10

Anomaly Detection

In anomaly detection, the goal is to find objects that are different from most other objects. Often, anomalous objects are known as **outliers**, since, on a scatter plot of the data, they lie far away from other data points. Anomaly detection is also known as **deviation detection**, because anomalous objects have attribute values that deviate significantly from the expected or typical attribute values, or as **exception mining**, because anomalies are exceptional in some sense. In this chapter, we will mostly use the terms *anomaly* or *outlier*.

There are a variety of anomaly detection approaches from several areas, including statistics, machine learning, and data mining. All try to capture the idea that an anomalous data object is unusual or in some way inconsistent with other objects. Although unusual objects or events are, by definition, relatively rare, this does not mean that they do not occur frequently in absolute terms. For example, an event that is "one in a thousand" can occur millions of times when billions of events are considered.

In the natural world, human society, or the domain of data sets, most events and objects are, by definition, commonplace or ordinary. However, we have a keen awareness of the possibility of objects that are unusual or extraordinary. This includes exceptionally dry or rainy seasons, famous athletes, or an attribute value that is much smaller or larger than all others. Our interest in anomalous events and objects stems from the fact that they are often of unusual importance: A drought threatens crops, an athlete's exceptional skill may lead to victory, and anomalous values in experimental results may indicate either a problem with the experiment or a new phenomenon to be investigated.

The following examples illustrate applications for which anomalies are of considerable interest.

- **Fraud Detection.** The purchasing behavior of someone who steals a credit card is probably different from that of the original owner. Credit card companies attempt to detect a theft by looking for buying patterns that characterize theft or by noticing a change from typical behavior. Similar approaches are used for other types of fraud.

- **Intrusion Detection.** Unfortunately, attacks on computer systems and computer networks are commonplace. While some of these attacks, such as those designed to disable or overwhelm computers and networks, are obvious, other attacks, such as those designed to secretly gather information, are difficult to detect. Many of these intrusions can only be detected by monitoring systems and networks for unusual behavior.

- **Ecosystem Disturbances.** In the natural world, there are atypical events that can have a significant effect on human beings. Examples include hurricanes, floods, droughts, heat waves, and fires. The goal is often to predict the likelihood of these events and the causes of them.

- **Public Health.** In many countries, hospitals and medical clinics report various statistics to national organizations for further analysis. For example, if all children in a city are vaccinated for a particular disease, e.g., measles, then the occurrence of a few cases scattered across various hospitals in a city is an anomalous event that may indicate a problem with the vaccination programs in the city.

- **Medicine.** For a particular patient, unusual symptoms or test results may indicate potential health problems. However, whether a particular test result is anomalous may depend on other characteristics of the patient, such as age and sex. Furthermore, the categorization of a result as anomalous or not incurs a cost—unneeded additional tests if a patient is healthy and potential harm to the patient if a condition is left undiagnosed and untreated.

Although much of the recent interest in anomaly detection has been driven by applications in which anomalies are the focus, historically, anomaly detection (and removal) has been viewed as a technique for improving the analysis of typical data objects. For instance, a relatively small number of outliers can distort the mean and standard deviation of a set of values or alter the set of clusters produced by a clustering algorithm. Therefore, anomaly detection (and removal) is often a part of data preprocessing.

In this chapter, we will focus on anomaly detection. After a few preliminaries, we provide a detailed discussion of some important approaches to anomaly detection, illustrating them with examples of specific techniques.

10.1 Preliminaries

Before embarking on a discussion of specific anomaly detection algorithms, we provide some additional background. Specifically, we (1) explore the causes of anomalies, (2) consider various anomaly detection approaches, (3) draw distinctions among approaches based on whether they use class label information, and (4) describe issues common to anomaly detection techniques.

10.1.1 Causes of Anomalies

The following are some common causes of anomalies: data from different classes, natural variation, and data measurement or collection errors.

Data from Different Classes An object may be different from other objects, i.e., anomalous, because it is of a different type or class. To illustrate, someone committing credit card fraud belongs to a different class of credit card users than those people who use credit cards legitimately. Most of the examples presented at the beginning of the chapter, namely, fraud, intrusion, outbreaks of disease, and abnormal test results, are examples of anomalies that represent a different class of objects. Such anomalies are often of considerable interest and are the focus of anomaly detection in the field of data mining.

The idea that anomalous objects come from a different source (class) than most of the data objects is stated in the often-quoted definition of an outlier by the statistician Douglas Hawkins.

Definition 10.1 (Hawkins' Definition of an Outlier). An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism.

Natural Variation Many data sets can be modeled by statistical distributions, such as a normal (Gaussian) distribution, where the probability of a data object decreases rapidly as the distance of the object from the center of the distribution increases. In other words, most of the objects are near a center (average object) and the likelihood that an object differs significantly from this average object is small. For example, an exceptionally tall person is not anomalous in the sense of being from a separate class of objects, but only

in the sense of having an extreme value for a characteristic (height) possessed by all the objects. Anomalies that represent extreme or unlikely variations are often interesting.

Data Measurement and Collection Errors Errors in the data collection or measurement process are another source of anomalies. For example, a measurement may be recorded incorrectly because of human error, a problem with the measuring device, or the presence of noise. The goal is to eliminate such anomalies, since they provide no interesting information but only reduce the quality of the data and the subsequent data analysis. Indeed, the removal of this type of anomaly is the focus of data preprocessing, specifically data cleaning.

Summary An anomaly may be a result of the causes given above or of other causes that we did not consider. Indeed, the anomalies in a data set may have several sources, and the underlying cause of any particular anomaly is often unknown. In practice, anomaly detection techniques focus on finding objects that differ substantially from most other objects, and the techniques themselves are not affected by the source of an anomaly. Thus, the underlying cause of the anomaly is only important with respect to the intended application.

10.1.2 Approaches to Anomaly Detection

Here, we provide a high-level description of some anomaly detection techniques and their associated definitions of an anomaly. There is some overlap between these techniques, and relationships among them are explored further in Exercise 1 on page 680.

Model-Based Techniques Many anomaly detection techniques first build a model of the data. Anomalies are objects that do not fit the model very well. For example, a model of the distribution of the data can be created by using the data to estimate the parameters of a probability distribution. An object does not fit the model very well; i.e., it is an anomaly, if it is not very likely under the distribution. If the model is a set of clusters, then an anomaly is an object that does not strongly belong to any cluster. When a regression model is used, an anomaly is an object that is relatively far from its predicted value.

Because anomalous and normal objects can be viewed as defining two distinct classes, classification techniques can be used for building models of these

two classes. Of course, classification techniques can only be used if class labels are available for some of the objects so that a training set can be constructed. Also, anomalies are relatively rare, and this needs to be taken into account when choosing both a classification technique and the measures to be used for evaluation. (See Section 5.7.)

In some cases, it is difficult to build a model; e.g., because the statistical distribution of the data is unknown or no training data is available. In these situations, techniques that do not require a model, such as those described below, can be used.

Proximity-Based Techniques It is often possible to define a proximity measure between objects, and a number of anomaly detection approaches are based on proximities. Anomalous objects are those that are distant from most of the other objects. Many of the techniques in this area are based on distances and are referred to as **distance-based outlier detection techniques**. When the data can be displayed as a two- or three-dimensional scatter plot, distance-based outliers can be detected visually, by looking for points that are separated from most other points.

Density-Based Techniques Estimates of the density of objects are relatively straightforward to compute, especially if a proximity measure between objects is available. Objects that are in regions of low density are relatively distant from their neighbors, and can be considered anomalous. A more sophisticated approach accommodates the fact that data sets can have regions of widely differing densities, and classifies a point as an outlier only if it has a local density significantly less than that of most of its neighbors.

10.1.3 The Use of Class Labels

There are three basic approaches to anomaly detection: unsupervised, supervised, and semi-supervised. The major distinction is the degree to which class labels (anomaly or normal) are available for at least some of the data.

Supervised anomaly detection Techniques for supervised anomaly detection require the existence of a training set with both anomalous and normal objects. (Note that there may be more than one normal or anomalous class.) As mentioned previously, classification techniques that address the so-called rare class problem are particularly relevant because

anomalies are relatively rare with respect to normal objects. See Section 5.7.

