648–657, 1994.

[254] S. Kasiviswanathan, P. Melville, and A. Banerjee. Emerging Topic Detection using Dictionary Learning, *CIKM Conference*, 2011.

[255] L. Kaufman, and P. Rousseeauw. Finding Groups in Data: An Introduction to Cluster Analysis, *Wiley-Interscience*, 1990.

[256] F. Keller, E. Muller, K. Bohm. HiCS: High-Contrast Subspaces for Density-based Outlier Ranking, *IEEE ICDE Conference*, 2012.

[257] E. Keogh, S. Lonardi, and B. Y.-C. Chiu. Finding Surprising Patterns in a Time Series Database in Linear Time and Space. *ACM KDD Conference*, 2002.

[258] E. Keogh, J. Lin, and A. Fu. HOT SAX: Finding the Most Unusual Time Series Subsequence: Algorithms and Applications, *ICDM Conference*, 2005.

[259] E. Keogh, S. Lonardi, and C. Ratanamahatana. Towards Parameter-Free Data Mining. *ACM KDD Conference*, 2004.

[260] D. Kifer, S. Ben-David, and J. Gehrke. Detecting Change in Data Streams, *VLDB Conference*, 2004.

[261] E. Knorr, and R. Ng. Algorithms for Mining Distance-based Outliers in Large Datasets. *VLDB Conference*, 1998.

[262] E. Knorr, and R. Ng. Finding Intensional Knowledge of Distance-Based Outliers. *VLDB Conference*, 1999.

[263] E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based Outliers: Algorithms and Applications, *VLDB Journal*, 8(3), pp. 237–253, February 2000.

[264] J. Koh, M.-L. Lee, W. Hsu, and W. Ang. Correlation-based Attribute Outlier Detection in XML. *ICDE Conference*, 2008.

[265] G. Kollios, D. Gunopulos, N. Koudas, and S. Berchtold. Efficient Biased Sampling for Approximate Clustering and Outlier Detection

in Large Data Sets, *IEEE Transactions on Knowledge and Data Engineering*, 15(5), pp. 1170–1187, 2003.

[266] M. Kontaki, A. Gounaris, A. Papadopoulos, K. Tsichlas, and Y. Manolopoulos. Continuous Monitoring of Distance-based Outliers over Data Streams, *ICDE Conference*, 2011.

[267] K. Kontonasios and T. Bie. An Information-Theoretic Approach to Finding Noisy Tiles in Binary Databases, *SIAM Conference on Data Mining*, 2003.

[268] Y. Kou, C. T. Lu, and D. Chen. Spatial Weighted Outlier Detection, *SDM Conference*, 2006.

[269] H.-P. Kriegel, M. Schubert, and A. Zimek. Angle-based Outlier Detection in High-Dimensional Data, *ACM KDD Conference*, 2008.

[270] H.-P. Kriegel, P. Kroger, and A. Zimek. Outlier Detection Techniques, *Conference Tutorial at SIAM Data Mining Conference*, 2010. Tutorial Slides at: `http://www.siam.org/meetings/sdm10/tutorial3.pdf`

[271] H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek. Interpreting and Unifying Outlier Scores. *SDM Conference*, pp. 13–24, 2011.

[272] H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek. A General Framework for Increasing the Robustness of PCA-Based Correlation Clustering Algorithms. *SSDBM Conference*, 2008.

[273] H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek. Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data. *PAKDD Conference*, 2009.

[274] H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek. Outlier Detection in Arbitrarily Oriented Subspaces, *ICDM Conference*, 2012.

[275] C. Kruegel, and G. Vigna. Anomaly-detection of Web-based Attacks, *CCS*, 2005.

[276] C. Kruegel, D. Mutz, W. Robertson, and F. Valeur. Bayesian Event Classification for Intrusion Detection. *Computer Security Applications Conference*, 2003.

[277] C. Kruegel, T. Toth, and E. Kirda. Service Specific Anomaly Detection for Network Intrusion Detection. *ACM symposium on Applied computing*, 2002.

[278] M. Kubat and S. Matwin. Addressing the Curse of Imbalanced Training Sets: One Sided Selection. *ICML Conference*, 1997.

[279] L. Kuncheva. Change Detection in Streaming Multivariate Data using Likelihood Detectors, *IEEE Transactions on Knowledge and Data Engineering*, Preprint, PP(99), 2011.

[280] A. Lakhina, M. Crovella, and C. Diot. Mining Anomalies using Traffic Feature Distributions, *ACM SIGCOMM Conference*, pp. 217–228, 2005.

[281] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. *ACM SIGCOMM Conference*, pp. 219–230, 2004.

[282] G. Lanckriet, L. Ghaoui, and M. Jordan. Robust Novelty Detection with Single Class MPM, *NIPS*, 2002.

[283] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *ICML Conference*, 2001.

[284] T. Lane and C. Brodley. Temporal Sequence Learning and Data Reduction for Anomaly Detection, *ACM Transactions on Information and Security*, 2(3), pp. 295–331, 1999.

[285] T. Lane and C. Brodley. An Application of Machine Learning to Anomaly Detection, *NIST-NCSC National Information Systems Security Conference*, 1997.

[286] T. Lane, and C. Brodley. Sequence matching and learning in anomaly detection for computer security. *AI Approaches to Fraud Detection and Risk Management*, pp. 43–49, 1997.

[287] R. Lasaponara. On the use of Principal Component Analysis (PCA) for Evaluating Interannual Vegetation Anomalies from SPOT/VEGETATION NDVI Temporal Series. *Ecological Modeling*, 194(4), pp. 429–434, 2006.

[288] J. Laurikkala, M. Juholal, and E. Kentala. Informal Identification of Outliers in Medical Data, *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, pp. 20–24, 2000.

[289] A. Lazarevic, and V. Kumar. Feature Bagging for Outlier Detection, *ACM KDD Conference*, 2005.

