异常值检测算法（二）

原创 2016-06-25 张戎 数学人生

前面一篇文章《异常点检测算法（一）》简要的介绍了如何使用概率统计的方法来计算异常点，本文将会介绍一种基于矩阵分解的异常点检测方法。在介绍这种方法之前，先回顾一下主成分分析（Principle Component Analysis）这一基本的降维方法。

（一）主成分分析（Principle Component Analysis）

对高维数据集合的简化有各种各样的原因，例如：

（1）使得数据集合更容易使用；

（2）降低很多算法的计算开销；

（3）去除噪声；

（4）更加容易的描述结果。

在主成分分析（PCA）这种降维方法中，数据从原来的坐标系转换到新的坐标系，新坐标系的选择是由数据集本身所决定的。第一个新坐标轴的方向选择的是原始数据集中方差最大的方向，第二个新坐标轴的选择是和第一个坐标轴正交并且具有最大方差的方向。该过程一直重复，重复的次数就是原始数据中特征的数目。如此操作下去，将会发现，大部分方差都包含在最前面的几个新坐标轴之中。因此，我们可以忽略余下的坐标轴，也就是对数据进行了降维的处理。

为了提取到第一个主成分（数据差异性最大）的方向，进而提取到第二个主成分（数据差异性次大）的方向，并且该方向需要和第一个主成分方向正交，那么我们就需要对数据集的协方差矩阵进行特征值的分析，从而获得这些主成分的方向。一旦我们计算出了协方差矩阵的特征向量，我们就可以保留最大的 N 个值。正是这 N 个值反映了 N 个最重要特征的真实信息，可以把原始数据集合映射到 N 维的低维空间。

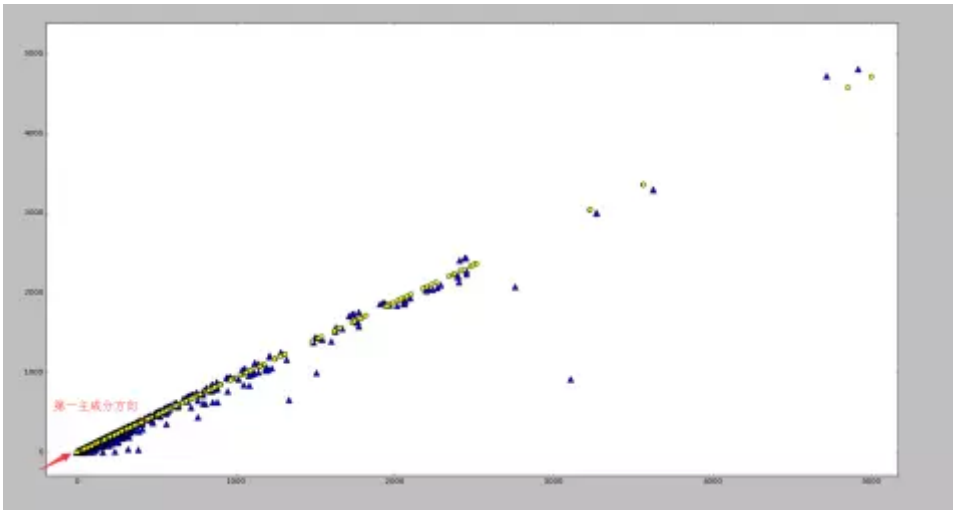
提取 N 个主成分的伪代码如下：

```
去除平均值

计算协方差矩阵

计算协方差矩阵的特征值和特征向量

将特征值从大到小排序
```

保留最大的N个特征值以及它们的特征向量

将数据映射到上述 N 个特征向量构造的新空间中

通过 Python 的 numpy 库和 matplotlib 库可以计算出某个二维数据集合的第一主成分如下：原始数据集使用蓝色的三角形表示，第一主成分使用黄色的圆点表示。



Principle Component Analysis 的基本性质：

Principle component analysis provides a set of eigenvectors satisfying the following properties:

（1）If the top-k eigenvectors are picked (by largest eigenvalue), then the k-dimensional hyperplane defined by these eigenvectors, and passing through the mean of the data, is a plane for which the mean square distance of all data points to it is as small as possible among all hyperplanes of dimensionality k.

（2）If the data is transformed to the axis-system corresponding to the orthogonal eigenvectors, the variance of the transformed data along each eigenvector dimension is equal to the corresponding eigenvalue. The covariances of the transformed data in this new representation are 0.

（3）Since the variances of the transformed data along the eigenvectors with small eigenvalues are low, significant deviations of the transformed data from the mean values along these directions may represent outliers.