Unsupervised anomaly detection In many practical situations, class labels are not available. In such cases, the objective is to assign a score (or a label) to each instance that reflects the degree to which the instance is anomalous. Note that the presence of many anomalies that are similar to each other can cause them all to be labeled normal or have a low outlier score. Thus, for unsupervised anomaly detection to be successful, anomalies must be distinct from one another, as well as normal objects.

Semi-supervised anomaly detection Sometimes training data contains labeled normal data, but has no information about the anomalous objects. In the semi-supervised setting, the objective is to find an anomaly label or score for a set of given objects by using the information from labeled normal objects. Note that in this case, the presence of many related outliers in the set of objects to be scored does not impact the outlier evaluation. However, in many practical situations, it can be difficult to find a small set of representative normal objects.

All anomaly detection schemes described in this chapter can be used in supervised or unsupervised mode. Supervised schemes are essentially the same as classification schemes for rare classes discussed in Section 5.7.

10.1.4 Issues

There are a variety of important issues that need to be addressed when dealing with anomalies.

Number of Attributes Used to Define an Anomaly The question of whether an object is anomalous based on a single attribute is a question of whether the object's value for that attribute is anomalous. However, since an object may have many attributes, it may have anomalous values for some attributes, but ordinary values for other attributes. Furthermore, an object may be anomalous even if none of its attribute values are individually anomalous. For example, it is common to have people who are two feet tall (children) or are 300 pounds in weight, but uncommon to have a two-foot tall person who weighs 300 pounds. A general definition of an anomaly must specify how the values of multiple attributes are used to determine whether or not an object is an anomaly. This is a particularly important issue when the dimensionality of the data is high.

Global versus Local Perspective An object may seem unusual with respect to all objects, but not with respect to objects in its local neighborhood. For example, a person whose height is 6 feet 5 inches is unusually tall with respect to the general population, but not with respect to professional basketball players.

Degree to Which a Point Is an Anomaly The assessment of whether an object is an anomaly is reported by some techniques in a binary fashion: An object is either an anomaly or it is not. Frequently, this does not reflect the underlying reality that some objects are more extreme anomalies than others. Hence, it is desirable to have some assessment of the degree to which an object is anomalous. This assessment is known as the **anomaly** or **outlier score**.

Identifying One Anomaly at a Time versus Many Anomalies at Once In some techniques, anomalies are removed one at a time; i.e., the most anomalous instance is identified and removed and then the process repeats. For other techniques, a collection of anomalies is identified together. Techniques that attempt to identify one anomaly at a time are often subject to a problem known as **masking**, where the presence of several anomalies masks the presence of all. On the other hand, techniques that detect multiple outliers at once can experience **swamping**, where normal objects are classified as outliers. In model-based approaches, these effects can happen because the anomalies distort the data model.

Evaluation If class labels are available to identify anomalies and normal data, then the effectiveness of an anomaly detection scheme can be evaluated by using measures of classification performance discussed in Section 5.7. But since the anomalous class is usually much smaller than the normal class, measures such as precision, recall, and false positive rate are more appropriate than accuracy. If class labels are not available, then evaluation is difficult. However, for model-based approaches, the effectiveness of outlier detection can be judged with respect to the improvement in the model once anomalies are eliminated.

Efficiency There are significant differences in the computational cost of various anomaly detection schemes. Classification-based schemes can require significant resources to create the classification model, but are usually inexpensive to apply. Likewise, statistical approaches create a statistical model and can

then categorize an object in constant time. Proximity-based approaches naturally have a time complexity of $O(m^2)$, where m is the number of objects, because the information they require can usually only be obtained by computing the proximity matrix. This time complexity can be reduced in specific cases, such as low-dimensional data, by the use of special data structure and algorithms. The time complexity of other approaches is considered in Exercise 6 on page 681.

Road Map

The next four sections describe several major categories of anomaly detection approaches: statistical, proximity-based, density-based, and cluster-based. One or more specific techniques are considered within each of these categories. In these sections, we will follow common practice and use the term outlier instead of anomaly.

10.2 Statistical Approaches

Statistical approaches are model-based approaches; i.e., a model is created for the data, and objects are evaluated with respect to how well they fit the model. Most statistical approaches to outlier detection are based on building a probability distribution model and considering how likely objects are under that model. This idea is expressed by Definition 10.2.

Definition 10.2 (Probabilistic Definition of an Outlier). An outlier is an object that has a low probability with respect to a probability distribution model of the data.

A probability distribution model is created from the data by estimating the parameters of a user-specified distribution. If the data is assumed to have a Gaussian distribution, then the mean and standard deviation of the underlying distribution can be estimated by computing the mean and standard deviation of the data. The probability of each object under the distribution can then be estimated.

A wide variety of statistical tests based on Definition 10.2 have been devised to detect outliers, or **discordant observations**, as they are often called in the statistical literature. Many of these discordancy tests are highly specialized and assume a level of statistical knowledge beyond the scope of this text. Thus, we illustrate the basic ideas with a few examples and refer the reader to the bibliographic notes for further pointers.

Issues

Among the important issues facing this approach to outlier detection are the following:

Identifying the specific distribution of a data set. While many types of data can be described by a small number of common distributions, such as Gaussian, Poisson, or binomial, data sets with non-standard distributions are relatively common. Of course, if the wrong model is chosen, then an object can be erroneously identified as an outlier. For example, the data may be modeled as coming from a Gaussian distribution, but may actually come from a distribution that has a higher probability (than the Gaussian distribution) of having values far from the mean. Statistical distributions with this type of behavior are common in practice and are known as **heavy-tailed distributions**.

The number of attributes used. Most statistical outlier detection techniques apply to a single attribute, but some techniques have been defined for multivariate data.

Mixtures of distributions. The data can be modeled as a mixture of distributions, and outlier detection schemes can be developed based on such models. Although potentially more powerful, such models are more complicated, both to understand and to use. For example, the distributions need to be identified before objects can be classified as outliers. See the discussion of mixture models and the EM algorithm in Section 9.2.2.

10.2.1 Detecting Outliers in a Univariate Normal Distribution

The Gaussian (normal) distribution is one of the most frequently used distributions in statistics, and we will use it to describe a simple approach to statistical outlier detection. This distribution has two parameters, μ and σ, which are the mean and standard deviation, respectively, and is represented using the notation $N(\mu, \sigma)$. Figure 10.1 shows the density function of $N(0, 1)$.

There is little chance that an object (value) from a $N(0, 1)$ distribution will occur in the tails of the distribution. For instance, there is only a probability of 0.0027 that an object lies beyond the central area between ± 3 standard deviations. More generally, if c is a constant and x is the attribute value of an object, then the probability that $|x| \geq c$ decreases rapidly as c increases. Let $\alpha = prob(|x| \geq c)$. Table 10.1 shows some sample values for c and the

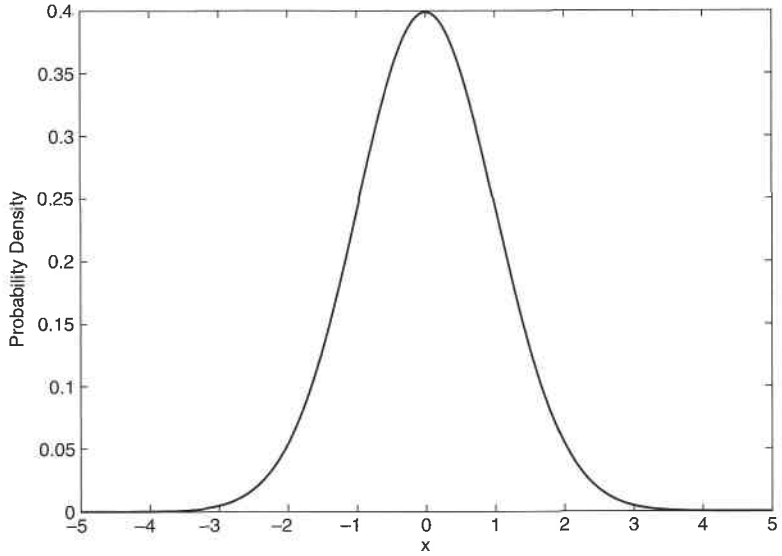

Figure 10.1. Probability density function of a Gaussian distribution with a mean of 0 and a standard deviation of 1.

corresponding values for α when the distribution is $N(0, 1)$. Note that a value that is more than 4 standard deviations from the mean is a one-in-ten-thousand occurrence.