[290] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava. A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection. *SIAM Conference on Data Mining*, 2003.

[291] S. Q. Le and T. B. Ho. An Association-based Dissimilarity Measure for Categorical Data. *Pattern Recognition Letters*, 26(16), pp. 2549–2557, 2005.

[292] J.-G. Lee, J. Han, and X. Li. Trajectory Outlier Detection: A Partition-and-detect Framework, *ICDE Conference*, 2008.

[293] W. Lee and B. Liu. Learning with Positive and Unlabeled Examples using Weighted Logistic Regression. *ICML Conference*, 2003.

[294] W. Lee, S. Stolfo, and P. Chan. Learning Patterns from Unix Execution Traces for Intrusion Detection, *AAAI workshop on AI methods in Fraud and Risk Management*, 1997.

[295] W. Lee, S. Stolfo, and K. Mok. Adaptive Intrusion Detection: A Data Mining Approach, *Artificial Intelligence Review*, 14(6), pp. 533–567, 2000.

[296] W. Lee, and S. Stolfo. Data Mining Approaches for Intrusion Setection. *Proceedings of the 7th USENIX Security Symposium*, 1998.

[297] W. Lee, and D. Xiang. Information Theoretic Measures for Anomaly Detection, *IEEE Symposium on Security and Privacy*, 2001.

[298] N. Lesh, M. J. Zaki, and M. Ogihara. Mining Features for Sequence Classification, *ACM KDD Conference*, 1999.

[299] C. Leslie, E. Eskin, and W. Noble. The Spectrum Kernel: A String Kernel for SVM Protein Classification, *Pacific Symposium on Biocomputing*, pp. 566–575, 2002.

[300] X. Li, J. Han, S. Kim, and H. Gonzalez. ROAM: Rule and Motif-based Anomaly Detection in Massive Moving Object Data Sets, *SDM Conference*, 2007.

[301] X. Li, B. Liu, and S. Ng. Negative Training Data can be Harmful to Text Classification, *EMNLP*, 2010.

[302] X. Li, Z. Li, J. Han, and J.-G. Lee. Temporal Outlier Detection in Vehicle Traffic Data. *ICDE Conference*, 2009.

[303] D. Lin. An Information-theoretic Definition of Similarity. *ICML Conference*, pp. 296–304, 1998.

[304] J. Lin, E. Keogh, A. Fu, and H. V. Herle. Approximations to Magic: Finding Unusual Medical Time Series, *Mining Medical Data (CBMS)*, 2005.

[305] J. Lin, E. Keogh, S. Lonardi, and B. Y.-C. Chiu. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. *DMKD Workshop*, 2003.

[306] B. Liu, W. S. Lee, P. Yu, and X. Li. Partially Supervised Text Classification, *ICML Conference*, 2002.

[307] B. Liu, Y. Dai, X. Li, W. S. Lee, P. Yu. Building Text Classifiers Using Positive and Unlabeled Examples. *ICDM Conference*, 2003.

[308] G. Liu, T. McDaniel, S. Falkow, and S. Karlin, Sequence Anomalies in the cag7 Gene of the Helicobacter Pylori Pathogenicity Island, *National Academy of Sciences of the United States of America*, 96(12), pp. 7011–7016, 1999.

[309] L. Liu, and X. Fern. Constructing Training Sets for Outlier Detection, *SDM Conference*, 2012.

[310] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation Forest. *ICDM Conference*, 2008.

[311] S. Lin, and D. Brown. An Outlier-based Data Association Method for Linking Criminal Incidents. *SIAM Conference On Data Mining*, 2003.

[312] X.-Y. Liu, J. Wu, and Z.-H. Zhou. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man and Cybernetics– Part B, Cybernetics*, 39(2), pp. 539–550, April 2009.

[313] X. Liu, X. Wu, H. Wang, R. Zhang, J. Bailey, and K. Ramamohanarao. Mining Distribution Change in Stock Order Data Streams, *ICDE Conference*, 2010.

[314] Z. Liu, W. Shi, D. Li, and Q. Qin. Partially Supervised Classification – based on Weighted Unlabeled Samples Support Vector Machine. *ADMA*, 2005.

[315] X. Liu, P. Zhang, and D. Zeng. Sequence Matching for Suspicious activity Detection in Anti-money Laundering. *Lecture Notes in Computer Science*, Vol. 5075, pp. 50–61, 2008.

[316] S. Loncarin. A Survey of Shape Analysis Techniques. *Pattern Recognition*, 31(5), pp. 983–1001, 1998.

[317] C.-T. Lu, D. Chen, and Y. Kou. Algorithms for Spatial Outlier Detection, *ICDM Conference*, 2003.

[318] J. Ma and S. Perkins. Online Novelty Detection on Temporal Sequences, *ACM KDD Conference*, 2003.

[319] J. Ma, L. Saul, S. Savage, and G. Volker. Learning to Detect Malicious URLs, *ACM Transactions on Intelligent Systems and Technology*, 2(3), Article 30, April 2011.

[320] M. Mahoney, and P. Chan. Learning Nonstationary Models of Normal Network Traffic for Detecting Novel Attacks, *ACM KDD Conference*, 2002.

[321] M. Mahoney, and P. Chan. Learning Rules for Anomaly Detection of Hostile Network Traffic, *ICDM Conference*, 2003.

[322] F. Malliaros, V. Megalooikonomou, and C. Faloutsos. Fast Robustness Estimation in Large Social Graphs: Communities and Anomaly Detection. *SDM Conference*, 2012.

[323] L. M. Manevitz and M. Yousef. One-class SVMs for Document Classification, *Journal of Machine Learning Research*, 2: pp, 139–154, 2001.

[324] C. Marceau. Characterizing the Behavior of a Program using Multiple-length n-grams, *Workshop on New Security Paradigms*, pp. 101–110, 2000.

[325] M. Markou and S. Singh. Novelty detection: A Review, Part 1: Statistical Approaches, *Signal Processing*, 83(12), pp. 2481–2497, 2003.

