

分类号_____

密级_____

U D C_____

编号_____

中国科学院自动化研究所

博士后研究工作报告

异常点检测和开放集分类问题的方法研究

王 硕

工作完成日期 2018 年 5 月

报告提交日期 2018 年 6 月

中国科学院自动化研究所

2018 年 6 月

摘 要

深度学习在人工智能领域的崛起起到重要的作用。基于深度学习的模型不仅在分类，分割，聚类和其他任务中具有最佳性能，并且可以依靠其强大的计算容量处理大规模的数据。在我们的报告中，我们研究了深度学习中的两个热门问题。一个是异常点检测，另一个是对抗机器学习。

异常点检测旨在通过正常数据建立模型来检测异常数据。它在网络入侵检测，金融领域，医学图像处理等领域具有重要的作用。常见的异常点检测算法主要分为五个种类，他们分别基于概率密度函数，距离，重构，区域划分以及信息论。在这个报告中，我们提出一个基于区域划分和深度对抗网络(GAN)的异常点检测方法。与基于静态样本生成的传统异常点检测方法相比，深度对抗网络所产生的动态样本可以更加有效的刻画正常数据的边界。在我们的方法中，生成器的极小似然正则化目标函数可以使生成器产生更加有效的异常样本并且阻止生成器收敛到真实的数据分布。同时，我们设计的集成方法可以提高判别器的稳定性。通过实验可以说明我们提出的方法在CIFAR10和UCI数据集中具有显著的优势。

对抗机器学习旨在生成令机器学习模型错误分类的对抗样本并且加强模型防御对抗样本的能力。近些年，很多研究工作指出机器学习模型在取得出色的性能同时会对一些简单的样本进行错误的预测。这些样本叫做对抗样本。研究广泛的一类对抗样本旨在对原数据集产生轻微的扰动使得分类器错误分类。在我们的报告中，我们考虑一种新型的对抗样本。我们说明机器学习模型可能把一些未知类样本以高置信度错分为已知类样本。我们首先利用基于生成器的方法来生成这种新型的对抗样本。然后利用这些数据对分类器进行对抗训练，从而提高对未知类的识别能力。我们利用特征匹配方法来训练生成器，使得正常数据在特征空间的输出与生成器输出的数据一致。同时，我们利用二分类器来训练生成器，使得在原始空间中，原始样本和生成器输出的样本可以被二分类器区分。通过实验可以说明我们提出的方法对提升分类器识别未知类的性能是有效的。

关键词：对抗机器学习，异常点检测，深度生成模型

Abstract

Deep learning plays a central role in the current rise of artificial intelligence. Models based on deep learning can not only achieve great performance on classification, segmentation, clustering and many other tasks, but also deal with rapidly growing size of accessible training data because of their high computation capacities. In this report, we study two popular problems in deep learning. One is anomaly detection, the other one is adversarial machine learning.

Anomaly detection aims to detect abnormal events by a model of normality. It plays an important role in many domains such as network intrusion detection, credit card and mortgage fraud detection, medical image and so on. Existing approaches for anomaly detection can be divided into five categories: probability-based, distance-based, reconstruction-based, domain-based and information-theory-based. In this report, we propose a new domain-based anomaly detection method based on generative adversarial networks (GAN). Compared with traditional anomaly detection methods which depends on static data generation, dynamic data generation based on GAN can describe the boundary of normal data more effectively. In our methods, minimum likelihood regularization which regularizes the generator can make it produce more anomalies and prevent it from converging to normal data distribution. We also propose proper ensemble of anomaly scores to improve the stability of discriminator effectively. In our experiment, we show that our method has achieved significant improvement than other anomaly detection methods on Cifar10 and UCI datasets.

Adversarial machine learning is also widely studied for treating the weakness of machine learning models. Recently, a lot of work has been done to demonstrate that machine learning models are vulnerable to make mistakes on some simple samples, although they make great performance in many tasks. Such simple samples are called adversarial samples. One well-studied kind of adversarial samples is defined as perturbing the inputs subtly to lead the models to make

wrong decisions. In this report, we treat another kind of adversarial samples which is closely related to classical adversarial samples. We demonstrate that machine learning models can also misclassify some data which are quite different from original ones as some known classes in high confidence scores. We first produce such data by a generator and then propose adversarial training for the classifier to improve the performance of recognition. We use feature matching to train the generator such that the output of real data in feature space are the same as that are produced by the generator. We also train a binary classifier whose positive class is real data and negative class is that produced by the generator. Our generator is trained to produce samples which are easily recognized by the binary classifier in original space and difficultly recognized in feature space by the original classifier. In our experiment, we show that our algorithm is effective to improve the performance of classifier to recognize unknown classes.

Key Words: adversarial machine learning, outlier detection, deep generative models

目 录

摘要	i
Abstract	iii
目录	v
第一章 引言	1
第二章 异常点检测方法综述	5
2.1 基于距离的异常点检测方法	5
2.1.1 基于最近邻距离的异常点检测方法	5
2.1.2 基于聚类的异常点检测方法	8
2.1.3 基于距离的开放集分类方法	8
2.2 基于概率密度函数的异常点检测方法	9
2.2.1 基于统计量分析的异常点检测方法	10
2.2.2 基于极值定理的开放集分类方法	11
2.2.3 基于概率密度函数的异常点检测方法	15
2.3 基于重构的异常点检测方法	17
2.3.1 基于重构的异常点检测方法	17
2.3.2 基于子空间的异常点检测方法	18
2.4 基于正规区域划分的异常点检测方法	19
2.4.1 基于OCSVM的单类分类器	19
2.4.2 基于SVDD的单类分类器	20
2.4.3 基于单类分类器的开放集识别问题	20
2.5 基于集成的异常点检测方法	21
2.6 基于信息论的异常点检测方法	22
2.7 异常点检测方法的比较方法举例	22

第三章 基于生成对抗网络的单类分类器	25
3.1 生成对抗网络	25
3.2 极小似然对抗生成网络	26
3.3 克服判别器的不稳定性	28
3.4 实验结果	28
3.4.1 玩具数据集的可视化	28
3.4.2 公开数据集的实验结果	29
第四章 基于对抗训练的开放集分类方法	35
4.1 对抗机器学习：生成对抗样本	35
4.2 对抗机器学习和异常点检测方法	41
4.3 基于对抗机器学习的开放集分类问题	42
4.4 基于提高分类器可分性的新算法	42
4.5 实验结果	45
第五章 总结和展望	49
参考文献	51
发表文章目录	61
致谢	63

表 格

3.1	Cifar10数据集的AUC值	30
3.2	UCI数据集的实验设置	33
3.3	UCI数据集的AUC值	33
4.1	SVHN数据集识别未知类的AUC值	46
4.2	SVHN数据集的分类正确率	47

插 图

3.1 生成器 G 在KL散度正则化下的性能对比。红色的点表示正常数据的流形，蓝色的点表示 G 的输出分布。图(a)和图(c)表示在没有KL散度正则化下，生成器 G 输出的分布。图(b)和图(d)表示在KL散度正则化下生成器 G 输出的分布。	29
3.2 Cifar10数据集集成的AUC值. 当基判别器的个数增多时，集成模型的AUC值逐渐增加。当集成个数达到10时，集成的性能达到稳定。	31
3.3 Cifar10数据集集成的AUC值. 当基判别器的个数增多时，集成模型的AUC值逐渐增加。当集成个数达到5时，集成的性能达到稳定。	31
3.4 Cifar10数据集的ROC曲线.	32
3.5 Cifar10数据集的异常值分布. 箱形图的顶端和低端分别代表第一和第三中位数。箱形图中的绿色线代表中位数。	32
4.1 SVHN数据集和新型对抗样本	43

第一章 引言

异常点检测技术旨在对正常数据进行建模来检测未知或者异常的数据特征。异常点检测技术是数据挖掘和机器学习领域的重要分支。与传统的监督学习，半监督学习和无监督方法相比，异常点检测最显著的特点是所处理的数据集缺乏异常数据的信息。在实际应用中，缺乏可以利用的异常数据是广泛存在的现象。人工收集和标注异常样本需要大量的资源和代价。同时，由于异常样本的多样性，收集充足的异常样本不可能完全实现。因此，异常点检测技术在数据挖掘领域中具有不可替代的地位。与机器学习领域的其它算法相比，异常点检测方法面临更多的技术困难和挑战。