Table 10.1. Sample pairs (c, α), $\alpha = prob(|x| \geq c)$ for a Gaussian distribution with mean 0 and standard deviation 1.

c	α for $N(0, 1)$
1.00	0.3173
1.50	0.1336
2.00	0.0455
2.50	0.0124
3.00	0.0027
3.50	0.0005
4.00	0.0001

Because a value's distance c from the center of the $N(0, 1)$ distribution is directly related to the value's probability, it can be used as the basis of a test for whether an object (value) is an outlier as indicated in Definition 10.3.

Definition 10.3 (Outlier for a Single *N(0,1)* Gaussian Attribute). An object with attribute value x from a Gaussian distribution with mean of 0 and standard deviation 1 is an outlier if

$$|x| \geq c, \tag{10.1}$$

where c is a constant chosen so that $prob(|x|) \geq c = \alpha$.

To use this definition it is necessary to specify a value for α. From the viewpoint that unusual values (objects) indicate a value from a different distribution, α indicates the probability that we mistakenly classify a value from the given distribution as an outlier. From the viewpoint that an outlier is a rare value of a $N(0, 1)$ distribution, α specifies the degree of rareness.

If the distribution of an attribute of interest (for the normal objects) has a Gaussian distribution with mean μ and a standard deviation σ, i.e., a $N(\mu, \sigma)$ distribution, then to use Definition 10.3, we need to transform the attribute x to a new attribute z, which has a $N(0, 1)$ distribution. In particular, the approach is to set $z = (x - \mu)/\sigma$. (z is typically called a z score.) However, μ and σ are typically unknown and are estimated using the sample mean $\overline{x}$ and sample standard deviation s_x. In practice, this works well when the number of observations is large. However, we note that the distribution of z is not actually $N(0, 1)$. A more sophisticated statistical procedure (Grubbs' test) is explored in Exercise 7 on page 681.

10.2.2 Outliers in a Multivariate Normal Distribution

For multivariate Gaussian observations, we would like to take an approach similar to that given for a univariate Gaussian distribution. In particular, we would like to classify points as outliers if they have low probability with respect to the estimated distribution of the data. Furthermore, we would like to be able to judge this with a simple test, for example, the distance of a point from the center of the distribution.

However, because of the correlation between the different variables (attributes), a multivariate normal distribution is not symmetrical with respect to its center. Figure 10.2 shows the probability density of a two-dimensional multivariate Gaussian distribution with mean of (0,0) and a covariance matrix of

$$\mathbf{\Sigma} = \begin{pmatrix} 1.00 & 0.75 \\ 0.75 & 3.00 \end{pmatrix}.$$

If we are to use a simple threshold for whether an object is an outlier, then we will need a distance measure that takes the shape of the data distribution into account. The Mahalanobis distance is such a distance. See Equation 2.14 on page 81. The Mahalanobis distance between a point $\mathbf{x}$ and the mean of the data $\overline{\mathbf{x}}$ is shown in Equation 10.2.

$$mahalanobis(\mathbf{x}, \overline{\mathbf{x}}) = (\mathbf{x} - \overline{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}})^T, \tag{10.2}$$

where $\mathbf{S}$ is the covariance matrix of the data.

It is easy to show that the Mahalanobis distance of a point to the mean of the underlying distribution is directly related to the probability of the point. In particular, the Mahalanobis distance is equal to the log of the probability density of the point plus a constant. See Exercise 9 on page 682.

Example 10.1 (Outliers in a Multivariate Normal Distribution). Figure 10.3 shows the Mahalanobis distance (from the mean of the distribution) for points in a two-dimensional data set. The points A $(-4, 4)$ and B $(5, 5)$ are outliers that were added to the data set, and their Mahalanobis distance is indicated in the figure. The other 2000 points of the data set were randomly generated using the distribution used for Figure 10.2.

Both A and B have large Mahalanobis distances. However, even though A is closer to the center (the large black x at (0,0)) as measured by Euclidean distance, it is farther away than B in terms of the Mahalanobis distance because the Mahalanobis distance takes the shape of the distribution into account. In particular, point B has a Euclidean distance of $5\sqrt{2}$ and a Mahalanobis distance of 24, while the point A has a Euclidean distance of $4\sqrt{2}$ and a Mahalanobis distance of 35. ■

10.2.3 A Mixture Model Approach for Anomaly Detection

This section presents an anomaly detection technique that uses a mixture model approach. In clustering (see Section 9.2.2), the mixture model approach assumes that the data comes from a mixture of probability distributions and that each cluster can be identified with one of these distributions. Similarly, for anomaly detection, the data is modeled as a mixture of two distributions, one for ordinary data and one for outliers.

For both clustering and anomaly detection, the goal is to estimate the parameters of the distributions in order to maximize the overall likelihood (probability) of the data. In clustering, the EM algorithm is used to estimate the parameters of each probability distribution. However, the anomaly detection technique presented here uses a simpler approach. Initially, all the

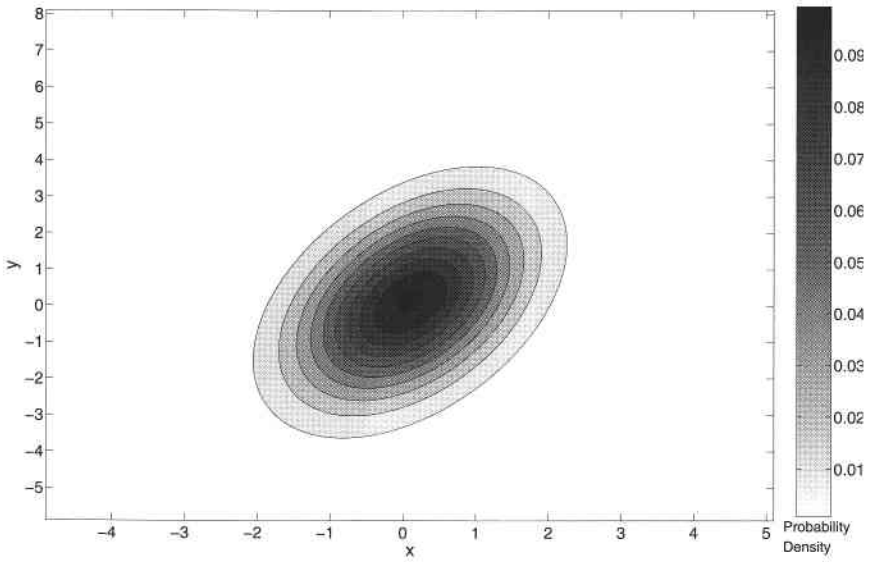

Figure 10.2. Probability density of points for the Gaussian distribution used to generate the points of Figure 10.3.

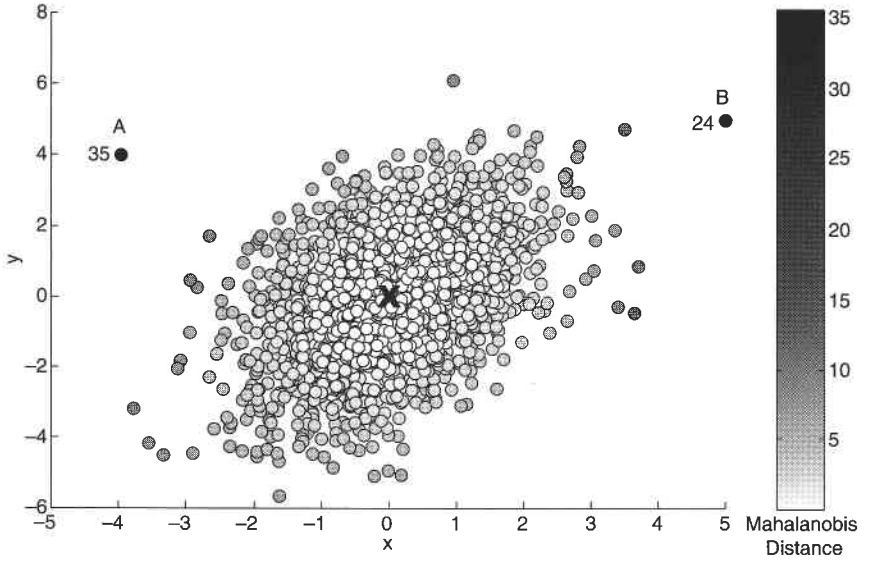

Figure 10.3. Mahalanobis distance of points from the center of a two-dimensional set of 2002 points.

objects are put in a set of normal objects and the set of anomalous objects is empty. An iterative procedure then transfers objects from the ordinary set to the anomalous set as long as the transfer increases the overall likelihood of the data.

