1、Input Data

Input is generally a collection of data instances (also referred as *object, record, point, vector, pattern, event, case, sample, observation, entity*) . Each data instance can be described using a set of attributes (also referred to as *variable, characteristic, feature, field, dimension*). The attributes can be of different types such as *binary*, *categorical* or *continuous*. Each data instance might consist of only one attribute (*univariate*) or multiple attributes (*multivariate*).

2、Type of Anomaly

一共有三种，但是第三种 collective anomaly 个人觉得与我们的方向无关，就不列出来了

（1） *Point Anomalies.* If an individual data instance can be considered as anomalous with respect to the rest of data, then the instance is termed as a point anomaly.
in Figure 1, points o_1 and o_2 as well as points in region O_3 lie outside the boundary of the normal regions, and hence are point anomalies since they are different from normal data points.

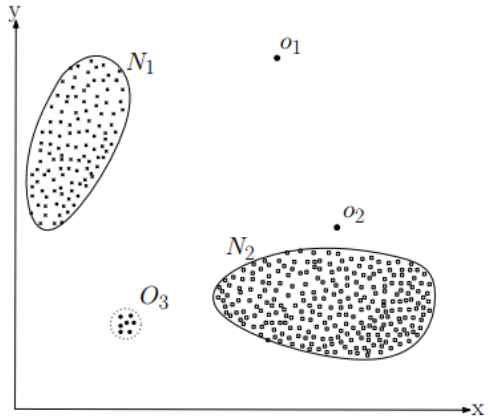

Fig. 1. A simple example of anomalies in a 2-dimensional data set.

（2） *Contextual Anomalies.* If a data instance is anomalous in a specific context (but not otherwise), then it is termed as a contextual anomaly . Each data instance is defined using following two sets of attributes:

(a) *Contextual attributes*. The contextual attributes are used to determine the context (or neighborhood) for that instance. For example, in spatial data sets, the longitude and latitude of a location are the contextual attributes. In time series data, time is a contextual attribute which determines the position of an instance on the entire sequence.

(b) *Behavioral attributes*. The behavioral attributes define the non-contextual characteristics of an instance. For example, in a spatial data set describing the average rainfall of the entire world, the amount of rainfall at any location is a behavioral attribute.

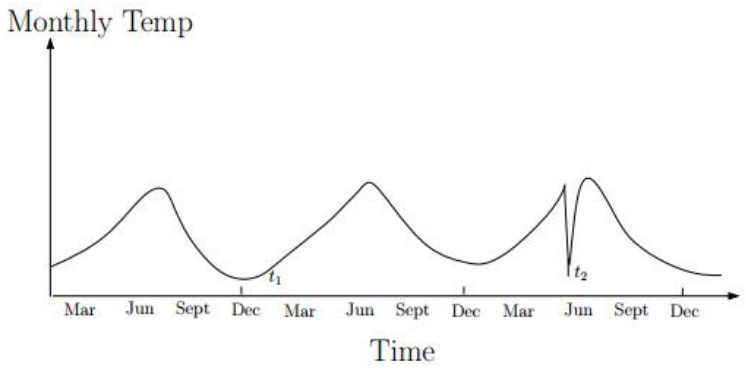

Fig. 3. Contextual anomaly t_2 in a temperature time series. Note that the temperature at time t_1 is same as that at time t_2 but occurs in a different context and hence is not considered as an anomaly.

3、Data Labels

The labels associated with a data instance denote if that instance is *normal* or *anomalous.* It should be noted that obtaining labeled data which is accurate as well as representative of all types of behaviors, is often prohibitively expensive. Labeling is often done manually by a human expert and hence requires substantial effort to obtain the labeled training data set.

根据是否有标签分为有监督，半监督，无监督

4、Output of Anomaly Detection

Typically, the outputs produced by anomaly detection techniques are one of the following two types:

（a） *Scores.* Scoring techniques assign an anomaly score to each instance in the test data depending on the degree to which that instance is considered an anomaly. Thus the output of such techniques is a ranked list of anomalies. An analyst may choose to either analyze top few anomalies or use a cut-off threshold to select the anomalies.

（b） *Labels.* Techniques in this category assign a label (*normal* or *anomalous*) to each test instance.

Scoring based anomaly detection techniques allow the analyst to use a domain specific threshold to select the most relevant anomalies. Techniques that provide binary labels to the test instances do not directly allow the analysts to make such a choice, though this can be controlled indirectly through parameter choices within each technique.

5、Techniques

方法太多，这里只简单提一下前三大类，个人觉得比较常用。

(1) Classification based anomaly detection techniques

(a) neural network based

A basic multi-class anomaly detection technique using neural networks

operates in two steps. First, a neural network is trained on the normal training data to learn the different normal classes. Second, each test instance is provided as an input to the neural network. If the network accepts the test input, it is normal and if the network rejects a test input, it is an anomaly

(b) SVM

(c) Bayesian networks based

(2) Nearest neighbor based anomaly detection techniques

Nearest neighbor based anomaly detection techniques can be broadly grouped into two categories:

(a) Techniques that use the distance of a data instance to its *kth* nearest neighbor as the anomaly score.

(b) Techniques that compute the relative density of each data instance to compute its anomaly score.

(3) Clustering based anomaly detection techniques

方法比较多，这里就不在列上来，一来有些方法比较老，二来这里的方法包揽的所有的异常检测领域，有的仅针对某种情况，没必要全部了解。

Anomaly Detection: A Survey

http://www.dtc.umn.edu/publications/reports/2008_16.pdf