[326] M. Markou and S. Singh. Novelty Detection: A Review, Part 2: Neural Network-based Approaches, *Signal Processing*, 83(12), pp. 2481–2497, 2003.

[327] M. Masud, Q. Chen, L. Khan, C. Aggarwal, J. Gao, J, Han, and B. Thuraisingham. Addressing Concept-Evolution in Concept-Drifting Data Streams. *ICDM Conference*, 2010.

[328] M. Masud, T. Al-Khateeb, L. Khan, C. Aggarwal, J. Gao, J. Han, and B. Thuraisingham. Detecting Recurring and Novel Classes in Concept-Drifting Data Streams. *ICDM Conference*, 2011.

[329] M. Masud, Q. Chen, L. Khan, C. Aggarwal, J. Gao, J. Han, A. Srivastava, and N. Oza. Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams, *IEEE Transactions on Knowledge and Data Engineering*, to appear, Online verion appeared on May 22, 2012.
`http://doi.ieeecomputersociety.org/10.1109/TKDE.2012.109`.

[330] C. McDiarmid. On the Method of Bounded Differences. In *Surveys in Combinatorics*, pp. 148–188, *Cambridge University Press*, Cambridge, 1989.

[331] P. Melville and R. Mooney. Diverse Ensembles for Active Learning, *ICML Conference*, 2004.

[332] C. Michael and A. Ghosh. Two State-based Approaches to Program-based Anomaly Detection, *Computer Security Applications Conference*, pp. 21, 2000.

[333] B. Miller, N. Bliss, and P. Wolfe. Subgraph Detection using Eigenvector L1-Norms. *NIPS Conference*, 2010.

[334] B. Miller, M. Beard, and N. Bliss. Eigenspace Analysis for Threat Detection in Social Networks. *International Conference on Information Fusion*, 2011.

[335] D. Mladenic and M. Grobelnik. Feature Selection for Unbalanced Class Distribution and Naive Bayes. *ICML Conference*, 1999.

[336] M. Mongiovi, P. Bogdanov, R. Ranca, A. Singh, E. Papalexakis, and C. Faloutsos. SIGSPOT: Mining Significant Anomalous Regions from Time-evolving Networks. *ACM SIGMOD Conference*, 2012.

[337] E. Muller, M. Schiffer, and T. Seidl. Statistical Selection of Relevant Subspace Projections for Outlier Ranking. *ICDE Conference*, pp, 434–445, 2011.

[338] E. Muller, M. Schiffer, P. Gerwert, M. Hannen, T. Jansen, and T. Seidl, SOREX: Subspace Outlier Ranking Exploration Toolkit, *Joint ECML PKDD Conference*, 2010.

[339] E. Muller, M. Schiffer, and T. Seidl. Adaptive Outlierness for Subspace Outlier Ranking, *CIKM Conference*, 2010.

[340] E. Muller, F. Keller, S. Blanc, and K. Bohm. OutRules: A Framework for Outlier Descriptions in Multiple Context Spaces. *ECML/PKDD Conference*, 2012.

[341] E. Muller, I. Assent, P. Iglesias, Y. Mulle, K. Bohm. Outlier Analysis via Subspace Analysis in Multiple Views of the Data, *ICDM Conference*, 2012.

[342] R. Motwani, and P. Raghavan. Randomized Algorithms, *Cambridge University Press*, 1995.

[343] A. Mueen, E. Keogh, and N. Young. Logical-Shapelets: An Expressive Primitive for Time Series Classification, *ACM KDD Conference*, 2011.

[344] A. Naftel and S. Khalid. Classifying Spatiotemporal Object Trajectories using Unsupervised Learning in the Coefficient Feature Space. *Multimedia Systems*, 12(3), pp. 227–238, 2006.

[345] K. Narita, and H. Kitagawa. Outlier Detection for Transaction Databases using Association Rules, *WAIM*, 2008.

[346] H. Nguyen, V. Gopalkrishnan, and I. Assent, An Unbiased Distance-based Outlier Detection Approach for High Dimensional Data, *DASFAA*, 2011.

[347] V. Niennattrakul, E. Keogh, and C. Ratanamahatana. Data Editing Techniques to Allow the Applicability of Distance-based Outlier Detection in Streams, *ICDM Conference*, 2010.

[348] H. Ning, W. Xu, Y. Chi, Y. Gong, and T. Huang. Incremental Spectral Clustering With Application to Monitoring of Evolving Blog Communities. *SDM Conference*, 2007.

[349] C. Noble, and D. Cook. Graph-based Anomaly Detection, *ACM KDD Conference*, 2003.

[350] P. Olmo Vaz de Melo, L. Akoglu, C. Faloutsos, and A. Loureiro. Surprising Patterns for the Call Duration Distribution of Mobile Phone Users, *ECML/PKDD Conference*, 2010.

[351] M. Otey, S. Parthasarathy, A. Ghoting, G. Li, S. Narravula, and D. Panda. Towards NIC-based Intrusion Detection. *ACM KDD Conference*, 2003.

[352] M. Otey, S. Parthasarathy, and A. Ghoting. Fast Distributed Outlier Detection in Mixed Attribute Data Sets, *Data Mining and Knowledge Discovery*, 12(2–3), pp. 203–228, 2006.

[353] C. R. Palmer and C. Faloutsos. Electricity based External Similarity of Categorical Attributes. *PAKDD Conference*, 2003.

[354] Y. Panatier. Variowin. Software For Spatial Data Analysis in 2D. *New York: Springer-Verlag*, 1996.

[355] C. Papadimitriou, P. Raghavan, H. Tamakai, and S. Vempala. Latent Semantic Indexing: A Probabilistic Analysis, *ACM PODS Conference*, 1998.

[356] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos. LOCI: Fast Outlier Detection using the Local Correlation Integral. *ICDE Conference*, 2003.