Assume that the data set D contains objects from a mixture of two probability distributions: M, the distribution of the majority of (normal) objects, and A, the distribution of anomalous objects. The overall probability distribution of the data can be written as

$$D(\mathbf{x}) = (1 - \lambda)M(\mathbf{x}) + \lambda A(\mathbf{x}). \tag{10.3}$$

where $\mathbf{x}$ is an object and λ is a number between 0 and 1 that gives the expected fraction of outliers. The distribution M is estimated from the data, while the distribution A is often taken to be uniform. Let M_t and A_t be the set of normal and anomalous objects, respectively, at time t. Initially, at time $t = 0$, $M_0 = D$ and A_0 is empty. At an arbitrary time t, the likelihood and log likelihood of the entire data set D are given by the following two equations, respectively:

$$L_t(D) = \prod_{\mathbf{x}_i \in D} P_D(\mathbf{x}_i) = \left((1 - \lambda)^{|M_t|} \prod_{\mathbf{x}_i \in M_t} P_{M_t}(\mathbf{x}_i)\right) \left(\lambda^{|A_t|} \prod_{\mathbf{x}_i \in A_t} P_{A_t}(\mathbf{x}_i)\right) \tag{10.4}$$

$$LL_t(D) = |M_t| \log(1 - \lambda) + \sum_{\mathbf{x}_i \in M_t} \log P_{M_t}(\mathbf{x}_i) + |A_t| \log \lambda + \sum_{\mathbf{x}_i \in A_t} \log P_{A_t}(\mathbf{x}_i) \tag{10.5}$$

where P_D, P_{M_t}, and P_{A_t} are the probability distribution functions for D, M_t and A_t, respectively. This equation can be derived from the general definition of a mixture model given in Equation 9.6 (Section 9.2.2). To do so, it is necessary to make the simplifying assumption that the probability is 0 for both of the following situations: (1) an object in A is a normal object, and (2) an object in M is an outlier. Algorithm 10.1 gives the details.

Because the number of normal objects is large compared to the number of anomalies, the distribution of the normal objects may not change much when an object is moved to the set of anomalies. In that case, the contribution of each normal object to the overall likelihood of the normal objects will remain relatively constant. Furthermore, if a uniform distribution is assumed for anomalies, then each object moved to the set of anomalies contributes a fixed amount to the likelihood of the anomalies. Thus, the overall change in the total likelihood of the data when an object is moved to the set of anomalies is roughly equal to the probability of the object under a uniform distribution

Algorithm 10.1 Likelihood-based outlier detection.

1: Initialization: At time $t = 0$, let M_t contain all the objects, while A_t is empty. Let $LL_t(D) = LL(M_t) + LL(A_t)$ be the log likelihood of all the data.
2: **for** each point $\mathbf{x}$ that belongs to M_t **do**
3: Move $\mathbf{x}$ from M_t to A_t to produce the new data sets A_{t+1} and M_{t+1}.
4: Compute the new log likelihood of D, $LL_{t+1}(D) = LL(M_{t+1}) + LL(A_{t+1})$
5: Compute the difference, $\Delta = LL_t(D) - LL_{t+1}(D)$
6: **if** $\Delta > c$, where c is some threshold **then**
7: $\mathbf{x}$ is classified as an anomaly, i.e., M_{t+1} and A_{t+1} are left unchanged and become the current normal and anomaly sets.
8: **end if**
9: **end for**

(weighted by λ) minus the probability of the object under the distribution of the normal data points (weighted by $1-\lambda$). Consequently, the set of anomalies will tend to consist of those objects that have significantly higher probability under a uniform distribution rather than under the distribution of the normal objects.

In the situation just discussed, the approach described by Algorithm 10.1 is roughly equivalent to classifying objects with a low probability under the distribution of normal objects as outliers. For example, when applied to the points in Figure 10.3, this technique would classify points A and B (and other points far from the mean) as outliers. However, if the distribution of the normal objects changes significantly as anomalies are removed or the distribution of the anomalies can be modeled in a more sophisticated manner, then the results produced by this approach will be different than the results of simply classifying low-probability objects as outliers. Also, this approach can work even when the distribution of objects is multimodal.

10.2.4 Strengths and Weaknesses

Statistical approaches to outlier detection have a firm foundation and build on standard statistical techniques, such as estimating the parameters of a distribution. When there is sufficient knowledge of the data and the type of test that should be applied these tests can be very effective. There are a wide variety of statistical outliers tests for single attributes. Fewer options are available for multivariate data, and these tests can perform poorly for high-dimensional data.

10.3 Proximity-Based Outlier Detection

Although there are several variations on the idea of proximity-based anomaly detection, the basic notion is straightforward. An object is an anomaly if it is distant from most points. This approach is more general and more easily applied than statistical approaches, since it is easier to determine a meaningful proximity measure for a data set than to determine its statistical distribution.

One of the simplest ways to measure whether an object is distant from most points is to use the distance to the k-nearest neighbor. This is captured by Definition 10.4. The lowest value of the outlier score is 0, while the highest value is the maximum possible value of the distance function—usually infinity.

Definition 10.4 (Distance to k-Nearest Neighbor). The outlier score of an object is given by the distance to its k-nearest neighbor.

Figure 10.4 shows a set of two-dimensional points. The shading of each point indicates its outlier score using a value of $k = 5$. Note that outlying point C has been correctly assigned a high outlier score.

The outlier score can be highly sensitive to the value of k. If k is too small, e.g., 1, then a small number of nearby outliers can cause a low outlier score. For example, Figure 10.5 shows a set of two-dimensional points in which another point is close to C. The shading reflects the outlier score using a value of $k = 1$. Note that both C and its neighbor have a low outlier score. If k is too large, then it is possible for all objects in a cluster that has fewer objects than k to become outliers. For example, Figure 10.6 shows a two-dimensional data set that has a natural cluster of size 5 in addition to a larger cluster of size 30. For $k = 5$, the outlier score of all points in the smaller cluster is very high. To make the scheme more robust to the choice of k, Definition 10.4 can be modified to use the average of the distances to the first k-nearest neighbors.

10.3.1 Strengths and Weaknesses

The distance-based outlier detection scheme described above, and other related schemes, are simple. However, proximity-based approaches typically take $O(m^2)$ time. For large data sets this can be too expensive, although specialized algorithms can be used to improve performance in the case of low-dimensional data. Also, the approach is sensitive to the choice of parameters. Furthermore, it cannot handle data sets with regions of widely differing densities because it uses global thresholds that cannot take into account such density variations.