识别异常的数据特征在现今社会的各个领域具有广泛应用。在互联网和计算机系统中，有多种类型的海量数据需要传播和处理。异常点检测可以通过检测数据流中的异常数据来识别入侵行为，进而维护互联网和计算机的安全信息不被盗用，防止信息处理系统被破坏。在 [34] [77] 中，异常点检测技术被用来检测诈骗行为，计算机攻击和盗用以及木马病毒的传播。信用卡诈骗是金融安全领域的一个重要威胁。由于信用卡用户的敏感信息被盗取而造成的信用卡盗用会给用户造成大量的损失。异常点检测技术可以通过检测用户的消费记录，消费模式和消费金额等信息来确认未经用户授权的不正当消费行为。在 [71] 等文献中，一些针对金融和财务系统中出现的海量数据的异常点检测方法被提出。这些方法可以有效的解决数据中的诈骗和虚假信息，提高了金融数据的安全性。在医学图像处理中，异常点检测作为辅助医疗技术可以有效的检测疾病特征。在 [69] 中，一个基于深度生成模型的方法被用来检测疾病的传播状况。异常点检测技术还可以应用于工业设备的检测和探伤 [15]，图像和音频监控 [60]，文本挖掘 [2] 以及传感器网络 [12] 等领域。

异常点检测方法的设计依赖于具体的数据类型。多数异常点检测方法基于多维的数值型数据。在现实数据中，异常点检测方法还需要考虑其他的数据类型，例如类别属性、文本属性、时间序列和离散序列属性、空间属性和图数据等。

根据对数据集的认知情况，异常点检测可以分为监督情形，正类与无标签(positive-unlabeled class)情形，半监督情形和无监督情形。监督情形是指已

知训练集中的所有样本是否为正常数据和异常数据，通过训练集中的异常数据来增强模型对异常数据的识别能力。由于异常数据所占的比例较小，所以这类情形以看作是类别不均匀的分类问题。当训练集中的标记样本只有正常数据，无标记样本既包含正常数据又包含异常数据，利用无标签数据来增强模型对异常数据的识别能力叫做正类与无标签情形。正类与无标签问题在互联网和社交网络的问题中具有广泛的应用。当训练数据只包含正常数据或者异常数据时，叫做半监督问题。无监督情形对应的训练数据无标记并且既包含正常数据又包含异常数据。

对于监督情形，一个基本的假设是将异常数据错误分类为正常数据比将正常数据错误分类成异常数据所付出的代价更大。所以针对类别不均匀的监督情形，算法的策略主要基于代价敏感(cost-sensitive)学习和自适应的重采样(adaptive re-sampling)。在代价敏感学习中，算法对正常数据和异常数据的分类误差加权处理，使得异常数据的分类误差具有更高的权值。基于自适应的重采样方法主要提高异常数据在数据集中的比例。这类方法主要分为两种。第一种方法是启发式地辨别无标签数据中的正常数据和异常数据，将问题转化成对正常数据和异常数据的分类问题。第二类方法将无标签训练数据进行加权。对于半监督情形和无监督情形，异常点检测的方法相似，主要分为5个类别，即基于统计分析的方法，基于距离的方法，基于概率密度函数的方法，基于区域划分的方法，基于重构的方法以及基于信息论的方法。其中基于统计分析、基于距离和信息论的方法旨在利用统计量和距离来定义数据间的相似度和数据的异常值，异常值的计算不需要训练模型。基于概率密度、区域划分以及重构的方法需要训练模型来计算异常值。

异常点检测的算法通过异常值(outlier score)来衡量测试样本的异常程度。模型根据异常值和给定的阈值关系来确定测试样本是否属于异常样本。衡量算法的有效性的指标主要是AUC(ROC)值和AUC(PPV)值。给定阈值和数据集D。被模型判定为异常的数据集合为S(t)，数据集中的异常数据集合为G。则精度(precision)的定义如下：

$$\text{Precision}(t) = 100 \cdot \frac{|S(t) \cap G|}{|S(t)|}. \quad (1.1)$$

精度不是阈值t的单调函数。召回率(recall)的定义如下：

$$\text{Recall}(t) = 100 \cdot \frac{|S(t) \cap G|}{|G|}. \quad (1.2)$$

通过改变阈值t的取值，可以绘制一条关于精度和召回率的曲线，这个曲线叫做PRC曲线。

ROC曲线与PRC曲线定义类似。ROC曲线取决于改变阈值t的取值，True Positive Rate(TPR(t))和False Positive Rate(FPR(t))的变化关系，其中

$$TPR(t) = \text{Recall}(t) = 100 \cdot \frac{|S(t) \cap G|}{|G|}, \quad FPR(t) = 100 \cdot \frac{|S(t) - G|}{|D - G|}. \quad (1.3)$$

给定ROC曲线，AUC值定义为ROC曲线下方的面积。AUC值衡量了阈值从负无穷到正无穷时模型识别异常数据的平均性能。

在这个报告中，我们介绍利用GAN来做正常区域划分的异常点检测方法。与之前基于GAN的异常点检测方法相比，我们的方法利用GAN的判别器来刻画正常数据的边界，利用GAN输出的大小来探测异常点。我们方法的核心思想是利用训练数据作为正常数据，将GAN训练的过程中生成器的输出作为异常数据来训练分类器。与传统的单类分类器相比，我们的异常数据是动态生成的，并且随着生成器输出分布接近于样本分布时，判别器所对应的分类边界会接近于真实的数据分布，从而提高刻画正常数据边界的质量。GAN在训练的过程中，随着生成器的分布收敛到真实的数据分布，判别器会接近于常值函数，从而影响刻画正常数据边界的性能。我们利用KL散度构造生成器的正则化目标函数，使得生成器不收敛到真实的数据分布并且在训练的过程中产生更多的异常样本，从而提高判别器的识别性能。由于GAN生成样本的随机性，生成器的性能会出现不稳定性。我们提出判别器的集成算法，从而克服判别器的不稳定性。我们把提出的方法同其他的的异常点检测方法OCSVM，Isolation Forest，VAE和AE相比较。实验结果说明我们的方法在CIFAR10和UCI数据集比比较方法取得更好的效果。

在这个报告中，我们介绍一个基于对抗机器学习来解决开放集分类问题的新方法。在经典的分类问题中，神经网络由于强大的特征提取能力在多种任务中取得了最佳的分类性能。但是基于softmax的经典分类目标函数将特征区域划分成和训练集类别个数相同的区域。因此，通过softmax的输出无法判别测试样本的类别是否与已知类别不同。已有的大部分开放集分类方法在神经网络的特征区域内设计更加合理的未知类识别指标。神经网络的特征空间没有经过训练来增强对未知类别的识别能力。通过正则化神经网络特征空间来增强神经网络识别未知类能力的工作主要基于对抗机器学习。对抗机器学习旨在查找和

原始数据相似，但是被分类器错分的样本。对抗机器学习中的对抗训练方法可以增加神经网络的光滑性。从而使已知类的样本聚集在已知类别区域的能力增强。我们的实验说明神经网络会把一些与原始样本相似度较低的数据映射到特征空间的已知类区域中。因此，神经网络会将一些未知类样本错误的判定为已知类。我们设计生成器来生成满足特定要求的样本。这些样本在原始空间和原始样本不同，但是在特征空间与原始样本类似。因此，我们利用特征匹配的目标函数正则化生成器，使得生成器输出的样本在特征空间中和原始样本的分布相同，利用二分类神经网络来区分原始样本和生成器输出的样本。实验结果说明我们的方法对增强分类器识别未知类的性能是有效的。

我们的报告结构如下：第二章介绍异常点检测和开放集分类的典型方法。第三章介绍基于GAN的新型单类分类器。第四章介绍对抗机器学习的典型算法以及基于对抗机器学习解决开放集分类问题的新方法。第五章对我们的工作进行总结和展望。

第二章 异常点检测方法综述

2.1 基于距离的异常点检测方法

基于距离的异常点检测方法是异常点检测算法中一类常见的算法。它主要包括基于最近邻的方法和基于聚类的方法。基于距离的异常点检测方法的优势是不需要知道数据分布的先验知识，不需要对数据的分布建模。基于最近邻的方法需要适合的距离来衡量数据点之间的相似度。与此同时，很多基于最近邻的方法不能识别全局的异常点并且不能辨别数据集中由于形状和密度不同而产生的局部异常点。基于聚类的方法在增类模型中更加适用。在高维空间中，基于距离的方法在估计距离的操作具有较高的时间复杂度。近些年，一些减少时间复杂度的方法被提出，从而增加了大规模问题的求解效率。

2.1.1 基于最近邻距离的异常点检测方法

基于最近邻距离方法的核心方法是k近邻域法。k近邻域法假设正常数据到训练集中的正常数据点较小，异常数据到训练集中的正常数据的距离较大。异常值的定义主要基于异常数据的k近邻距离。距离越大，则异常值越高。欧氏距离和马尔科夫距离是常用的距离。此外，对于类别属性等非数值属性，匹配个数是常用的距离。基于最近邻距离的异常值主要有三种定义方式，分别为准确k近邻，平均k近邻以及调和k近邻：

定义 2.1. 给定数据集 $D = \{x_1, \dots, x_n\}$ ，数据 x_i 的异常值定义为下面三种形式：

- 准确k近邻： x_i 到 $D - \{x_i\}$ 的第k个最近的数据点的距离。
- 平均k近邻： x_i 到 $D - \{x_i\}$ 的前k个最近的数据点的距离的平均值。
- 调和k近邻： x_i 到 $D - \{x_i\}$ 的第k个最近的数据点的距离的调和平均值。