[357] S. Papadimitriou, J. Sun, and C. Faloutsos. SPIRIT: Streaming pattern discovery in multiple time-series. *VLDB Conference*, 2005.

[358] L. Parra, G. Deco, and S. Andmiesbach. Statistical Independence and Novelty Detection with Information Preserving Nonlinear Maps. *Neural Computation*, 8(2), pp. 260–269, 1996.

[359] Y. Pei, O. Zaiane, and Y. Gao. An Efficient Reference-based Approach to Outlier Detection in Large Datasets. *ICDM Conference*, 2006.

[360] D. Pelleg, and A. Moore. Active Learning for Anomaly and Rare Category Detection, *NIPS Conference*, 2004.

[361] Z. Peng, and F. Chu. Review Application of the Wavelet Transform in Machine Condition Monitoring and Fault Diagnostics, *Mechanical Systems and Signal Processing*, 18(2), pp. 199–221, March 2004.

[362] S. Petrovic, M. Osborne, and V. Lavrenko. Streaming First Story Detection with Application to Twitter. *Proceedings of the ACL Conference*, pp. 181–189, 2010.

[363] N. Pham, and R. Pagh. A Near-linear Time Approximation Algorithm for Angle-based Outlier Detection in High-dimensional Data, *ACM KDD Conference*, 2012.

[364] J. Pickands. Statistical Inference using Extreme Order Statistics. *The Annals of Statistics*, 3(1), pp. 119–131, 1975.

[365] B. Pincombe. Anomaly Detection in Time Series of Graphs using ARMA Processes. *ASOR Bulletin*, 24(4): 2–10, 2005.

[366] M. Pinsky. Introduction to Fourier Analysis and Wavelets, *American Mathematical Society*, 2009.

[367] C. Phua, V. Lee, K. Smith, and R. Gayler. A Comprehensive Survey of Data Mining-based Fraud Detection Research. `http://arxiv.org/abs/1009.6119`.

[368] C. Phua, D. Alahakoon, and V. Lee. Minority Report in Fraud Detection: Classification of Skewed Data, *ACM SIGKDD Explorations Newsletter*, 6(1), pp. 50–59, 2004.

[369] D. Pokrajac, A. Lazerevic, and L. Latecki. Incremental Local Outlier Detection for Data Streams, *CIDM Conference*, 2007.

[370] C. Potter, P. N. Tan, M. Steinbach, S. Klooster, V. Kumar, R. Myneni, and V. Genovese. Major Disturbance Events in Terrestrial Ecosystems detected using Global Satellite Data Sets. *Global Change Biology*, pp. 1005–1021, 2003.

[371] A. Pires, and C. Santos-Pereira. Using Clustering and Robust Estimators to Detect Outliers in Multivariate Data. *International Conference on Robust Statistics*, 2005.

[372] L. Portnoy, E. Eskin, and S. Stolfo. Intrusion Detection with Unlabeled Data using Clustering. *ACM Workshop on Data Mining Applied to Security*, 2001.

[373] B. Prakash, N. Valler, D. Andersen, M. Faloutsos, and C. Faloutsos. BGP-lens: Patterns and Anomalies in Internet Routing Updates. *ACM KDD Conference*, 2009.

[374] M. Prastawa, E. Bullitt, S. Ho, and G. Gerig. A Brain Tumor Segmentation Framework based on Outlier Detection, *Medical Image Analysis*, 8, pp. 275–283, 2004.

[375] C. Priebe, J. Conroy, D. Marchette, and Y. Park. Scan Statistics on Enron Graphs, *Computational and Mathematical Organizational Theory*, 11(3), pp. 229–247, 2005.

[376] F. Provost, and T. Fawcett. Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions, *ACM KDD Conference*, 1997.

[377] F. Provost, T. Fawcett, and R. Kohavi. The Case against Accuracy Estimation while Comparing Induction Algorithms, *ICML Conference*, 1998.

[378] G. Qi, C. Aggarwal, and T. Huang. On Clustering Heterogeneous Social Media Objects with Outlier Links, *WSDM Conference*, 2012.

[379] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, 77(2), pp. 257–285, Feb. 1989.

[380] M. Radovanovic, A. Nanopoulos, and M. Ivanovic. On the Existence of Obstinate Results in Vector Space Models, *ACM SIGIR Conference*, 2010.

[381] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient Algorithms for Mining Outliers from Large Data Sets. *ACM SIGMOD Conference*, pp. 427–438, 2000.

[382] B. Raskutti and A. Kowalczyk. Extreme Rebalancing for SVMS: A Case Study. *SIGKDD Explorations*, 6(1): pp. 60–69, 2004.

[383] S. Roberts. Novelty Detection using Extreme Value Statistics, *IEEE Proceedings on Vision, Image and Signal Processing*, 146(3). pp. 124–129, 1999.

[384] S. Roberts. Extreme Value Statistics for Novelty Detection in Biomedical Signal Processing. *International Conference on Advances in Medical Signal and Information Processing.* pp. 166–172, 2002.

[385] J. Rogan, J. Miller, D. Stow, J. Franklin, L. Levien, and C. Fischer. Land-Cover Change Monitoring with Classification Trees Using Landsat TM and Ancillary Data. *Photogrammetric Engineering and Remote Sensing*, 69(7), pp. 793–804, 2003.

[386] D. Ron, Y. Singer, and N. Tishby. The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length, *Machine Learning*, 25(2–3) pp. 117–149, 1996.

[387] P. Rousseeuw and A. Leroy. Robust Regression and Outlier Detection. *Wiley*, 2003.

[388] I. Ruts, and P. Rousseeuw, Computing Depth Contours of Bivariate Point Clouds. *Computational Statistics and Data Analysis*, 23, pp. 153–168, 1996.