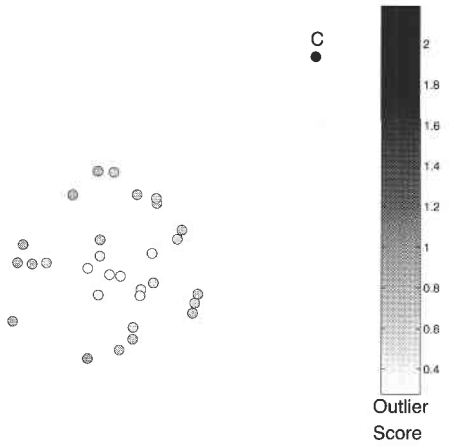

Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

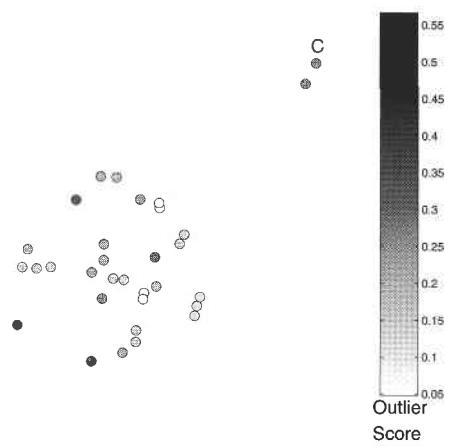

Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

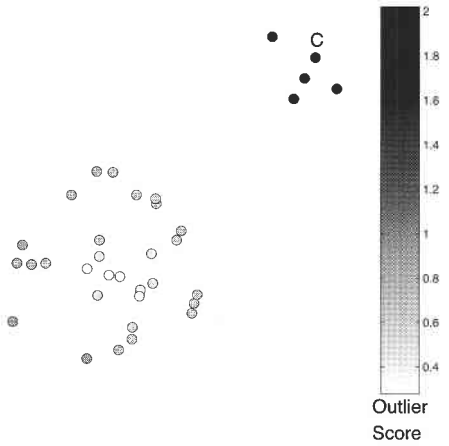

Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

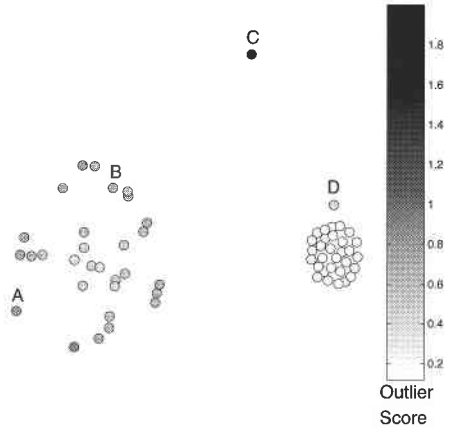

Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

To illustrate this, consider the set of two-dimensional points in Figure 10.7. This figure has one rather loose cluster of points, another dense cluster of points, and two points, C and D, that are quite far from these two clusters. Assigning the outlier score to points according to Definition 10.4 for $k = 5$, correctly identifies point C to be an outlier, but shows a low outlier score for

point D. In fact, the outlier score for D is much lower than many points that are part of the loose cluster.

10.4 Density-Based Outlier Detection

From a density-based viewpoint, outliers are objects that are in regions of low density.

Definition 10.5 (Density-Based Outlier). The outlier score of an object is the inverse of the density around an object.

Density-based outlier detection is closely related to proximity-based outlier detection since density is usually defined in terms of proximity. One common approach is to define density as the reciprocal of the average distance to the k nearest neighbors. If this distance is small, the density is high, and vice versa. This is captured by Definition 10.6.

Definition 10.6 (Inverse Distance).

$$density(\mathbf{x}, k) = \left(\frac{\sum_{\mathbf{y} \in N(\mathbf{x},k)} distance(\mathbf{x}, \mathbf{y})}{|N(\mathbf{x}, k)|} \right)^{-1} \tag{10.6}$$

where $N(\mathbf{x}, k)$ is the set containing the k-nearest neighbors of $\mathbf{x}$, $|N(\mathbf{x}, k)|$ is the size of that set, and $\mathbf{y}$ is a nearest neighbor.

Another definition of density is the one used by the DBSCAN clustering algorithm. See Section 8.4.

Definition 10.7 (Count of Points within a Given Radius). The density around an object is equal to the number of objects that are within a specified distance d of the object.

The parameter d needs to be chosen carefully. If d is too small, then many normal points may have low density and thus a high outlier score. If d is chosen to be large, then many outliers may have densities (and outlier scores) that are similar to normal points.

Detecting outliers using any of the definitions of density has similar characteristics and limitations to those of the proximity-based outlier schemes discussed in Section 10.3. In particular, they cannot identify outliers correctly when the data contains regions of differing densities. (See Figure 10.7.) To correctly identify outliers in such data sets, we need a notion of density that is relative to the neighborhood of the object. For example, point D in Figure

10.7 has a higher absolute density, according to Definitions 10.6 and 10.7, than point A, but its density is lower relative to its nearest neighbors.

There are many ways to define the relative density of an object. One method that is used by the SNN density-based clustering algorithm is discussed in Section 9.4.8. Another method is to compute the relative density as the ratio of the density of a point $\mathbf{x}$ and the average density of its nearest neighbors $\mathbf{y}$ as follows:

$$average\ relative\ density(\mathbf{x}, k) = \frac{density(\mathbf{x}, k)}{\sum_{\mathbf{y} \in N(\mathbf{x},k)} density(y, k)/|N(\mathbf{x}, k)|}. \quad (10.7)$$

10.4.1 Detection of Outliers Using Relative Density

In this section, we describe a technique that is based on the notion of relative density. This technique, which is a simplified version of the Local Outlier Factor (LOF) technique (see bibliographic notes), is described in Algorithm 10.2. The details of the algorithm are examined in more detail below, but in summary, it works as follows. We calculate the outlier score for each object for a specified number of neighbors (k) by first computing the density of an object $density(\mathbf{x}, k)$ based on its nearest neighbors. The average density of the neighbors of a point is then calculated and used to compute the average relative density of the point as indicated in Equation 10.7. This quantity provides an indication of whether $\mathbf{x}$ is in a denser or sparser region of the neighborhood than its neighbors and is taken as the outlier score of $\mathbf{x}$.

Algorithm 10.2 Relative density outlier score algorithm.

1: {k is the number of nearest neighbors}
2: **for all** objects $\mathbf{x}$ **do**
3: Determine $N(\mathbf{x}, k)$, the k-nearest neighbors of $\mathbf{x}$.
4: Determine $density(\mathbf{x}, k)$, the density of $\mathbf{x}$ using its nearest neighbors, i.e., the objects in $N(\mathbf{x}, k)$.
5: **end for**
6: **for all** objects $\mathbf{x}$ **do**
7: Set the $outlier\ score(\mathbf{x}, k) = average\ relative\ density(\mathbf{x}, k)$ from Equation 10.7.
8: **end for**

Example 10.2 (Relative Density Outlier Detection). We illustrate the performance of the relative density outlier detection method by using the example data set shown in Figure 10.7. Here, $k = 10$. The outlier scores for

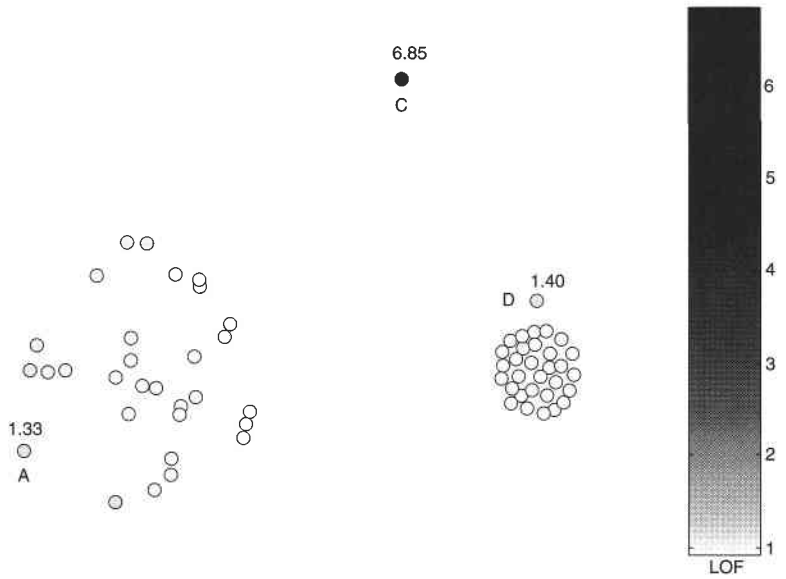

Figure 10.8. Relative density (LOF) outlier scores for two-dimensional points of Figure 10.7.

these points are shown in Figure 10.8. The shading of each point is determined by its score; i.e., points with a high score are darker. We have labeled points A, C, and D, which have the largest outlier scores, with these values. Respectively, these points are the most extreme outlier, the most extreme point with respect to the compact set of points, and the most extreme point in the loose set of points. ■

10.4.2 Strengths and Weaknesses

Outlier detection based on relative density gives a quantitative measure of the degree to which an object is an outlier and can work well even if data has regions of differing density. Like distance-based approaches, these approaches naturally have $O(m^2)$ time complexity (where m is the number of objects), although this can be reduced to $O(m \log m)$ for low-dimensional data by using special data structures. Parameter selection can also be difficult, although the standard LOF algorithm addresses this by looking at a variety of values for k and then taking the maximum outlier scores. However, the upper and lower bounds of these values still need to be chosen.