（二）基于矩阵分解的异常点检测方法

基于矩阵分解的异常点检测方法的关键思想是利用主成分分析去寻找那些违背了数据之间相关性的异常点。为了发现这些异常点，基于主成分分析（PCA）的算法会把原始数据从原始的空间投影到主成分空间，然后再把投影拉回到原始的空间。如果只使用第一主成分来进行投影和重构，

对于大多数的数据而言，重构之后的误差是小的；但是对于异常点而言，重构之后的误差依然相对大。这是因为第一主成分反映了正常值的方差，最后一个主成分反映了异常点的方差。

假设 dataMat 是一个 p 维的数据集合，有 N 个样本，它的协方差矩阵是 X。那么协方差矩阵就通过奇异值分解写成：

其中 P 是一个 (p,p) 维的正交矩阵，它的每一列都是 X 的特征向量。D 是一个 (p,p) 维的对角矩阵，包含了特征值　　。从图像上看，一个特征向量可以看成 2 维平面上面的一条线，或者高维空间里面的一个超平面。特征向量所对应的特征值反映了这批数据在这个方向上的拉伸程度。通常情况下，可以把对角矩阵 D 中的特征值进行从大到小的排序，矩阵 P 的每一列也进行相应的调整，保证 P 的第 i 列对应的是 D 的第 i 个对角值。

这个数据集 dataMat 在主成分空间的投影可以写成

需要注意的是做投影可以只在部分的维度上进行，如果使用 top-j 的主成分的话，那么投影之后的数据集是

其中　　是矩阵 P 的前 j 列，也就是说　　是一个 (p,j) 维的矩阵，　　是一个 (N,j) 维的矩阵。如果考虑拉回映射的话（也就是从主成分空间映射到原始空间），重构之后的数据集合是