[389] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz. A Bayesian approach to filtering junk e-mail. *AAAI Workshop on Learning for Text Categorization. Tech. Rep. WS-98-05.* `http://robotics.stanford.edu/users/sahami/papers.html`

[390] R. K. Sahoo, A. J. Oliner, I. Rish, M. Gupta, J. E. Moreira, S. Ma, R. Vilalta, and A. Sivasubramaniam. Critical Event Prediction for Proactive Management in Large-scale Computer Clusters. *ACM KDD Conference*, 2003.

[391] G. Salton, and M. J. McGill. Introduction to Modern Information Retrieval, *McGraw Hill*, 1986.

[392] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven Exploration of OLAP Data Cubes. *EDBT Conference*, 1998.

[393] G. Scarth, M. McIntyre, B. Wowk, and R. Somorjai. Detection of Novelty in Functional Images using Fuzzy Clustering. *Meeting of International Society for Magnetic Resonance in Medicine*, 1995.

[394] R. Schapire and Y. Singer. Improved Boosting Algorithms using Confidence-rated Predictions. *Annual Conference on Computational Learning Theory*, 1998.

[395] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel. On Evaluation of Outlier Rankings and Outlier Scores, *SDM Conference*, 2012.

[396] B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7), pp. 1443–1472, 2001.

[397] B. Scholkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt. Support-vector Method for Novelty Detection, *NIPS Conference*, 2000.

[398] R. Schoen, T. Habetler, F. Kamran, and R. Bartfield. Motor Bearing Damage Detection using Stator Current Monitoring. *IEEE Transactions on Industry Applications*, 31(6), pp. 1275–1279, 1995.

[399] K. Sequeira, and M. Zaki. ADMIT: Anomaly-based Data Mining for Intrusions, *ACM KDD Conference*, 2002.

[400] H. Seung, M. Opper, and H. Sompolinsky. Query by Committee. *ACM Workshop on Computational Learning Theory*, 1992.

[401] S. Shekhar, C. T. Lu, and P. Zhang. Detecting Graph-based Spatial Outliers: Algorithms and Applications, *ACM KDD Conference*, 2001.

[402] S.Shekhar, C. T. Lu, and P. Zhang. A Unified Approach to Detecting Spatial Outliers, *Geoinformatica*, 7(2), pp. 139–166, 2003.

[403] S. Shekhar and S. Chawla. A Tour of Spatial Databases. *Prentice Hall*, 2002.

[404] S. Shekhar, C. T. Lu, and P. Zhang. Detecting Graph-based Spatial Outliers, *Intelligent Data Analysis*, 6, pp. 451–468, 2002.

[405] P. Showbridge, M. Kraetzl, and D. Ray. Detection of Abnormal Change in Dynamic Networks. *Proceedings of the Intl. Conf. on Information, Decision and Control*, pp. 557–562, 1999.

[406] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn, and L. Chang. A Novel Anomaly Detection Scheme based on Principal Component Classifier, *ICDM Conference*, 2003.

[407] A. Siebes, J. Vreeken, and M. van Leeuwen. Itemsets than Compress, *SIAM Conference on Data Mining*, 2006.

[408] J. Silva, and R. Willett. Detection of Anomalous Meetings in a Social Network, *SocialCom*, 2008.

[409] B. W. Silverman. Density Estimation for Statistics and Data Analysis. *Chapman and Hall*, 1986.

[410] K. Smets and J. Vreeken. The Odd One Out: Identifying an Characterising Anomalies, *SIAM Conference on Data Mining*, 2011.

[411] E. S. Smirnov. On exact methods in systematics. *Systematic Zoology*, 17(1), pp. 1–13, 1968.

[412] R. Smith, A. Bivens, M. Embrechts, C. Palagiri, and B. Szymanski. Clustering Approaches for Anomaly Based Intrusion Detection. *Intelligent Engineering Systems through Artificial Neural Networks*, 2002.

[413] P. Smyth. Clustering Sequences with Hidden Markov Models, *Neural Information Processing*, 1997.

[414] P. Smyth. Markov Monitoring with Unknown States. *IEEE Journal on Selected Areas in Communications*, 12(9), pp. 1600-1612, 1994.

[415] H. Solberg, and A. Lahti. Detection of Outliers in Reference Distributions: Performance of Horn's Algorithm, *Clinical Chemistry*, 51(12), pp. 2326–2332, 2005.

[416] X. Song, M. Wu, C. Jermaine, and S. Ranka. Conditional Anomaly Detection, *IEEE Transaction on Knowledge and Data Engineering*, 19(5), pp. 631–645, 2007.

[417] X. Song, M. Wu, C. Jermaine, and S. Ranka. Statistical Change Detection for Multidimensional Data, *ACM KDD Conference*, 2007.

[418] C. Spence, L. Parra, and P. Sajda. Detection, Synthesis and Compression in Mammographic Image Analysis with a Hierarchical Image Probability Model. *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, 2001.

[419] A. Srivastava. Discovering System Health Anomalies using Data Mining Techniques, *Joint Army Navy NASA Airforce Conference on Propulsion*, 2005.

[420] A. Srivastava. Enabling the Discovery of Recurring Anomalies in Aerospace Problem Reports using High-dimensional Clustering Techniques. *Aerospace Conference*, 2006.

[421] A. Srivastava, and B. Zane-Ulman. Discovering Recurring Anomalies in Text Reports regarding Complex Space Systems. *Aerospace Conference*, 2005.

[422] A. Srivastava, A. Kundu, S. Sural, and A. Majumdar. Credit card fraud detection using Hidden Markov Model. *IEEE Transactions on Dependable and Secure Computing*, 5(1), pp. 37–48, 2008.

[423] P. Srivastava, D. Desai, S. Nandi, and A. Lynn. HMM-ModE-Improved Classification using Profile Hidden Markov Models by Optimizing the Discrimination Threshold and Modifying Emission Probabilities with Negative Training Sequences, *BMC Bioinformatics*, 8 (104), 2007.

[424] M. Stephens. Use of the Kolmogorov-Smirnov, Cramer-von Mises and Related Statistics without Extensive Tables, *Journal of the Royal Statistical Society*. Series B, pp. 115–122, 1970.