除了常用的距离，共享的k近邻距离(shared KNN) [30]和逆k近邻距离(ODIN) [26]可以度量数据点之间的相似度。共享的k近邻距离计算两个数据点的k近邻

域中公共数据的个数。它不仅与两个点之间的距离有关，而且与整个数据集中数据点分布的统计性质有关。逆k近邻距离和k近邻距离定义相似。给定数据点p 和q， p是q的逆k近邻当且仅当q是p的k近邻。

基于k近邻距离的异常点检测方法衡量了给定的数据点所在区域密度的大小。局部异常因子(LOF)衡量数据点所在区域密度变化的快慢。给定数据点x和y，令 $D^k(x)$ 为x的k近邻距离， $L_k(x)$ 为x的k近邻距离中的数据点。在LOF方法中，异常值 LOF_k 通过x关于y的可达距离 $R_k(x, y)$ 来定义。可达距离 $R_k(x, y)$ 是欧式距离的平滑化，其定义如下：

$$R_k(x, y) = \max\{\text{dist}(x, y), D^k(x)\}, \quad (2.1)$$

其中 $\text{dist}(x, y)$ 为x与y的欧式距离。平均的可达距离 $AR_k(x)$ 定义为

$$AR_k(x) = \text{MEAN}_{y \in L_k(x)} R_k(x, y). \quad (2.2)$$

LOF方法中，异常值定义为

$$LOF_k(x) = \text{MEAN}_{y \in L_k(x)} \frac{AR_k(x)}{AR_k(y)} = AR_k(x) \cdot \text{MEAN}_{y \in L_k(x)} \frac{1}{AR_k(y)} \quad (2.3)$$

$$= \frac{AR_k(x)}{\text{HMEAN}_{y \in L_k(x)} AR_k(y)}, \quad (2.4)$$

其中HMEAN表示调和平均。当x为正常数据时， $LOF_k(x) \approx 1$ 。当x为异常数据时， $LOF_k(x) > 1$ 。异常值 $LOF_k(x)$ 有很多变体。可达距离可以替换成原始的欧式距离，调和平均可以换成算术平均值和两个点的成对距离(LDOF)。当数据集中存在距离非常小的数据点或者重复的数据点时，可以利用很小的常数 $a > 0$ 来正则化LOF：

$$LOF_k(x) = \frac{a + AR_k(x)}{a + \text{HMEAN}_{y \in L_k(x)} AR_k(y)}. \quad (2.5)$$

LOF算法的计算复杂度较高。因此，减小LOF距离的计算量具有重要的意义。在 [60]中，作者提出了LOF的增量形式。根据 $LOF_k(x)$ 的定义， $LOF_k(x)$ 与数据点的局部邻域有关。给定测试样本，只有该测试样本邻域内的样本可以影响该样本的 LOF_k 。在 [76]中，作者提出减少冗余的LOF计算策略。由于LOF方法需要选取具有前N个最高异常值的数据，计算所有数据的异常值需要大量的计算量。因此，作者提出多细粒度剪枝策略来将数据集划分成若干个

子区域。方法在每一个子区域子区域估计LOF的上界，去除LOF上界充分小的数据。经过论证，这些数据点不可能具有前N个最高的异常值。在 [75]中，作者提出了LOF的分布式计算方法。

最近邻距离法利用k个最近数据点的距离来衡量数据x所在区域的密集程度，局部关联积分(LOCI) [54]利用数据x的邻域内数据的个数来衡量所在区域的密集程度。LOCI方法中，数据点x关于半径 ϵ 的密度 $M(x, \epsilon)$ 定义为x的 ϵ 邻域中点的个数。利用 $M(x, \epsilon)$ ， x 邻域内的平均密度定义为

$$AM(x, \epsilon, \delta) = \text{MEAN}_{y: \text{dist}(x, y) \leq \delta} M(y, \epsilon). \quad (2.6)$$

水平为 δ 的多细粒度偏移因子 $MDEF(x, \epsilon, \delta)$ 定义为

$$MDEF(x, \epsilon, \delta) = 1 - \frac{M(x, \epsilon)}{AM(x, \epsilon, \delta)}. \quad (2.7)$$

作为异常值指标， $MDEF(x, \epsilon, \delta)$ 越大，数据x是异常数据的可能性越大。 $MDEF(x, \epsilon, \delta)$ 的具有较高的计算复杂度。在邻域中抽样计算，对领域中点的个数采用更加高效的统计方法是减小 $MDEF(x, \epsilon, \delta)$ 计算复杂度的有效方法。

2.1.1.1 其它的异常点检测方法

除了基于点与点的距离，点到子空间距离也可以作为异常点检测的准则。在 [53]中，作者假设正常数据到流形邻域的点生成的线性子空间的距离较小。作者在数据点x的邻域中寻找若干个稠密点集，然后利用x到这些稠密点集所生成的子空间的距离作为异常值。距离越大，x是异常数据的可能性越高。

在 [78]中，作者通过稀疏线性表示的系数来度量数据点之间的相似度。令 X 为数据点 x_j 为列的矩阵。作者利用下面的优化问题求解数据点每一个数据点 x_j 基于其余数据点的线性表示：

$$\min \lambda \|r_j\|_1 + \frac{1-\lambda}{2} \|r_j\|_2^2 + \frac{\lambda}{2} \|x_j - Xr_j\|_2^2 \quad s.t. \quad r_{jj} = 0. \quad (2.8)$$

基于优化问题的求解系数 r_i ，作者构造马尔科夫转移矩阵：

$$p_{ij} = \|r_{ji}\| / \|r_i\|_1 \text{ 对所有的 } \{i, j\} \subset 1, 2, \dots, N. \quad (2.9)$$

最后，作者通过计算马尔科夫转移矩阵的稳定分布来计算测试数据的异常值。

基于夹角的异常点检测方法(ABOD)是衡量数据点相似度的常见算法。ABOD通过数据点和其他两个数据点组成向量的夹角来衡量数据的异常值。给定数据集 D 和数据点 x, y, z 。 x 关于 y 和 z 的夹角定义为： $\frac{\langle \vec{xy}, \vec{xz} \rangle}{\|\vec{xy}\|^2 \cdot \|\vec{xz}\|^2}$ 。异常值ABOF(x)为 x 关于 y, z 夹角的方差，其定义如下：

$$VAR_{y,z \in D} \left(\frac{\langle \vec{xy}, \vec{xz} \rangle}{\|\vec{xy}\|^2 \cdot \|\vec{xz}\|^2} \right). \quad (2.10)$$

ABOD假设正常数据聚集在一个类区域。对于正常数据，其夹角随着数据 y, z 的选取变化较大。由于异常数据距离正常数据的类区域较远，其夹角与 y, z 的选取不敏感，因此夹角的方差较小。

2.1.2 基于聚类的异常点检测方法

基于聚类的异常点检测方法旨在检测测试数据是否属于某一个已知类别(cluster)。异常值的定义根据它到最近类别的距离以及它到其它类别的距离。利用聚类算法训练数据集分解成若干个稠密子集，异常检测主要检测与这些稠密子集距离较远的点，任何一个聚类算法都可以作为异常点检测方法。与基于最近邻域距离的方法相比，基于聚类的异常点检测方法具有更低的计算复杂度，但是需要更多的样本来进行训练。基于聚类的异常点检测方法性能对聚类算法中超参数的选取比较敏感。

在基于聚类的异常点检测方法中，异常值的设定对检测效果影响很大。如果异常值设定不合理，成功检测的异常数据中包含了大量的弱异常点或者噪声。常用的异常值将测试样本到各类中心的马氏距离作为指标。异常值的设定也可以根据聚类算法的优化方算法来设定。例如在k-means聚类算法中，测试样本的类别可以根据该测试样本到类别中心的马氏距离来确定。除了基于距离的准则，给定邻域内所含有的样本个数也可以作为异常值的指标。

谱聚类是基于矩阵特征值分解的聚类方法。谱聚类作用于图上，图中的每一个结点都代表一个数据点。数据点之间的相似度利用仿射矩阵来表示。在 [28] 中，作者首先解释了谱聚类在进行聚类的同时对异常点的影响。同时，作者说明异常点所在类别的奇异性与异常点类别个数的关系。最后作者提出有效的准则来估计异常点类别的个数。

2.1.3 基于距离的开放集分类方法

基于距离的异常点检测方法在开放集分类的问题之中具有广泛的应用。

在 [33] 中，作者提出基于最近邻距离的检测异常类别的 OSNN 准则和 NNDRS 准则。在 OSNN 准则中，给定测试样本 x 和二近邻点 y, z 。如果 y 和 z 属于相同类别，则 s 的类别被判定为 y 和 z 的类别。否则，判定 s 为未知类。NNDRS 准则首先计算距离 x 最近的点 y ，然后计算距离 x 最近但是类别与 y 不同的数据 z 。最后，将距离的比值 $R = d(x, y)/d(x, z)$ 作为 x 的异常值。 R 越大，则 x 是异常数据的可能性越高。