10.5 Clustering-Based Techniques

Cluster analysis finds groups of strongly related objects, while anomaly detection finds objects that are not strongly related to other objects. It should not be surprising, then, that clustering can be used for outlier detection. In this section, we will discuss several such techniques.

One approach to using clustering for outlier detection is to discard small clusters that are far from other clusters. This approach can be used with any clustering technique, but requires thresholds for the minimum cluster size and the distance between a small cluster and other clusters. Often, the process is simplified by discarding all clusters smaller than a minimum size. This scheme is highly sensitive to the number of clusters chosen. Also, it is hard to attach an outlier score to objects using this scheme. Note that considering groups of objects as outliers extends the notion of outliers from individual objects to groups of objects, but does not change anything essential.

A more systematic approach is to first cluster all objects and then assess the degree to which an object belongs to any cluster. For prototype-based clustering, the distance of an object to its cluster center can be used to measure the degree to which the object belongs to a cluster. More generally, for clustering techniques that are based on an objective function, we can use the objective function to assess how well an object belongs to any cluster. In particular, if the elimination of an object results in a substantial improvement in the objective, then we would classify the object as an outlier. To illustrate, for K-means, eliminating an object that is far from the center of its associated cluster can substantially improve the sum of the squared error (SSE) of the cluster. In summary, clustering creates a model of the data and anomalies distort that model. This idea is captured in Definition 10.8.

Definition 10.8 (Clustering-Based Outlier). An object is a cluster-based outlier if the object does not strongly belong to any cluster.

When used with clustering schemes that have an objective function, this definition is a special case of the definition of a model-based anomaly. Although Definition 10.8 is more natural for prototype-based schemes or schemes that have an objective function, it can also encompass density- and connectivity-based clustering approaches to outlier detection. In particular, for density-based clustering, an object does not strongly belong to any cluster if its density is too low, while for connectivity-based clustering, an object does not strongly belong to any cluster if it is not strongly connected.

Below, we will discuss issues that need to be addressed by any technique for clustering-based outlier detection. Our discussion will focus on prototype-based clustering techniques, such as K-means.

10.5.1 Assessing the Extent to Which an Object Belongs to a Cluster

For prototype-based clusters, there are several ways to assess the extent to which an object belongs to a cluster. One method is to measure the distance of the object to the cluster prototype and take this as the outlier score of the object. However, if the clusters are of differing densities, then we can construct an outlier score that measures the relative distance of an object from the cluster prototype with respect to the distances of the other objects in the cluster. Another possibility, provided that the clusters can be accurately modeled in terms of Gaussian distributions, is to use the Mahalanobis distance.

For clustering techniques that have an objective function, we can assign an outlier score to an object that reflects the improvement in the objective function when that object is eliminated. However, assessing the degree to which a point is an outlier based on the objective function can be computationally intensive. For that reason, the distance-based approaches of the previous paragraph are often preferred.

Example 10.3 (Clustering-Based Example). This example is based on the set of points shown in Figure 10.7. Prototype-based clustering uses the K-means algorithm, and the outlier score of a point is computed in two ways: (1) by the point's distance from its closest centroid, and (2) by the point's relative distance from its closest centroid, where the relative distance is the ratio of the point's distance from the centroid to the median distance of all points in the cluster from the centroid. The latter approach is used to adjust for the large difference in density between the compact and loose clusters.

The resulting outlier scores are shown in Figures 10.9 and 10.10. As before, the outlier score, measured in this case by the distance or relative distance, is indicated by the shading. We use two clusters in each case. The approach based on raw distance has problems with the differing densities of the clusters, e.g., D is not considered an outlier. For the approach based on relative distances, the points that have previously been identified as outliers using LOF (A, C, and D) also show up as outliers here. ■

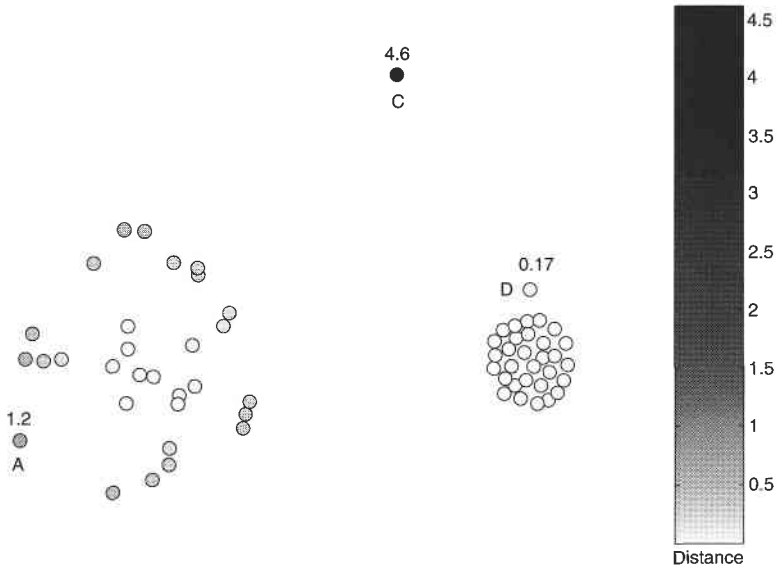

Figure 10.9. Distance of points from closest centroid.

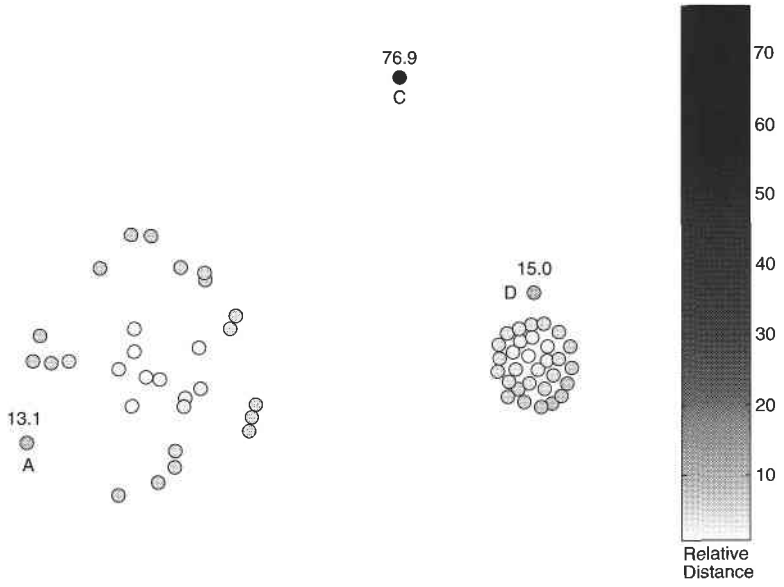

Figure 10.10. Relative distance of points from closest centroid.

10.5.2 Impact of Outliers on the Initial Clustering

If outliers are detected by clustering, there is a question of whether the results are valid since outliers affect the clustering. To address this issue, the following approach can be used: objects are clustered, outliers are removed, and then the objects are clustered again. While there is no guarantee that this approach will yield optimal results, it is easy to use. A more sophisticated approach is to have a special group for objects that do not currently fit well in any cluster. This group represents potential outliers. As the clustering process proceeds, clusters change. Objects that no longer belong strongly to any cluster are added to the set of potential outliers, while objects currently in the set are tested to see if they now strongly belong to a cluster and can be removed from the set of potential outliers. The objects remaining in the set at the end of the clustering are classified as outliers. Again, there is no guarantee of an optimal solution or even that this approach will work better than the simpler one described previously. For example, a cluster of noise points may look like a real cluster with no outliers. This problem is particularly serious if the outlier score is computed using the relative distance.

10.5.3 The Number of Clusters to Use

Clustering techniques such as K-means do not automatically determine the number of clusters. This is a problem when using clustering in outlier detection, since whether an object is considered an outlier or not may depend on the number of clusters. For instance, a group of 10 objects may be relatively close to one another, but may be included as part of a larger cluster if only a few large clusters are found. In that case, each of the 10 points could be regarded as an outlier, even though they would have formed a cluster if a large enough number of clusters had been specified.

As with some of the other issues, there is no simple answer to this problem. One strategy is to repeat the analysis for different numbers of clusters. Another approach is to find a large number of small clusters. The idea here is that (1) smaller clusters tend to be more cohesive and (2) if an object is an outlier even when there are a large number of small clusters, then it is likely a true outlier. The downside is that groups of outliers may form small clusters and thus escape detection.