[425] S. Stolfo, D. Fan, W. Lee, A.L. Prodromidis, and P. Chan. Credit Card Fraud Detection Using Meta-Learning: Issues and Initial Results, *AAAI Workshop AI Methods in Fraud and Risk Management*, pp. 83–90, 1997.

[426] S. Stolfo, D. Fan, W. Lee, A. Prodromidis, and P. Chan. Cost-Based Modeling for Fraud and Intrusion Detection: Results from the JAM Project, *DARPA Information Survivability Conf. and Exposition*, 2, pp. 130–144, 2000.

[427] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos. Online Outlier Detection in Sensor Data using Non-parametric Models. *VLDB Conference*, 2006.

[428] H. Sun, Y. Bao., F. Zhao, G. Yu, and D. Wang. CD-Trees: An Efficient Index Structure for Outlier Detection. *Web-Age Information Management (WAIM)*, pp. 600–609, 2004.

[429] J. Sun, S. Papadimitriou, P. Yu, and C. Faloutsos. Graphscope: Parameter-free Mining of Large Time-Evolving Graphs, *ACM KDD Conference*, 2007.

[430] J. Sun, D. Tao, and C. Faloutsos. Beyond Streams and Graphs: Dynamic Tensor Analysis, *ACM KDD Conference*, 2006.

[431] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood Formation and Anomaly Detection in Bipartite Graphs. *ICDM Conference*, 2005.

[432] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos. Less is More: Compact Matrix Representation of Large Sparse Graphs. *SIAM Conference on Data Mining*, 2007.

[433] P. Sun, and S. Chawla. On Local Spatial Outliers, *IEEE ICDM Conference*, 2004.

[434] P. Sun, S. Chawla, and B. Arunasalam. Mining for Outliers in Sequential Databases, *SIAM International Conference on Data Mining*, 2006.

[435] Y. Sun, Y. Yu, and J. Han. Ranking-based Clustering of Heterogeneous Information Networks with Star Network Schema. *ACM KDD Conference*, 2009.

[436] C. Surace, and K. Worden. A Novelty Detection Method to Diagnose Damage in Structures: An Application to an Offshore Platform. *International Conference of Offshore and Polar Engineering*, 4, pp. 64–70, 1998.

[437] C. Surace, K. Worden, and G. Tomlinson. A Novelty Detection Approach to Diagnose Damage in a Cracked Beam. *Proceedings of SPIE*, Vol. 3089, pp. 947–953, 1997.

[438] E. Suzuki, T. Watanabe, H. Yokoi, and K. Takabayashi. Detecting Interesting Exceptions from Medical Test Data with Visual Summarization. *International Conference on Data Mining*, pp. 315–322, 2003.

[439] T. Takahashi, R. Tomioka, and K. Yamanishi. Discovering Emerging Topics in Social Streams via Link Anomaly Detection. *ICDM Conference*, 2011.

[440] P.-N. Tan and V. Kumar. Discovery of Web Robot Sessions based on their Navigational Patterns, *Data Mining and Knowledge Discovery*, 6(1), pp. 9–35, 2002.

[441] Y. Tao, X. Xiao, and S. Zhou. Mining Distance-based Outliers from Large Databases in any Metric Space. *ACM KDD Conference*, 2006.

[442] J. Tang, Z. Chen, A. W.-C. Fu, D. W. Cheung. Enhancing Effectiveness of Outlier Detections for Low Density Patterns. *PAKDD Conference*, 2002.

[443] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser. SVMs Modeling for Highly Imbalanced Classification, *IEEE Transactions on Systems, Man and Cybernetics- Part B: Cybernetics*, 39(1), pp. 281–288, 2009.

[444] M. Taniguchi, M. Haft, J. Hollmen, and V. Tresp. Fraud Detection in Communications Networks using Neural and Probabilistic Methods. *IEEE International Conference in Acoustics, Speech and Signal Processing*, 2, pp. 1241–1444, 1998.

[445] D. Tax. One Class Classification: Concept-learning in the Absence of Counter-examples, *Doctoral Dissertation, University of Delft*, Netherlands, 2001.

[446] L. Tarassenko. Novelty detection for the Identification of Masses in Mammograms. *IEEE International Conference on Artificial Neural Networks*, 4, pp. 442–447, 1995.

[447] P. Thompson, D. MacDonald, M. Mega, C. Holmes, A. Evans, and A. Toga. Detection and Mapping of Abnormal Brain Structure with a Probabilistic Atlas of Cortical Surfaces. *Journal of Computer Assisted Tomography*, 21(4), pp. 567–581, 1997.

[448] M. Thottan, and C. Ji. Anomaly Detection in IP Networks. *IEEE Transactions on Signal Processing*, 51(8), pp. 2191–2204, 2003.

[449] S. Tian, S. Mu, and C. Yin. Sequence-similarity Kernels for SVMs to Detect Anomalies in System Calls, *Neurocomputing*, 70(4–6), pp. 859–866, 2007.

[450] K. M. Ting. An Instance-weighting Method to Induce Cost-sensitive Trees. *IEEE Transaction on Knowledge and Data Engineering*, 14: pp. 659–665, 2002.

[451] M. E. Tipping, and C. M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society*, B 61, pp. 611–622, 1999.

[452] H. Tong, and C.-Y. Lin. Non-Negative Residual Matrix Factorization with Application to Graph Anomaly Detection. *SDM Conference*, 2011.

[453] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri. A Review of Process Fault Detection and Diagnosis Part I: Quantitative Model-based Methods. *Computers and Chemical Engineering*, 27(3), pp. 293–311, 2003.

[454] J. S. Vitter. Random sampling with a reservoir, *ACM Trans. Math. Softw.*, vol. 11(1), pp. 37–57, 1985.

[455] W. Tobler. Cellular geography. In *Philosophy in Geography, Dordrecht Reidel Publishing Company*, pp. 379–386, 1979.