在 [63] 中，作者首先预训练一个分类器，对于每一个类别 l ，作者利用类模型 $M_l = (c_l, r_l)$ 学习该类别样本的类中心 c_l 和类半径 r_l 。这里，类中心 c_l 定义为该类别样本的平均值，类半径 r_l 定义为同类别样本到 c_l 的最大距离。样本 x 对于类别 l 的异常值 ϕ 定义为

$$\phi(x, l) = \frac{d(x, c_l)}{r_l}. \quad (2.11)$$

异常值越大，则样本 x 为异常点的可能性越高。

在 [32] 中，作者提出具有拒识能力的二分类支持向量机。该支持向量机确保在原始空间中，正类区域是有界的。作者针对高斯函数核函数论证了正类区域有界的充分必要条件是分类超平面位移小于零。作者同时提出新的优化算法来确保支持向量机分类超平面的位移小于零。

基于距离的开放集分类方法主要以特征空间中点之间的距离为度量。特征空间的训练和距离度量是独立的。在 [25] 中，作者提出特征空间正则化的方法来提高分类器的拒识能力。作者正则化特征空间的准则是令已知类训练数据的类内距离最小，类间距离最大。异常值通过测试样本到已知类的类中心距离的最小值来定义。距离越大，测试样本是异常点的可能性越大。

2.2 基于概率密度函数的异常点检测方法

基于概率密度函数的异常点检测方法旨在估计数据的生成概率密度函数。通过生成概率密度函数来估计正常数据的边界。基于概率密度函数的异常点检测方法具有良好的数学框架并且在实际中具有更高的性能。但是当样本数量较少时，基于概率密度函数的异常点检测方法的性能会受到限制。在高维空间下，这一缺点尤为显著。拟合数据生成分布的模型主要基于统计学中的统计量以及机器学习中的概率生成模型。

2.2.1 基于统计量分析的异常点检测方法

基于统计分析的异常点检测方法不需要数据分布的全部信息，只要求数据分布中极值点的分布情况。已有工作主要基于统计学的极值理论。极值理论是统计学中的一个重要分支，主要研究数据分布中的极大值和极小值的分布情况。如果数据满足常见的分布，例如高斯分布，t分布和卡方分布，则具有标准的尾分布估计方法。对于一般的分布，常见的极值理论主要包括估计尾分布的概率不等式。例如马尔科夫不等式：

$$P(X > a) \leq E[X]/a \quad (2.12)$$

切比雪夫不等式：

$$P(|X - E(X)| > a) < \text{Var}[X]/a^2 \quad (2.13)$$

切尔诺夫不等式：

$$P(X < (1 - \delta) \cdot E[X]) < \exp^{-E[X] \cdot \delta^2/2} \quad (2.14)$$

霍夫丁不等式：

$$P(X > (1 + \delta) \cdot E[X]) < \exp^{-E[X] \cdot \delta^2/4} \quad (2.15)$$

等。

除了常见的不等式，基于统计分析的方法还利用统计学中的极值定理(EVT)来拟合统计学中的尾分布。下面两个极值定理被应用到开放集分类的问题之中并且取得了良好的效果。

定理 2.1. EVT第一定理：令 $\{s_i\}$ 为独立同分布的样本， $M_n = \max\{s_1, \dots, s_n\}$ 。假设存在 a_n, b_n ，使得对每一个 $a_n > 0$ ，

$$\lim_{x \rightarrow \infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x). \quad (2.16)$$

则 F 是一个非退化的数据分布，并且属于三个极值分布其中的一个。极值分布主要为耿贝尔分布，弗雷歇分布和威布尔分布。

极值定理中的三个极值分布可以融合成广义的极值分布(GEV)：

$$GEV(t) = \begin{cases} \frac{1}{\lambda} \exp^{-v^{-\frac{1}{k}}} v^{-\frac{1}{k}+1}, & k \neq 0 \\ \frac{1}{\lambda} \exp^{-(x+\exp(-x))}, & k = 0 \end{cases} \quad (2.17)$$

这里， $x = \frac{t-\tau}{\lambda}$, $v = (1 + k \frac{t-\tau}{\lambda})$ 。 k, λ 和 τ 分别表示形状，尺度和位置参数。

定理 2.2. *EVT第二定理：* 给定极值分布 G_γ :

$$G_\gamma : x \rightarrow \exp(-(1 + \gamma x)^{-\frac{1}{\gamma}}), \quad \gamma \in \mathbb{R}, \quad 1 + \gamma x > 0. \quad (2.18)$$

分布函数 $F_t(x)$ 依分布收敛到 G_γ 当且仅当存在 σ , 使得对所有满足条件 $1 + \gamma x > 0$ 的样本 x , 下面的等式成立:

$$\frac{\bar{F}(t + \sigma(t)x)}{\bar{F}(t)} \xrightarrow{t \rightarrow \tau} (1 + \gamma x)^{-\frac{1}{\gamma}}. \quad (2.19)$$

EVT第二极值定理可以表述为:

$$\text{当 } X \text{ 充分大时, } F_t(x) = P(X - t > x | X > t) \sim_{t \rightarrow \tau} (1 + \frac{\gamma x}{\sigma(t)})^{-\frac{1}{\gamma}}. \quad (2.20)$$

在 [72]中, 作者利用EVT第二极值定理来拟合数据分布X的尾分布。作者利用极大似然估计估计参数 γ 和 σ 。令 z_q 为使得条件 $P(X > z_q) < q$ 成立的最小值。则

$$z_q \simeq t + \frac{\gamma}{\sigma} \left(\left(\frac{qn}{N_t} \right)^{-\gamma} - 1 \right), \quad (2.21)$$

其中, t 是使得数据分布充分接近极值分布的阈值, q 是概率阈值, n 是拟合样本的个数, N_t 是取值大于 t 的样本的个数。

2.2.2 基于极值定理的开放集分类方法

基于极值定理的开放集分类方法中, 极值定理用来判定异常值的合理性。基于极值定理的方法估计正常数据的尾分布指标, 并且识别异常值不合理的数据。在 [68]中, 作者提出利用极值定理来判定未知类数据的重要方法。

算法 2.1. 输入: 数据集 S 。

步骤 1 查找数据集 S 中的 n 个最大值。按照从小到大排列, 记作 $s_1, \dots, s_n \in S$ 。

步骤 2 利用广义极值分布拟合 s_2, \dots, s_n , 记作 W 。

步骤 3 如果 $W(s_1)$ 大于给定阈值 δ 。则 s_1 被判定为异常点。

在 [29]中, 作者利用极值理论提出SVM的未知类识别准则。具有未知类识别机制的SVM叫做PI-SVM。对于多分类SVM, 作者采用one-vs-rest策略将多

分类转化成二分类问题。对于每一个二分类问题，作者考虑二分类问题中的正类样本的未知类识别问题。作者利用样本点到分类超平面的有向距离作为未知类的指标。距离分类边界越近，样本点属于未知类的可能性越大。PI-SVM首先计算正类数据与分类边界的距离，并且取出距离较小的数据。根据极值理论，将这些数据拟合成维尔布分布： $1 - \exp^{-\left(\frac{x-\tau_y}{\lambda_y}\right)^{\kappa_y}}$ 。给定测试数据x，其后验概率密度函数定义为

$$P_I(y|x, \theta_y) = \rho(y)P_I(x|y, \theta_y) = \rho(y)(1 - \exp^{-\left(\frac{x-\tau_y}{\lambda_y}\right)^{\kappa_y}}), \quad (2.22)$$

其中 $\rho(y)$ 表示先验概率密度函数。给定测试样本x，PI-SVM利用下面的表达式来确定x是否属于未知类：

$$y^* = \arg \max P_I(y|x, \theta_y), \quad s.t. \quad P_I(y^*|x, \theta_{y^*}) \geq \delta. \quad (2.23)$$

除了距离边界较近的数据，距离正类样本分布较远的点属于未知类的可能性也较大。在 [67]中，作者提出了CAP模型来刻画数据点到正类样本的距离。CAP模型与测试随着数据点与正类样本的距离增加而逐渐减小。作者将CAP模型极值应用在PI-SVM模型之中，从而增强PI-SVM识别未知类的能力。作者将新的模型命名为W-SVM模型。

