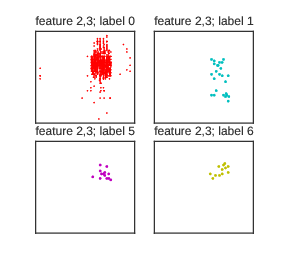
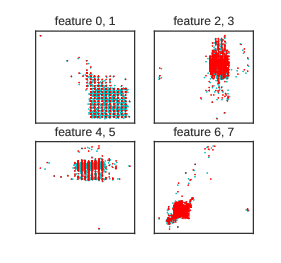
# 图&描述

图1：数据集的特征可视化



我们首先选取了数据比较有代表性的第2、3维特征进行可视化，如图。途中，四种颜色分别代表第0、1、5、6种类别，其中第0种为无故障，从图中可以很容易地看出，我们的分类任务是非常严重的不均衡分类，而且特征与故障之间联系不是十分显著。

图2：自动编码器的训练效果



我们训练自动编码器来使输出尽可能地还原输入。图中红色代表原始数据，即自动编码器的输入；蓝色代表自动编码器的输出，即还原数据。 从图中可以看出，还原的数据都在原始数据附近，表明自动编码器学习到了数据之间的关联。通过自动编码器对数据进行了有效的数据压缩即降维，可以为后续分类作准备。

图3：经过自动编码器编码后的特征

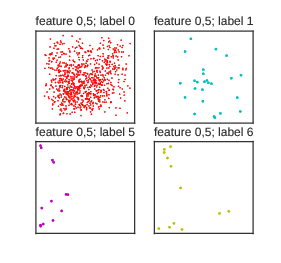


图3为自动编码器编码后的特征可视化图，其中四种颜色分别代表第0、5维特征的第0、1、5、6种类别，其中第0种类别为无故障。从图1与图3的对比可以看出，原来较为集中难以区分的数据经过自编码的特征编码之后，分散了开来，并且第5、6类故障的特征从原本与无故障的这一类几乎完全重叠变为了几乎不重叠，这样会使故障的分类更加容易，将会大幅提高准确率。

# t-SNE降维分析

## 简介：

t-SNE(t-distributed stochastic neighbor embedding)是用于降维的一种机器学习算法，是由 Laurens van der Maaten 和 Geoffrey Hinton在08年提出来。此外，t-SNE 是一种非线性降维算法，由SNE(Stochastic Neighbor Embedding, SNE; Hinton and Roweis, 2002)发展而来。

## 基本原理：

SNE是通过仿射变换将数据点映射到概率分布上，主要包括两个步骤：

1. SNE构建一个高维对象之间的概率分布，使得相似的对象有更高的概率被选择，而不相似的对象有较低的概率被选择。
2. SNE在低维空间里在构建这些点的概率分布，使得这两个概率分布之间尽可能的相似。

t-SNE是在SNE的基础上使用t分布作为选择的概率，极大地在保证数据空间关系不变的情况下进行降维，该算法通常用于将数据降维到2或3维来可视化，而本文使用该算法进行数据压缩。

## 方法：

本文使用t-SNE将数据压缩，压缩方法为：

1. 使用线性分类器SoftMax对数据D拟合，通过测试集测试分类准确率
2. 使用t-SNE将数据集D降维，得到D‘
3. 将数据集D更换为D’，重复第一步，直到SoftMax分类器分类准确率出现明显降低。

本文采用该方法，将数据集从30维压缩到24维的时候，而SoftMax分类器的分类准确率并没有出现显著降低，因此可以认为，t-SNE过滤掉了数据集中无用的信息，有效地将数据集降维。

来源：http://www.datakit.cn/blog/2017/02/05/t\_sne\_full.html#11基本原理