[456] S. Viaene, R. Derrig, B. Baesens, and G. Dedene. A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection, *Journal of Risk and Insurance*, 69(3), pp. 373–421, 2002.

[457] N. Wale, X. Ning, and G. Karypis. Trends in Chemical Data Mining, *Managing and Mining Graph Data*, *Springer*, 2010.

[458] X. Wang, C. Zhai, X. Hu and R. Sproat. Mining Correlated Bursty Topic Patterns from Coordinated Text Streams. *ACM KDD Conference*, 2007.

[459] B. Wang, G. Xiao, H. Yu, and X. Yang, Distance-based Outlier Detection on Uncertain Data, *International Conference on Computer and Information Technology*, 2009.

[460] B. Wang, X. Yang, G. Wang, and G. Yu. Outlier Detection over Sliding Windows for Probabilistic Data Streams, *Journal of Computer Science and Technology*, 25(3), pp. 389–400, 2010.

[461] L. Wei, W. Qian, A. Zhou, and W. Jin. HOT: Hypergraph-based Outlier Test for Categorical Data. *PAKDD Conference*, 2007.

[462] L. Wei and E. Keogh. Semi-supervised Time Series Classification, *ACM KDD Conference*, 2006.

[463] G. Weiss and F. Provost. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction, *Journal of Artificial Intelligence Reserach*, 19: pp. 315–354, 2003.

[464] G. Williams, R. Baxter, H. He, S. Hawkings, and L. Gu. A Comparative Study of RNN for Outlier Detection in Data Mining. *IEEE ICDM Conference*, 2002.

[465] W.-K. Wong, A. Moore, G. Cooper, and M. Wagner. Rule-based Anomaly Pattern Detection for Detecting Disease Outbreaks, *National Conference on Artificial Intelligence*, 2002.

[466] K. van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens. Automated Segmentation of Multiple Sclerosis Lesions by Model Outlier Detection, *IEEE Transactions on Medical Imaging*, vol. 20, pp. 677–688, August 2001.

[467] T. De Vries, S. Chawla, and M. Houle. Finding Local Anomalies in Very High Dimensional Space, *ICDM Conference*, 2010.

[468] M. Wang, C. Zhang, and J. Yu. Native API-based Windows Anomaly Intrusion Detection Method using SVM, *International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, 2006.

[469] L. Wei, E. Keogh, and X. Xi. SAXually Exaplicit Images: Finding Unusual Shapes, *ICDM Conference*, 2006.

[470] G. Wu and E. Y. Chang. Class-boundary Alignment for Imbalanced Dataset Learning. *Proceedings of the ICML Workshop on Learning from Imbalanced Data Sets*, 2003.

[471] M. Wu, and C. Jermaine. Outlier Detection by Sampling with Accuracy Guarantees. *ACM KDD Conference*, 2006.

[472] E. Wu, W. Liu, and S. Chawla. Spatio-temporal Outlier Detection in Precipitation Data, *Knowledge Discovery from Sensor Data, Springer, LNCS 5840*, 2008.

[473] M. Wu, X. Song, C. Jermaine, S. Ranka, and J. Gums. A LRT Framework for Fast Spatial Anomaly Detection. *ACM KDD Conference*, 2009.

[474] X. Xi, E. Keogh, C. Shelton, L.Wei, and C. Ratanamahatana. Fast Time Series Classification using Numerosity Reduction, *ICML Conference*, 2006.

[475] Z. Xing, J. Pei, and E. Keogh. A Brief Survey on Sequence Classification, *ACM SIGKDD Explorations*, 12(1), 2010.

[476] L. Xiong, X. Chen, and J. Schneider. Direct Robust Matrix Factorization for Anomaly Detection. *ICDM Conference*, 2011.

[477] L. Xiong, B. Poczos, J. Schneider, A. Connolly, and J. VanderPlas. Hierarchical Probabilistic Models for Group Anomaly Detection, *Artificial Intelligenece and Statitistics*, 2011.

[478] K. Yaminshi, J. Takeuchi, and G. Williams. Online Unsupervised Outlier Detection using Finite Mixtures with Discounted Learning Algorithms, *ACM KDD Conference*, 2000.

[479] K. Yaminshi, and J. Takeuchi. A Unified Framework for Detecting Outliers and Change Points from Time Series Data, *ACM KDD Conference*, 2002.

[480] R. Yan, Y. Liu, R. Jin, and A. Hauptmann. On Predicting Rare Classes with SVM Ensembles in Scene Classification. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003.

[481] J. Yang, and W. Wang. CLUSEQ: Efficient and Effective Sequence Clustering, *ICDE Conference*, 2003.

[482] P. Yang, and Q. Zhu. Finding Key Outlying Subspaces for Outlier Detection, *Knowledge-based Systems*, 24(2), pp. 269–274, 2011.

[483] X. Yang, L. Latecki, and D. Pokrajac. Outlier Detection with Globally Optimal Exemplar-based GMM. *SDM Conference*, 2009.

[484] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. T. Archibald, and X. Liu. Learning Approaches for Detecting and Tracking News Events. *IEEE Intelligent Systems*, 14(4):32–43, 1999.

[485] Y. Yang, T. Pierce, and J. Carbonell. A Study on Retrospective and On-line Event Detection. *ACM SIGIR Conference*, 1998.

[486] Y.Yang, J. Zhang, J. Carbonell, and C. Jin. Topic-conditioned Novelty Detection. *ACM KDD Conference*, 2002.

[487] D. Yankov, E. Keogh, and U. Rebbapragada. Disk-aware Discord Discovery: Finding Unusual Time Series in Terabyte Sized Data Sets, *ICDM Conference*, 2007.

[488] N. Ye. A Markov Chain Model of Temporal Behavior for Anomaly Detection, *IEEE Information Assurance Workshop*, 2004.