CAP模型没有特定的函数形式。在传统方法中，分类问题主要基于SVM，基于SVM的CAP模型的基本形式如下：

$$M(x) = p_f(F(K(x, x_1) \dots K(x, x_m))), \quad (2.24)$$

这里K表示核函数， p_f 是单调递减的概率密度函数。F表示具有衰减性质的函数：存在 $x' \in X$ 和 $A_{x'}$ ，使得下面的不等式成立：

$$F(K(x, x_1) \dots K(x, x_m)) \leq A_{x'}(\|x' - x\|). \quad (2.25)$$

下面的定理中，作者证明基于径向基函数的单类支持向量机(OCSVM)属于CAP模型。

定理 2.3. 给定数据点 x_i 。令OCSVM为以径向基函数为核函数的单类SVM。令 a_i 为对应的拉格朗日乘子，则 $\sum_i a_i y_i K(x, x_i)$ 构成CAP模型。

W-SVM首先对正类利用单类SVM做建模，建立CAP模型。如果CAP模型判定样本为未知类，则算法判定样本为未知类。如果CAP模型判定样本为已知类，W-SVM利用极值理论来判定是否过于接近分类边界。由于正类距离分类边界的距离具有下界，作者利用维尔布 P_η 来说明数据点是正类样本的分布并根据分布来判定是否为异常数据，其中

$$P_\eta(y|f(x)) = 1 - e^{-(\frac{-f(x)-v_\eta}{\lambda_\eta})^{\kappa_\psi}}, \quad (2.26)$$

$v_\eta, \lambda_\eta, \kappa_\psi$ 为待拟合的参数。由于负类距离分类边界的距离具有上界，利用逆维尔布分布 P_ψ 来计算样本不属于负类样本的概率，这里

$$P_\psi(y|f(x)) = e^{-(\frac{f(x)-v_\eta}{\lambda_\eta})^{\kappa_\psi}}, \quad (2.27)$$

$v_\eta, \lambda_\eta, \kappa_\psi$ 为待拟合的参数。

因此，W-SVM的目标函数如下：

$$y^* = \arg \max P_{\eta,y}(x) \times P_{\psi,y} \times \iota_y \quad s.t. \quad P_{\eta,y}(x) \times P_{\psi,y} \geq \theta_R. \quad (2.28)$$

如果OVSVM的CAP模型判定为正类，则 $\iota_y = 1$ 。否则， $\iota_y = 0$ 。

在 [5]中，作者将CAP模型添加到NCM分类器 [44]中，提出了新的开放集分类算法(NNO)。NCM是基于距离的分类器。假设分类器含有C个类别。给定测试样本x，分类算法如下：

$$c^* = \operatorname{argmin}_{c \in \{1, \dots, C\}} d(x, \mu_c), \quad s.t. \quad \mu_c = \frac{1}{N_c} \sum_{i:y_i=c} x_i. \quad (2.29)$$

作者将CAP模型设置为 $\nu(x) = \max_c 1 - d(x, \mu_c)$ 。当 $\nu(x) < 0$ 时，算法判定x为未知类。

在 [61]中，作者提出极值机模型(EVM)。EVM的原理是考虑不同类别之间点的距离的分布。给定数据集 $\{(x_i, y_i)\}$ 和数据集中的两个数据 x_i 和 x_j 并且满足 $y_i \neq y_j$ 。定义边缘距离 $m_{ij} = \|x_i - x_j\|/2$ 。作者利用极值理论证明边缘分布 m_{ij} 的极小值满足维尔布分布。给定测试样本 x' ，作者利用维尔布分布构造 x' 的概率密度函数：

$$\Psi(x_i, x', \kappa_i, \lambda_i) = \exp^{-\left(\frac{\|x_i-x'\|}{\lambda_i}\right)^{\kappa_i}}, \quad (2.30)$$

其中 κ_i 和 λ_i 是拟合的维尔布分布参数。 x' 属于某一类别 C_l 的概率定义为

$$\hat{P}(C_l|x') = \arg \max_{i:y_i=C_l} \Psi(x_i, x'; \kappa_i, \lambda_i). \quad (2.31)$$

分类决策定义为

$$y^* = \begin{cases} \arg \max_{l \in \{1, \dots, M\}} \hat{P}(C_l | x') & \text{如果 } \hat{P}(C_l | x') \geq \delta, \\ \text{'未知类'} & \text{否则.} \end{cases} \quad (2.32)$$

在 [6] 中，作者将CAP模型应用到深度神经网络分类器中，得到具有拒识能力的分类器openmax。作者考虑softmax前一层(logit层)的分布情况。深度神经网络中CAP模型的核心思想是该层每一个结点的输出不能距离正常数据集在该结点上输出均值过大。CAP模型随着该距离的增大而逐渐衰减。CAP值越小，说明该数据是异常数据的可能性越高。算法的步骤如下：

算法 2.2. 输入： 数据集 S 。

步骤 1 计算第 i 类数据 x_i 在 logit 层第 i 个结点的输出 v_1, \dots, v_n 和均值 μ_i 。

步骤 2 利用维尔布分布来拟合 $\|x_i - \mu_i\|$ 的尾分布 w_i 。

步骤 3 给定测试样本 x ，计算 $v(x), w(x)$ ，并且计算

$$\hat{v}(x) = v(x) \circ w(x) \hat{v}_0(x) = \sum_i v(x)(1 - w_i(x)). \quad (2.33)$$

步骤 4 $\hat{P}(y = j | x) = \frac{e^{\hat{v}_j(x)}}{\sum_i^N e^{\hat{v}_j(x)}}$ 。

步骤 5 计算 $y^* = \arg \max_j \hat{P}(y = j | x)$ 。

步骤 6 如果 $y^* == 0$ 或者 $\hat{P}(y = y^* | x) < \epsilon$ ，则判定 x 为未知类。

在 [20] 中，作者利用将 openmax 和生成对抗网络(GAN)相结合。作者利用预训练的GAN网络生成器生成样本。生成的样本输入openmax分类器中进行拒识。利用拒识的样本来训练openmax分类器，提升openmax分类器识别未知类的性能。

在 [81] 中，作者将第二极值理论应用于基于稀疏表示的分类算法(SRC)之中。基于稀疏表示的分类算法的原理是测试样本被同类别的样本线性表示所需要的数量较少。给定第 i 类测试样本所组成的向量 y_t 和训练数据 Y 。则 $y_t = Yx$ ，稀疏表示系数 x 中只有对应第 i 类样本的位置非零。

$$\hat{x} = \arg \min_x \|x\|_1 \quad s.t. \quad \|y_t - Yx\|_2 < \epsilon. \quad (2.34)$$

利用稀疏表示系数，可以计算 y_t 被第*i*类训练样本表示的剩余误差：

$$r_k = \|y_t - Y_k \hat{x}_k\|, \quad k = 1, \dots, K, \quad (2.35)$$

\hat{x}_k 为稀疏表示系数中属于第*k*类的子向量， Y_k 为属于第*k*类的数据样本。SRC算法将 y_t 的类别定义为

$$k^* = \arg \min_k r_k. \quad (2.36)$$

利用第二极值定理，作者估计训练集中第*i*类样本对应的剩余 r_i 以及 $\sum_{k \neq i} r_k$ 的尾分布。给定测试样本，如果对应的剩余向量 r_k 较大或者 $\sum_{k \neq i} r_k$ 较小，则测试样本被判定为未知类。

2.2.3 基于概率密度函数的异常点检测方法

机器学习中的概率生成模型利用模型概率分布拟合真实的数据分布。模型概率分布分为参数分布和非参数分布两种形式。极大似然函数是拟合过程中使用的基本目标函数。在传统方法中，常用来拟合数据分布的是GMM模型。EM算法是用来拟合GMM的常用算法。在深度学习中，深度概率生成模型主要包括深度玻尔兹曼机，变分自编码器和深度对抗网络。变分自编码器克服了当真实数据分布复杂时，混合高斯密度函数拟合性能不足的问题以及EM算法因缺乏后验概率密度函数的解析形式所引起的技术困难。概率生成模型利用对抗学习的方式来代替极大似然估计，增强了生成器生成图片的多样性和清晰度。

2.2.3.1 非参数估计

直方图密度估计和核函数估计是基本的非参数概率密度估计方法。核密度函数估计是直方图密度估计方法的光滑化。基于核密度函数估计的异常点检测方法，旨在建立关于训练样本的概率密度函数，其密度函数如下：

$$f(x) = \frac{1}{N} \cdot \sum_{i=1}^N K'_h(x - x_i), \quad (2.37)$$

其中 $K'_h(x - x_i) = \left(\frac{1}{h\sqrt{(2\pi)}}\right) \cdot \exp^{-\frac{\|x-x_i\|^2}{2h^2}}$ 。核密度函数估计需要计算所有的数据生成的高斯密度函数，因此空间复杂度较高。在[70]中，作者将核密度函数估计扩展到期望的核密度相似度估计。期望的相似度估计定义为

$$\eta(z) = E[\phi(z)] = \int k(z, x) dP(x), \quad (2.38)$$

其中k为核函数。当数据个数有限时， $\eta(z) \approx \langle \phi(z), \frac{1}{n} \sum_{i=1}^n \phi(x_i) \rangle$ 。作者利用分布式计算降低了 $\sum_{i=1}^n \phi(x_i)$ 的计算复杂度。