其中　　是使用 top-j 的主成分进行重构之后形成的数据集，是一个 (N,p) 维的矩阵。

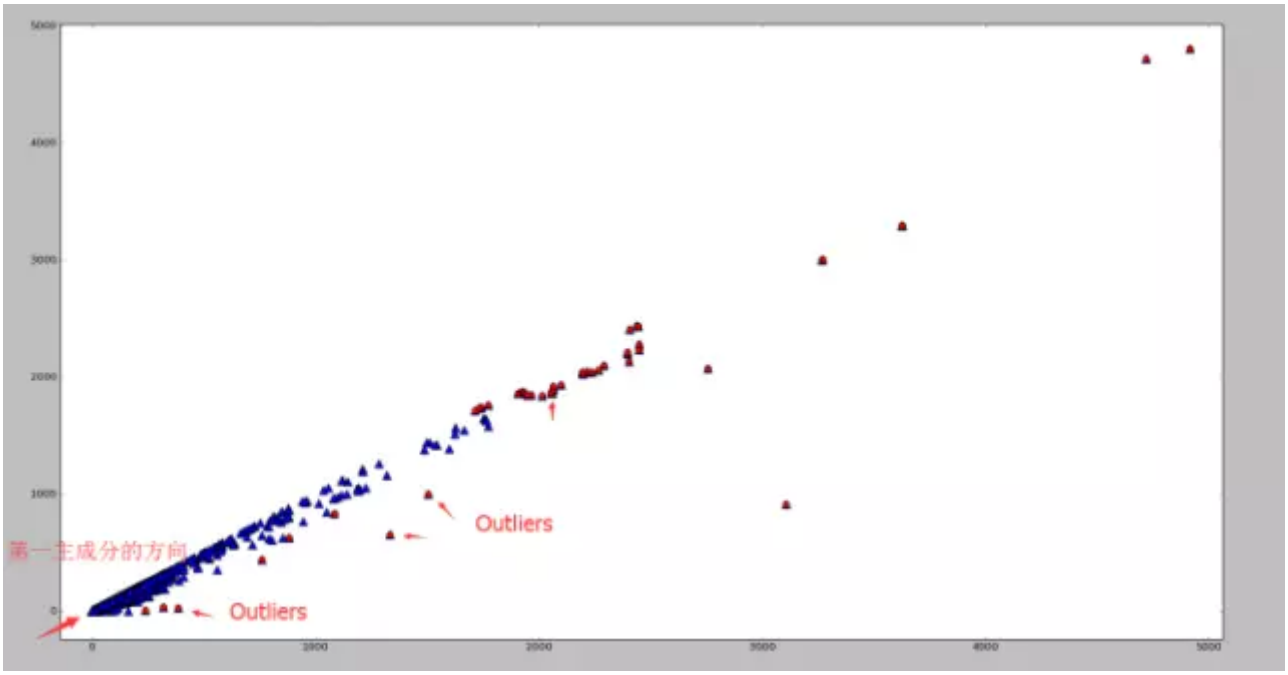
下面可以定义数据　　的异常值分数（outlier score）如下：

注意到　　指的是 Euclidean 范数， ev(j) 表示的是 top-j 的主成分在所有主成分中所占的比例，并且特征值是按照从大到小的顺序排列的。因此，ev(j) 是递增的序列，这就表示 j 越高，越多的方差就会被考虑在 ev(j) 中，因为是从 1 到 j 的求和。在这个定义下，偏差最大的第一个主成分获得最小的权重，偏差最小的最后一个主成分获得了最大的权重 1。根据 PCA 的性质，异常点在最后一个主成分上可能有着较大的偏差，因此可以获得更高的分数。

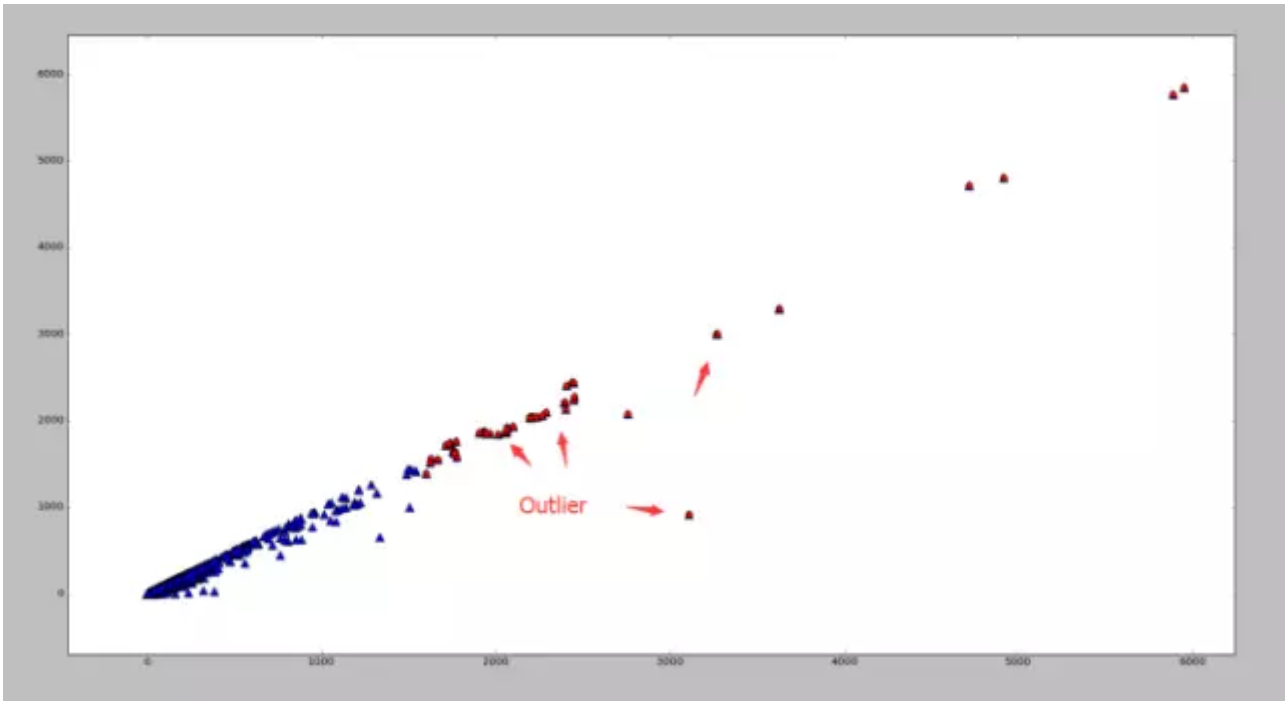
整个算法的结构如图所示：

（三）效果展示

下面两幅图使用了同一批数据集，分别采用了基于矩阵分解的异常点检测算法和基于高斯分布的概率模型的异常点算法。



基于矩阵分解的异常点检测



基于高斯分布的概率模型的异常点检测

根据图像可以看出，如果使用基于矩阵分解的异常点检测算法的话，偏离第一主成分较多的点都被标记为异常点，其中包括部分左下角的点。需要注意的是如果使用基于高斯分布的概率模型的话，是不太可能标记出左下角的点的，两者形成鲜明对比。

END

相关文章推荐：

1. 量子计算（一）
2. 特征工程简介
3. 聚类算法（一）
4. 异常点检测算法（一）