10.5.4 Strengths and Weaknesses

Some clustering techniques, such as K-means, have linear or near-linear time and space complexity and thus, an outlier detection technique based on such

algorithms can be highly efficient. Also, the definition of a cluster is often complementary to that of an outlier and thus, it is usually possible to find both clusters and outliers at the same time. On the negative side, the set of outliers produced and their scores can be heavily dependent upon the number of clusters used as well as the presence of outliers in the data. For example, clusters produced by prototype-based algorithms can be distorted by the presence of outliers. The quality of outliers produced by a clustering algorithm is heavily impacted by the quality of clusters produced by the algorithm. As discussed in Chapters 8 and 9, each clustering algorithm is suitable only for a certain type of data; hence the clustering algorithm needs to be chosen carefully.

10.6 Bibliographic Notes

Anomaly detection has a long history, particularly in statistics, where it is known as outlier detection. Relevant books on the topic are those of Barnett and Lewis [464], Hawkins [483], and Rousseeuw and Leroy [513]. The article by Beckman and Cook [466] provides a general overview of how statisticians look at the subject of outlier detection and provides a history of the subject dating back to comments by Bernoulli in 1777. Also see the related articles [467, 484]. Another general article on outlier detection is the one by Barnett [463]. Articles on finding outliers in multivariate data include those by Davies and Gather [474], Gnanadesikan and Kettenring [480], Rocke and Woodruff [511], Rousseeuw and van Zomerenand [515], and Scott [516]. Rosner [512] provides a discussion of finding multiple outliers at the same time.

An extensive survey of outlier detection methods is provided by Hodge and Austin [486]. Markou and Singh [506, 507] give a two-part review of techniques for novelty detection that covers statistical and neural network techniques, respectively. Grubbs' procedure for detecting outliers was originally described in [481]. The mixture model outlier approach discussed in Section 10.2.3 is from Eskin [476]. The notion of a distance-based outlier and the fact that this definition can include many statistical definitions of an outlier was described by Knorr et al. [496–498]. The LOF technique (Breunig et al. [468, 469]) grew out of DBSCAN. Ramaswamy et al. [510] propose a distance-based outlier detection procedure that gives each object an outlier score based on the distance of its k-nearest neighbor. Efficiency is achieved by partitioning the data using the first phase of BIRCH (Section 9.5.2). Chaudhary et al. [470] use k-d trees to improve the efficiency of outlier detection, while Bay and Schwabacher [465] use randomization and pruning to improve performance. Aggarwal and Yu [462] use projection to address outlier detection for high-

dimensional data, while Shyu et al. [518] use an approach based on principal components. A theoretical discussion of outlier removal in high-dimensional space can be found in the paper by Dunagan and Vempala [475]. The use of information measures in anomaly detection is described by Lee and Xiang [504], while an approach based on the χ^2 measure is given by Ye and Chen [520].

Many different types of classification techniques can be used for anomaly detection. A discussion of approaches in the area of neural networks can be found in papers by Hawkins et al. [485], Ghosh and Schwartzbard [479], and Sykacek [519]. Recent work on rare class detection includes the work of Joshi et al. [490–494]. The rare class problem is also sometimes referred to as the imbalanced data set problem. Of relevance are an AAAI workshop (Japkowicz [488]), an ICML workshop (Chawla et al. [471]), and a special issue of SIGKDD Explorations (Chawla et al. [472]).

Clustering and anomaly detection have a long relationship. In Chapters 8 and 9, we considered techniques, such as BIRCH, CURE, DENCLUE, DBSCAN, and SNN density-based clustering, which specifically include techniques for handling anomalies. Statistical approaches that discuss this relationship are described in papers by Scott [516] and Hardin and Rocke [482].

In this chapter, we have focused on basic anomaly detection schemes. We have not considered schemes that take into account the spatial or temporal nature of the data. Shekhar et al. [517] provide a detailed discussion of the problem of spatial outliers and present a unified approach to spatial outlier detection. The issue of outliers in time series was first considered in a statistically rigorous way by Fox [478]. Muirhead [508] provides a discussion of different types of outliers in time series. Abraham and Chuang [461] propose a Bayesian approach to outliers in time series, while Chen and Liu [473] consider different types of outliers in time series and propose a technique to detect them and obtain good estimates of time series parameters. Work on finding deviant or surprising patterns in time series databases has been performed by Jagadish et al. [487] and Keogh et al. [495]. Outlier detection based on geometric ideas, such as the depth of convex hulls, has been explored in papers by Johnson et al. [489], Liu et al. [505], and Rousseeuw et al. [514].

An important application area for anomaly detection is intrusion detection. Surveys of the applications of data mining to intrusion detection are given by Lee and Stolfo [502] and Lazarevic et al. [501]. In a different paper, Lazarevic et al. [500] provide a comparison of anomaly detection routines specific to network intrusion. A framework for using data mining techniques for intrusion detection is provided by Lee et al. [503]. Clustering-based approaches in the

area of intrusion detection include work by Eskin et al. [477], Lane and Brodley [499], and Portnoy et al. [509].

Bibliography

[461] B. Abraham and A. Chuang. Outlier Detection and Time Series Modeling. *Technometrics*, 31(2):241–248, May 1989.

[462] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *Proc. of 2001 ACM-SIGMOD Intl. Conf. on Management of Data*, pages 37–46. ACM Press, 2001.

[463] V. Barnett. The Study of Outliers: Purpose and Model. *Applied Statistics*, 27(3): 242–250, 1978.

[464] V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, 3rd edition, April 1994.

[465] S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proc. of the 9th Intl. Conf. on Knowledge Discovery and Data Mining*, pages 29–38. ACM Press, 2003.

[466] R. J. Beckman and R. D. Cook. 'Outlier.....s'. *Technometrics*, 25(2):119–149, May 1983.

[467] R. J. Beckman and R. D. Cook. ['Outlier.....s']: Response. *Technometrics*, 25(2): 161–163, May 1983.

[468] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. OPTICS-OF: Identifying Local Outliers. In *Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery*, pages 262–270. Springer-Verlag, 1999.

[469] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proc. of 2000 ACM-SIGMOD Intl. Conf. on Management of Data*, pages 93–104. ACM Press, 2000.

[470] A. Chaudhary, A. S. Szalay, and A. W. Moore. Very fast outlier detection in large multidimensional data sets. In *Proc. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)*, 2002.

[471] N. V. Chawla, N. Japkowicz, and A. Kolcz, editors. *Workshop on Learning from Imbalanced Data Sets II, 20th Intl. Conf. on Machine Learning*, 2000. AAAI Press.

[472] N. V. Chawla, N. Japkowicz, and A. Kolcz, editors. *SIGKDD Explorations Newsletter, Special issue on learning from imbalanced datasets*, volume 6(1), June 2004. ACM Press.

[473] 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):284–297, March 1993.

[474] L. Davies and U. Gather. The Identification of Multiple Outliers. *Journal of the American Statistical Association*, 88(423):782–792, September 1993.

[475] J. Dunagan and S. Vempala. Optimal outlier removal in high-dimensional spaces. *Journal of Computer and System Sciences, Special Issue on STOC 2001*, 68(2):335–373, March 2004.

[476] E. Eskin. Anomaly Detection over Noisy Data using Learned Probability Distributions. In *Proc. of the 17th Intl. Conf. on Machine Learning*, pages 255–262, 2000.

[477] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. J. Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of Data Mining in Computer Security*, pages 78–100. Kluwer Academics, 2002.

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

[479] A. Ghosh and A. Schwartzbard. A Study in Using Neural Networks for Anomaly and Misuse Detection. In *8th USENIX Security Symposium*, August 1999.

[480] R. Gnanadesikan and J. R. Kettenring. Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. *Biometrics*, 28(1):81–124, March 1972.

[481] F. Grubbs. Procedures for Testing Outlying Observations. *Annal of Mathematical Statistics*, 21(1):27–58, March 1950.

[482] J. Hardin and D. M. Rocke. Outlier Detection in the Multiple Cluster Setting using the Minimum Covariance Determinant Estimator. *Computational Statistics and Data Analysis*, 44:625–638, 2004.