2.2.3.2 参数估计

深度生成模型利用基于神经网络的概率密度模型来拟合与生成数据，在近几年取得了迅速发展。深度生成模型在异常点检测领域具有广泛的应用。在 [1] 中，作者提出了基于变分自编码器的异常点检测方法并且提出了利用变分自编码器和自编码器的区别与优势。作者通过测试样本的概率密度函数值作为异常值的指标。变分自编码器和自编码器模型结构类似。定义异常值的区别如下：

- 变分自编码器计算概率密度函数值。自编码器利用重构误差。概率密度函数值具有统一的尺度，而重构误差与样本本身的尺度有关。
- 变分自编码器计算概率密度函数值通过隐藏层中采取随机抽样计算得到。自编码器的重构误差的计算通过编码器的输出来完成。在异常值的计算过程中，由于采样的多样性，变分自编码器比自编码器更加稳定。

作者通过实验说明基于变分自编码器的异常点检测模型比自编码器更具有优势。

深度玻尔兹曼机(DSEBMs)是基于势函数建模的网络。在 [80] 中，作者基于得分匹配方法来训练DSEBM，从而拟合真实的数据分布。深度玻尔兹曼机利用下面两个方法确定样本的异常值：

- 测试样本在深度玻尔兹曼机的概率密度函数值。
- 基于得分匹配训练的重构误差。

在 [16] 中，作者提出了基于DSEVM的集成方法。作者利用深度递增的概率密度网络做集成，每一个神经网络代表一种概率密度函数。作者提出的集成算法旨在使得数据在不同的抽象程度上得到更多数据结构的信息。

在 [69] 中，作者预训练GAN模型来拟合数据的概率密度函数。在测试时，作者在数据生成器的输出样本中寻找与测试样本距离最近的点。如果该样本距离原始样本的距离较远或者在判别器的特征空间中距离也较远，则该样本被判定为未知类样本。

2.3 基于重构的异常点检测方法

基于重构的异常点检测方法在训练模型时旨在减小模型输出和输入数据的重构误差。异常值的设置可以基于重构误差，基于模型的结构和训练机制。基于重构的异常点检测方法主要包含自编码器和基于子空间的方法。基于重构的方法训练模型的方式比较灵活。但是自编码器的性能对其结构的选取比较敏感，基于子空间方法的性能对子空间维数等超参数的选取比较敏感。

2.3.1 基于重构的异常点检测方法

在深度学习中，自编码器是基于重构的主要异常点检测方法。自编码器利用重构误差作为异常值的指标。重构误差越大，样本是异常数据的可能性越高。已有的工作在下面两个方面改进了自编码器的性能：

- 在训练自编码器时，利用其它目标函数来替代二范数重构误差。
- 正则化自编码器特征层的输出，调整数据在特征层的分布。

在 [38] 中，自编码器的二范数剩余误差被替换成对抗生成网络的目标函数，利用对抗训练来使得自编码器的输出和输入一致。在异常点检测中，基于对抗训练的自编码器的异常值可以通过判别器的输出来定义。当原始样本和重构的样本在判别器中的输出一致时，判别器判定为样本来自真实数据。否则，判定测试样本为异常数据。

在 [3] 中，作者利用 l_2 范数正则化编码器的输出。正则化的自编码器网络结构如下：

$$L_c = \frac{1}{|J|} \sum_{j \in J} (I_j - D(E_c(I_j)))^2, \quad E_c(I) = \frac{E(I)}{\|E(I)\|_2}. \quad (2.39)$$

在测试时，作者首先对二范数正则化的特征 $E_c(I)$ 利用 K 均值做聚类，得到类中心 C_j 。对于每一个测试数据 I_i ，利用 $E_c(I_i)$ 与类中心的夹角 v_j 作为异常值的指标：

$$v_j = \max_j (E_c(I_i) \cdot \frac{C_j}{\|C_j\|_2}). \quad (2.40)$$

v_j 越大，测试数据是异常点的可能性越高。

在 [83] 中，作者提出将 GMM 概率密度函数估计模型融合到自编码器中。作者的想法是对于高维数据的异常点检测问题，维归约是非常有效的处理方

法。简单的维归约方法可能丢失数据的部分信息并且提取一些无用的特征。自编码器虽然是有效的维归约方法，但是基于重构的目标函数会使模型过拟合，在编码器输出的特征层和重构误差的分布中不能与异常数据有效的区分。作者在自编码器的基础上增加了基于GMM混合密度函数的正则项，使得编码器特征层和重构误差向量的分布满足GMM分布。由于维归约和异常点检测方法的相互独立性影响了异常点检测的性能，作者采用端对端的方法同时训练自编码器的重构性能和GMM分布。

2.3.2 基于子空间的异常点检测方法

基于子空间的方法将数据投影到一个低维的子空间。在低维子空间中，正常数据能够和异常数据更加有效的区分。子空间的构造方法主要基于矩阵分解，非线性曲线因子分析(CCA)和随机投影方法。

基于矩阵分解的异常点检测方法主要基于主成分分析(PCA)，奇异值分解(SVD)和非负秩分解(NMF)等矩阵分解方法。作为经典的机器学习方法，矩阵分解方法具有下面的优化形式：

$$\min \|D - UV^T\|^2 \quad \text{s.t.} \quad U \text{ 和 } V \text{ 的列相正交} \quad U = V \text{ (PCA)} \quad U, V \geq 0 \text{ (NMF)}. \quad (2.41)$$

作为经典的维归约方法，PCA计算最优k维超平面，使得数据点映射到最优超平面方差最大，映射到最优超平面的正交补空间中方差最小。由于正常数据在正交补空间中的投影方差较小，这些投影会聚集到一个区域中。当数据的均值为0时，投影会集中到原点的邻域中。在异常点检测中，异常数据在正交补空间中的投影会远离正常数据所在的区域。因此，异常数据在正交补空间中投影的大小可以作为数据的异常值。距离越大，数据是异常点的可能性越大。

已有的工作主要改进了特征空间的选取和提高PCA的稳健性。核PCA首先利用核函数进行特征提取，然后在核空间中利用标准的PCA方法进行异常点检测。稳健PCA旨在减小PCA分解对数据的扰动敏感性，减小训练集中的异常点对特征空间的构造的影响。在 [82] 中，稳健PCA被扩展到稳健的张量PCA之中。在 [7] 中作者提出利用正交补空间来做异常点检测缺乏明确的解释。作者在PCA算法中加入了一个稀疏性约束来增加表示的可解释性。其优化算法如

下：

$$\arg \max_{v_i} v_i^T A v_i \quad \text{使得} \quad v_i^T v_i = 1, \quad v_i^T v_j = 0 \quad \forall 1 \leq j < i, \quad \text{Card}(v_i) \leq k. \quad (2.42)$$

2.4 基于正规区域划分的异常点检测方法

基于正规区域划分的异常点检测方法旨在建立正规数据的边界。基于正规区域划分方法刻画数据的边界，所以对数据抽样和数据的密度不敏感。给定测试样本，异常值的定义基于到边界的有向距离，有向距离越大，则该样本是异常数据的可能性越大。基于正规区域划分的方法不依赖于数据的分布。但是这些方法的计算复杂度比较高。如何有效的控制边界的大小和紧致程度是该类方法的一个难点。

异常点检测的难度高于一般分类问题的原因是数据属于同一类别，缺乏异常数据的信息来训练模型。因此，传统分类方法不能直接应用到异常点检测问题中。这类异常点检测问题的方法主要基于两个思想。第一个思想是构造属于异常数据，将单类数据边界的划分问题转换为二分类问题。基于这个思想的基本方法是单类支持向量机(OCSVM) [27]。第二个思想是利用给定形状的最小曲面覆盖正常数据。基于这个思想的基本方法是支持向量数据描述(SVDD) [8]。已有的工作基于这两个思想进行不断改进和完善。

2.4.1 基于OCSVM的单类分类器

对单类SVM的改进主要分为两个方面。第一个方面是如何生成适合的异常样本，第二个方面是如何对数据提取适当的特征。在 [18] 中，作者预训练自编码器和深度信念网络作为特征提取器。基于编码器输出和深度信念网络隐藏层的输出特征，作者利用OCSVM刻画正规数据边界。在 [10] 中，作者的基本想法是特征提取和单类分类器的相互独立性限制了异常点检测的性能。作者利用单类支持向量机的目标函数和神经网络进行端到端训练。在 [79] 中，作者改进了异常样本的生成方式。假设正样本为正常数据，负样本为异常数据，作者基于下面三个准则来构造负样本：