[489] N. Ye, and Q. Chen. An Anomaly Detection Technique based on a Chi-square Statistic for Detecting Intrusions into Information Systems. *Quality and Reliability Engineering International*, 17, pp. 105–112, 2001.

[490] L. Ye and E. Keogh. Time Series Shapelets: a New Primitive for Data Mining. *ACM KDD Conference*, 2009.

[491] B.-K. Yi, N. D. Sidiropoulos, T. Johnson, H. Jagadish, C. Faloutsos, and A. Biliris. Online Data Mining for Co-evolving Time Sequences. *ICDE Conference*, 2000.

[492] D. Yu, G. Sheikholeslami, and A. Zhang. Findout: Finding Outliers in Very Large Datasets. *Knowledge And Information Systems*, 4(4), pp. 387–412, 2002.

[493] H. Yu, J. Han, and K. C.-C. Chang. PEBL: Web Page Classification without Negative Examples. *IEEE Transactions on Knowledge and Data Engineering*, 16(1), pp. 70–81, 2004.

[494] J. X. Yu, W. Qian, H. Lu, and A. Zhou. Finding Centric Local Outliers in Categorical/Numeric Spaces. *Knowledge and Information Systems*, 9(3), pp. 309–338, 2006.

[495] B. Zadrozny, and C. Elkan. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. *ACM KDD Conference*, 2002.

[496] B. Zadrozny, J. Langford, and N. Abe. Cost-Sensitive Learning by Cost-Proportionate Example Weighting, *ICDM Conference*, 2003.

[497] B. Zadrozny, and C. Elkan. Learning and Making Decisions when Costs and Probabilities are Unknown, *KDD Conference*, 2001.

[498] J. Zhang, Q. Gao, and H. Wang. SPOT: A System for Detecting Projected Outliers from High-Dimensional Data Stream, *ICDE Conference*, 2008.

[499] J. Zhang, M. Lou, T. W. Ling and H. Wang. HOS-Miner: A System for Detecting Outlying Subspaces of High-dimensional Data. *VLDB Conference*, 2004.

[500] J. Zhang, Q. Gao and H. Wang. A Novel Method for Detecting Outlying Subspaces in High-dimensional Databases Using Genetic Algorithm. *ICDM Conference*, 2006.

[501] J. Zhang and H. Wang. Detecting Outlying Subspaces for High-Dimensional Data: the New Task, Algorithms and Performance. *Knowledge and Information Systems*, 10(3), pp. 333–355, 2006.

[502] Y. Zhang, P. Meratnia, and P. Havinga. Outlier Detection for Wireless Sensor Networks: A Survey. *IEEE Communications Surveys and Tutorials*, 12(2), 2010.

[503] J. Zhang, Z. Ghahramani, and Y. Yang. A Probabilistic Model for Online Document Clustering with Application to Novelty Detection. *NIPS*, 2005.

[504] D. Zhang, and G. Lu. Review of Shape Representation and Description Techniques. *Pattern Recognition*, 37(1), pp. 1–19, 2004.

[505] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: An Efficient Data Clustering Method for Very Large Databases. *ACM SIGMOD Conference*, 1996.

[506] J. Zhang and I. Mani. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. *Proceedings of the ICML Workshop on Learning from Imbalanced Datasets*, 2003.

[507] D. Zhang and W. S. Lee. A Simple Probabilistic Approach to Learning from Positive and Unlabeled Examples. *Annual UK Workshop on Computational Intelligence*, pp. 83–87, 2005.

[508] X. Zhang, P. Fan, and Z. Zhu. A New Anomaly Detection Method based on Hierarchical HMM, *International Conference on Parallel and Distrbuted Computing, Applications, and Technologies*, 2003.

[509] Y. Zhang, J. Hong, and L. Cranor. CANTINA: A Content-based Approach to Detecting Phishing Web Sites. *WWW Conference*, 2007.

[510] J. Zhao, C.-T. Lu, and Y. Kou. Detecting Region Outliers in Meteorological Data. *ACM GIS Conference*, 2003.

[511] Z. Zheng, X. Wu, and R. Srihari. Feature Selection for Text Categorization on Imbalanced Data. *SIGKDD Explorations*, 6(1), pp. 80–89, 2004.

[512] C. Zhu, H. Kitagawa, S. Papadimitriou, and C. Faloutsos. OBE: Outlier by Example, *PAKDD Conference*, 2004.

[513] C. Zhu, H. Kitagawa, and C. Faloutsos. Example-based Robust Outlier Detection in High Dimensional Data Sets, *ICDM Conference*, 2005.

[514] A. Zimek, A. Schubert, and H.-P. Kriegel. A Survey on Unsupervised Outlier Detection in High-dimensional Numerical Data, *Journal on Statistical Analysis and Data Mining*, Preprint available at the online Wiley library: `http://onlinelibrary.wiley.com/doi/10.1002/sam.11161/abstract`, 2012.

[515] http://www.itl.nist.gov/iad/mig/tests/tdt/tasks/fsd.html

[516] D. D. Lewis. Reuters-21578 Data Set. http://www.daviddlewis.com/resources/test-collections/reuters21578.

[517] http://kdd.ics.uci.edu/databases/20newsgroups

[518] http://www.informatik.uni-trier.de/~ley/db/

[519] http://www.kdnuggets.com/software/deviation.html

[520] http://www.kdnuggets.com/software/index.html

[521] http://www.cs.waikato.ac.nz/ml/weka/

[522] http://www.cs.ucr.edu/~eamonn/time_series_data/

[523] http://www.cs.ucr.edu/~eamonn/SAX.htm

[524] http://www-935.ibm.com/services/nz/en/it-services/ibm-proventia-network-anomaly-detection-system-ads.html

[525] http://www.ibm.com/software/analytics/spss

[526] http://www-01.ibm.com/software/analytics/spss/products/statistics/

[527] http://www.sas.com/

[528] http://www.sas.com/software/security-intelligence/index.html

[529] http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/index.html

[530] http://www.wizsoft.com