[483] D. M. Hawkins. *Identification of Outliers*. Monographs on Applied Probability and Statistics. Chapman & Hall, May 1980.

[484] D. M. Hawkins. '[Outlier.....s]': Discussion. *Technometrics*, 25(2):155–156, May 1983.

[485] S. Hawkins, H. He, G. J. Williams, and R. A. Baxter. Outlier Detection Using Replicator Neural Networks. In *DaWaK 2000: Proc. of the 4th Intnl. Conf. on Data Warehousing and Knowledge Discovery*, pages 170–180. Springer-Verlag, 2002.

[486] V. J. Hodge and J. Austin. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22:85–126, 2004.

[487] H. V. Jagadish, N. Koudas, and S. Muthukrishnan. Mining Deviants in a Time Series Database. In *Proc. of the 25th VLDB Conf.*, pages 102–113, 1999.

[488] N. Japkowicz, editor. *Workshop on Learning from Imbalanced Data Sets I, Seventeenth National Conference on Artificial Intelligence, Published as Technical Report WS-00-05*, 2000. AAAI Press.

[489] T. Johnson, I. Kwok, and R. T. Ng. Fast Computation of 2-Dimensional Depth Contours. In *KDD98*, pages 224–228, 1998.

[490] M. V. Joshi. On Evaluating Performance of Classifiers for Rare Classes. In *Proc. of the 2002 IEEE Intl. Conf. on Data Mining*, pages 641–644, 2002.

[491] M. V. Joshi, R. C. Agarwal, and V. Kumar. Mining needle in a haystack: Classifying rare classes via two-phase rule induction. In *Proc. of 2001 ACM-SIGMOD Intl. Conf. on Management of Data*, pages 91–102. ACM Press, 2001.

[492] M. V. Joshi, R. C. Agarwal, and V. Kumar. Predicting rare classes: can boosting make any weak learner strong? In *Proc. of 2002 ACM-SIGMOD Intl. Conf. on Management of Data*, pages 297–306. ACM Press, 2002.

[493] M. V. Joshi, R. C. Agarwal, and V. Kumar. Predicting Rare Classes: Comparing Two-Phase Rule Induction to Cost-Sensitive Boosting. In *Proc. of the 6th European Conf. of Principles and Practice of Knowledge Discovery in Databases*, pages 237–249. Springer-Verlag, 2002.

[494] M. V. Joshi, V. Kumar, and R. C. Agarwal. Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements. In *Proc. of the 2001 IEEE Intl. Conf. on Data Mining*, pages 257–264, 2001.

[495] E. Keogh, S. Lonardi, and B. Chiu. Finding Surprising Patterns in a Time Series Database in Linear Time and Space. In *Proc. of the 8th Intl. Conf. on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 2002.

[496] E. M. Knorr and R. T. Ng. A Unified Notion of Outliers: Properties and Computation. In *Proc. of the 3rd Intl. Conf. on Knowledge Discovery and Data Mining*, pages 219–222, 1997.

[497] E. M. Knorr and R. T. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. In *Proc. of the 24th VLDB Conf.*, pages 392–403, 24–27 1998.

[498] E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: algorithms and applications. *The VLDB Journal*, 8(3-4):237–253, 2000.

[499] T. Lane and C. E. Brodley. An Application of Machine Learning to Anomaly Detection. In *Proc. 20th NIST-NCSC National Information Systems Security Conf.*, pages 366–380, 1997.

[500] A. Lazarevic, L. Ertöz, V. Kumar, A. Ozgur, and J. Srivastava. A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection. In *Proc. of the 2003 SIAM Intl. Conf. on Data Mining*, 2003.

[501] A. Lazarevic, V. Kumar, and J. Srivastava. Intrusion Detection: A Survey. In *Managing Cyber Threats: Issues, Approaches and Challenges*, pages 19–80. Kluwer Academic Publisher, 2005.

[502] W. Lee and S. J. Stolfo. Data Mining Approaches for Intrusion Detection. In *7th USENIX Security Symposium*, pages 26–29, January 1998.

[503] W. Lee, S. J. Stolfo, and K. W. Mok. A Data Mining Framework for Building Intrusion Detection Models. In *IEEE Symposium on Security and Privacy*, pages 120–132, 1999.

[504] W. Lee and D. Xiang. Information-theoretic measures for anomaly detection. In *Proc. of the 2001 IEEE Symposium on Security and Privacy*, pages 130–143, May 2001.

[505] R. Y. Liu, J. M. Parelius, and K. Singh. Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Annals of Statistics*, 27(3):783–858, 1999.

[506] M. Markou and S. Singh. Novelty detection: A review–part 1: Statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[507] M. Markou and S. Singh. Novelty detection: A review–part 2: Neural network based approaches. *Signal Processing*, 83(12):2499–2521, 2003.

[508] C. R. Muirhead. Distinguishing Outlier Types in Time Series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(1):39–47, 1986.

[509] L. Portnoy, E. Eskin, and S. J. Stolfo. Intrusion detection with unlabeled data using clustering. In *In ACM Workshop on Data Mining Applied to Security*, 2001.

[510] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proc. of 2000 ACM-SIGMOD Intl. Conf. on Management of Data*, pages 427–438. ACM Press, 2000.

[511] D. M. Rocke and D. L. Woodruff. Identification of Outliers in Multivariate Data. *Journal of the American Statistical Association*, 91(435):1047–1061, September 1996.

[512] B. Rosner. On the Detection of Many Outliers. *Technometrics*, 17(3):221–227, 1975.

[513] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* Wiley Series in Probability and Statistics. John Wiley & Sons, September 2003.

[514] P. J. Rousseeuw, I. Ruts, and J. W. Tukey. The Bagplot: A Bivariate Boxplot. *The American Statistician*, 53(4):382–387, November 1999.

[515] P. J. Rousseeuw and B. C. van Zomeren. Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association*, 85(411):633–639, September 1990.

[516] D. W. Scott. Partial Mixture Estimation and Outlier Detection in Data and Regression. In M. Hubert, G. Pison, A. Struyf, and S. V. Aelst, editors, *Theory and Applications of Recent Robust Methods*, Statistics for Industry and Technology. Birkhauser, 2003.

[517] S. Shekhar, C.-T. Lu, and P. Zhang. A Unified Approach to Detecting Spatial Outliers. *GeoInformatica*, 7(2):139–166, June 2003.

[518] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn, and L. Chang. A Novel Anomaly Detection Scheme Based on Principal Component Classifier. In *Proc. of the 2003 IEEE Intl. Conf. on Data Mining*, pages 353–365, 2003.

[519] P. Sykacek. Equivalent error bars for neural network classifiers trained by bayesian inference. In *Proc. of the European Symposium on Artificial Neural Networks*, pages 121–126, 1997.

[520] N. Ye and Q. Chen. Chi-square Statistical Profiling for Anomaly Detection. In *Proc. of the 2000 IEEE Workshop on Information Assurance and Security*, pages 187–193, June 2000.

10.7 Exercises

1. Compare and contrast the different techniques for anomaly detection that were presented in Section 10.1.2. In particular, try to identify circumstances in which the definitions of anomalies used in the different techniques might be equivalent or situations in which one might make sense, but another would not. Be sure to consider different types of data.

2. Consider the following definition of an anomaly: An anomaly is an object that is unusually influential in the creation of a data model.

   (a) Compare this definition to that of the standard model-based definition of an anomaly.

   (b) For what sizes of data sets (small, medium, or large) is this definition appropriate?

3. In one approach to anomaly detection, objects are represented as points in a multidimensional space, and the points are grouped into successive shells, where each shell represents a layer around a grouping of points, such as a convex hull. An object is an anomaly if it lies in one of the outer shells.

   (a) To which of the definitions of an anomaly in Section 10.1.2 is this definition most closely related?

   (b) Name two problems with this definition of an anomaly.

4. Association analysis can be used to find anomalies as follows. Find strong association patterns, which involve some minimum number of objects. Anomalies are those objects that do not belong to any such patterns. To make this more concrete, we note that the hyperclique association pattern discussed in Section 6.8 is particularly suitable for such an approach. Specifically, given a user-selected h-confidence level, maximal hyperclique patterns of objects are found. All objects that do not appear in a maximal hyperclique pattern of at least size three are classified as outliers.