- 新产生的负样本与原始数据距离较近。
- 新产生的负样本与已有的负样本距离较远。

- 新产生的负样本被二分类判别器判定为负样本。

在 [62] 中，作者利用自编码器的输出作为异常样本。自编码器和二分类判别器进行类似于生成对抗网络的训练。在训练结束后，二分类判别器可以作为单类分类器。二分类判别器的输出作为异常值的判定指标。

在 [59] 中，作者的基本想法是由于数据集中只有正常类别的样本，无法通过分类任务来解决。作者的想法是在保持已知类在特征空间一个小邻域内的同时，避免神经网络的输出趋近于一点。基于这个目的，作者引入紧致损失和描述损失来分别描述训练数据在特征空间中的分布方差和与潜在未知类的距离。由于训练数据集中缺失异常数据，作者选取一些已知的公共多类数据集作为异常样本。作者利用这些负样本的标签进行分类，同时扩大与已知类数据的距离。

2.4.2 基于SVDD的单类分类器

改进SVDD的方法主要基于高斯核函数宽度的选取，超球面的形状以及在线和分布式计算等方面。在 [35] 中，作者利用基于峰值的准则来估计核宽度。在 [58] 中，作者将随机采样策略应用到该准则，从而提高核宽度估计的准确度。作者选取高斯核的宽度 s ，使得SVDD的目标函数关于高斯核宽度的二阶导数为0。

在 [13] 中作者提出高斯核宽度选取的准则。给定核矩阵 $K_N(s)$ ，第 ij 个元素为 $\exp(-\frac{\|x_i - x_j\|^2}{2s^2})$ 。当核宽度选取过小时，则核矩阵 $K_N(s)$ 会收敛到单位矩阵。这时，SVDD算法中所有的原始数据都会变成支撑向量并且导致过拟合。作者提出下面的不等式来选取核宽 s 防止过拟合：

$$\|K_N(s) - I_N\| \geq \delta \|I_N\|. \quad (2.43)$$

在 [31] 中，作者提出了SVDD的在线增量形式。作者的算法基于特种空间中的支持向量到高维球中心具有相同的距离的事实。当新数据点出现时，只有支撑向量和新数据点参与到迭代更新高维球中。新数据被判定为高维球空间的内点时，则停止计算。否则，根据递推公式进一步更新高维球的中心和半径。

2.4.3 基于单类分类器的开放集识别问题

在 [5] 中，作者提出使得SVM具有开放集识别的能力的模型，叫作1-vs-Set学习机。作者指出由于SVM分类器在特征空间中的分类边界是一条直线，

所以类别空间分别是半空间。当数据点距离训练数据集较远时，数据点是异常点的可能性增加。作者利用平行于分类边界的超平面来缩小正类的空间。在1-vs-Set学习机中，正类区域是分类边界和构造超平面之间的带状区域。带状区域的宽度由事先给定的风险函数所决定。

2.5 基于集成的异常点检测方法

集成分析是提升数据挖掘算法性能的重要方法。集成方法的基本思想是不同的算法在相同的数据集上表现不同。通过集成模型，可以发挥每一个模型的优势。集成算法可以减少模型预测的方差，从而增强模型预测的稳定性。集成算法的步骤分为两个。第一个是基模型的选取，第二个是模型评分的正则化和结合。集成算法可以分成两种分类。第一种分类是基于模型的集成与基于数据集的集成。基于模型的集成旨在利用不同的模型或者算法进行集成，包括不同的参数选取与不同的初始化。基于数据集的集成表现在相同的模型在不同的数据集进行训练。集成方法也可以分成独立集成和顺序集成的方式。对于相互独立的集成，不同的基模型在训练过程中互不影响。模型的评分会以某种方式进行合并。对于顺序集成，基模型会顺序的进行集成，后一个模型会改进前面基模型的输出结果，最后逐步提高模型的准确度。对于异常点检测方法，基于顺序集成方式的方法比较少，主要采用独立集成的方式。

在 [39] 中，作者提出了基于分割的方法，叫作隔离森林(Isolation Forest)。该方法的基本思想是孤立点与正常数据分布不同，可以通过更少的超平面将孤立点同正常数据分离。当数据集的结构比较复杂时，将给定点与其余数据隔离所需超平面的最小个数无法精确计算。作者采用随机超平面和集成的方法来降低计算超平面最小个数的复杂度。隔离森林表示逐步分割数据集直至隔离数据集中所有点的超平面序列。隔离森林的构造算法如下：

算法 2.3. 输入： 数据集 S 。

步骤 1 随机选取数据的一个属性 q 。

步骤 2 计算数据集在属性 q 的最大值 \max 和最小值 \min ，随机选取在最大值和最小值之间的分割 p 。

步骤 3 将数据集中的点通过分割划分成两部分，直至数据集中只有一个点。

在近似计算所需要的最小超平面时，只需要计算隔离森林隔离该点时所需要的随机分割 p 的个数。由于分割的随机性，隔离森林算法采用独立集成的方法。随机生成若干个隔离森林，分割的个数去每一个隔离森林的平均值。

2.6 基于信息论的异常点检测方法

基于信息论的方法通过熵和相对熵等统计方法计算数据集中所含有的信息量。这类方法假设异常点会显著改变正常数据的信息量。基于信息论的方法对数据的分布没有要求，但是它需要一个足够敏感的测度来度量异常点对数据集信息量的影响。通常的信息论统计量只能度量出数据集中的异常点比例不是很稀少的情形。

在 [41]中，作者提出新的信息论度量，记作敏感度(sensitivity)。敏感度定义为对数据分类目标函数的最坏影响。计算敏感度具有较高的计算复杂度，作者计算了敏感度的上界，记作影响(influence)。作者说明估计影响只需要线性时间。

2.7 异常点检测方法的比较方法举例

异常点检测方法的种类较多，不同类别的方法擅长检测的异常点各异。在比较试验中，只要说明在给定条件下，与已有方法的优势即可。下面，我们列举两个比较典型的实验例子，一个来自于machine learning期刊的文献 [70]，一个来自于ICLR2018的文献 [84]。

文献 [70]主要提出基于LOF的改进异常点检测方法。文章主要选取的比对方法是baseline LOF来展示提升的性能，同时选取IFOREST以及KDE作为比对方法。KDE是基于概率密度函数的经典方法。文章的数据集均为实验数据集。包含若干个UCI数据集作为小数据集以及MNIST和SVHN数据集作为大数据集。所有的实验均选取实验数据集中的其中一类为正常数据，将数据集中的其它类别作为异常数据。由于文章针对的是传统方法，作者对MNIST和SVHN数据集提取hog特征和适当的数据增广。

文献 [84]是ICLR2018审稿意见评分较高的异常点检测方法。在审稿意见中，评分较高的原因是实验比对方法完善，具有说服力。在 [84]中，作者提出基于自编码器的异常点检测方法的改进版本。作者将重构误差和编码器的输出进行特征融合，并且利用GMM模型将融合特征拟合正则化为混合高斯密度函

数。作者采用的数据集是KDDCUP, Thyroid以及Arrhythmia。它的比较方法主要分为不同类别方法性能的比较以及基于特征融合和高斯混合密度函数分布的创新方法的优越性。对不同类别的方法，文章主要选取了OCSVM单类分类器，基于深度生成网络的玻尔兹曼机模型DSEBM-e, DSEBM-r以及基于聚类的深度学习DCN方法。针对创新性能优越性的实验，作者设计的比对方法如下：

- GMM-EN：去掉重构误差，只要求encoder输出满足混合高斯密度函数。
- PAE：去掉混合高斯密度函数拟合的约束。这时，模型退化为标准的深度自编码器。这个方法利用预训练来训练自编码器。
- E2E-AE：深度自编码器的端到端训练。
- PAE-GMM-EM：首先训练自编码器，利用EM算法拟合训练GMM。
- PAE-GMM：将深度自编码器和GMM拟合分别训练。

这些比对方法说明特征融合以及高斯混合密度函数拟合的必要性。同时，也说明了在GMM拟合过程中，方法选用的合理性。因此，审稿意见认为文章比较方法可以充分论证方法中每一个环节的有效性。

作为总结，异常点检测方法因环境的差异，方法种类的多样性和问题的高难度，比较实验中没有特定的比较方法和标准实验数据集。异常点检测的文章中，作者叙述自己的实验结果时，很少说明本文章的方法具有state of the art性能。因此，比较实验需要大量复现与文章提出的方法相类似的其它异常点检测方法。只要能说明在某些实验条件下，提出的方法比其它的方法具有更好的性能即可。

第三章 基于生成对抗网络的单类分类器

单类分类器是一类重要的异常点检测方法。它的典型方法是包含OCSVM和SVDD。OCSVM方法的基本思想是生成异常样本作为负样本，将单类分类问题转换成二分类问题。作为第一个基于异常样本生成的方法，OCSVM在特征空间的原点当成唯一的异常样本，进行二分类训练。由于异常样本的个数较少，与正常数据的距离无法衡量，二分类的分类边界不能有效的逼近真实的数据边界。SVDD的基本思想是利用最小体积的球体来包含特征空间的已知类区域。在实际应用中，已知类别分布区域的形状不规则，利用固定形状的球体不能高质量的刻画数据边界。已有的OCSVM改进方法是静态的生成异常样本，虽然提高了OVSVM的性能，但是有限的静态样本不能从实质上提高正常数据边界刻画的性能。

为了克服传统单类分类器的缺点，我们提出基于动态样本生成来识别异常数据的单类分类器。我们利用逐渐接近真实数据的动态样本来减小样本数量不足所造成的分类误差并且克服分类边界与真实数据距离较大的缺点。作为生成真实数据的深度生成模型，VAE和GAN在训练的过程中可以产生收敛到真实数据的大量样本。因此，我们的方法基于GAN的生成器输出的动态样本来训练判别器，利用判别器的输出来刻画真实数据的边界。由于GAN中的生成器输出分布收敛到真实的数据分布，随着训练的进行，判别器识别未知类的能力会随着生成器输出与真实样本过于接近而下降。因此，我们利用KL正则散度来在保持生成器输出与真实样本分布在一定距离范围内的前提下阻止生成器的输出收敛到真实的数据分布。

3.1 生成对抗网络

在这一节，我们主要介绍基于GAN的单类分类器及其改进。作为深度生成模型，GAN [21]在深度学习的许多领域具有广泛应用。GAN的基本思想是训练一个生成器 G 和一个判别器 D ，使得 D 判定样本是否来自真实数据， G 生成使得 D 无法区分是否来自真实数据的样本。GAN 的目标函数如下：

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}(x)} \log D(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D(G(z))). \quad (3.1)$$

特征匹配 [64] 是生成器目标函数的一个变体，它可以防止生成器 G 过拟合。令 $f(x)$ 为判别器中间层激活函数的输出。则 G 的目标函数定义为：

$$\|\mathbb{E}_{x \sim p_{data}} f(x) - \mathbb{E}_{z \sim p(z)} f(G(z))\|. \quad (3.2)$$

在介绍我们的工作之前，我们说明利用GAN的判别器作为单类分类器的合理性。在GAN训练的初始过程中，生成器输出的样本与随机样本接近。这些样本定义为弱异常数据。在训练的过程中，生成器输出的分布收敛到真实的数据分布。判别器在训练过程中在正常数据中取得较高的值，在生成器的输出中取得较低的值。通过判别其输出所决定的正规数据边界随着训练而逐渐减小并且接近于正常数据的真实边界。

尽管判别器可以检测异常样本，但是仍然有不足之处。由于判别器输出的分布 p_G 逐渐收敛到真实的数据分布 p_{data} ，判别器最终收敛到 $\frac{1}{2}$ 。因此，判别器 D 识别未知类的性能在训练的过程中逐渐下降。

基于上述考虑，我们提出关于生成器的正则化目标函数，使得生成器 G 输出的分布 p_G 无法收敛到真实的数据分布，进而生成强异常数据。与此同时，我们设计一个集成算法来克服判别器性能的不稳定性。

3.2 极小似然对抗生成网络

为了防止在训练过程中判别器的退化和提高判别器的识别能力，我们正则化 G ，使得

- G 在训练中产生更多的异常数据。
- G 的输出分布 p_G 不收敛到 p_{data} 。

为了达到这个目的，我们提出KL散度来阻止 p_G 收敛到 p_{data} 。给定先验分布 $p(z)$ 和 $z \sim p(z)$ 。由于 p_G 是生成器输出的分布， p_G 的支撑是高维空间中的子流形。这时， $KL(p_{data} \| p_G)$ 的定义无意义。因此，我们定义 $x : G(z) + n$ ，这里 n 与随机变量 z 相互独立。 n 的分布是高斯分布或者拉普拉斯分布。定义 \tilde{p}_G 为 x 的分布。对任意的 x 与 z ，下面的不等式成立：

$$p(x|z) > 0, \quad p(z) > 0. \quad (3.3)$$

因此，对任意的 x ，

$$\tilde{p}_G(x) = \int p(x|z)p(z)dz > 0. \quad (3.4)$$

\tilde{p}_G 的支撑属于整个全空间。此外，当n适当选定时， $\tilde{p}_G(x) \approx p_G$ 。在我们的方法中， G 的目标函数定义如下：

$$\|\mathbb{E}_{x \sim p_{data}} f(x) - \mathbb{E}_{z \sim p(z)} f(G(z))\| - aKL(p_{data} \| \tilde{p}_G). \quad (3.5)$$

极小化 $-KL(p_{data} \| \tilde{p}_G)$ 等价于降低正常数据的概率密度函数值 $\tilde{p}_G(x)$ 。基于KL散度正则化的GAN为极小似然GAN(MinLGAN)。

由于 $KL(p_{data} \| \tilde{p}_G) = \int p_{data} \log p_{data} - \int p_{data} \log \tilde{p}_G$ 并且 \tilde{p}_G 没有闭合形式，直接计算 $KL(p_{data} \| \tilde{p}_G)$ 无法实现。因此，我们利用变分推断将 $\log \tilde{p}_G$ 替换成 $\max_{\vartheta} \mathcal{L}(x, \theta, q(z|x, \vartheta))$ ，其中

$$\mathcal{L}(x, \theta, q) := \log p_{data}(x|\theta) - D_{KL}(q(z|x) \| p(z|\theta)) \quad (3.6)$$

$$= \int q(z|x) \log p(x|z, \theta) - D_{KL}(q(z|x) \| p(z)). \quad (3.7)$$

$KL(p_{data} \| \tilde{p}_G)$ 具有下面的近似表达式：

$$\int p_{data} \log p_{data} - \int p_{data} \max_{\vartheta} \mathcal{L}(x, \theta, q(z|x, \vartheta)). \quad (3.8)$$

我们的方法包括了迭代优化判别器D， $q(z|x, \vartheta)$ 和生成器G的参数。判别器D的目标函数是

$$\max_D \mathbb{E}_{x \sim p_{data}} \log D(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D(G(z))). \quad (3.9)$$

生成器G的目标函数是

$$\min_G \|\mathbb{E}_{x \sim p_{data}} f(x) - \mathbb{E}_{z \sim p(z)} f(G(z))\| + a \int q(z|x, \vartheta) \log p(x|z, \theta). \quad (3.10)$$

$q(z|x, \vartheta)$ 的目标函数是

$$\max_{\vartheta} \int q(z|x, \vartheta) \log p(x|z, \theta) - D_{KL}(q(z|x, \vartheta) \| p(z)). \quad (3.11)$$

当 $q(z|x, \vartheta)$ 和 $p(x|Z)$ 为高斯或者拉普拉斯分布时，变分推断具有显著的几何解释。在 $q(z|x, \vartheta)$ 的目标函数中，极大化 $\int q(z|x, \vartheta) \log p(x|z, \theta)$ 意味着 $q(z|x, \vartheta)$ 输出 z ，使得 $G(z)$ 与 x 接近。在 G 的目标函数中，极小化

$$\int q(z|x, \vartheta) \log p(x|z, \theta) \quad (3.12)$$

意味着 $G(z)$ 与 x 的距离增加。因此，KL正则化阻止 G 收敛到真实的数据分布。

3.3 克服判别器的不稳定性

判别器D的性能依赖于在训练过程中G产生的异常点序列。由于GAN训练的随机性，异常点序列也会具有随机性，例如随机初始的权值，先验分布的随机采样。这些随机性会引起判别器性能的不稳定性。

集成学习是处理随机性的有效方法。集成学习包含多个基学习器来减小偏差与方差。两种常见的集成方法是bagging法和boosting法。Bagging法主要基于多个模型在数据集的不同子集进行训练，总得分依赖于各个子模型的得分。Boosting法旨在通过模型序列逐步提高总体模型的判别能力。

与bagging法相似，我们独立训练 N 个判别器。给定测试样本 x ，计算 $D_i(x)$ 。样本 x 的异常值 s 的定义主要分为两种方式。第一个方式叫做集成GAN，其形式如下：

$$s = -\frac{1}{N} \sum D_i(x). \quad (3.13)$$

当验证集 S 可以利用时，令 $m_i := \max_{x \in S} D_i(x)$ 和 $n_i := \min_{x \in S} D_i(x)$ 。第二个方式叫做尺度化集成GAN，其定义形式如下：

$$s = -\frac{1}{N} \sum (D_i(x) - n_i)/(m_i - n_i). \quad (3.14)$$

3.4 实验结果

在这一节，我们首先在circle和moon数据集上可视化KL散度的性能。同时，我们利用CIFAR10和UCI数据集的实验结果说明我们的方法比其他的方法具有更好的识别性能。

3.4.1 玩具数据集的可视化

我们选取circle和moon数据集来可视化KL散度正则化在生成器 G 的作用。在图片3.1(a)和图片3.1(c)中，绝大多数蓝点分布在正常数据的流形中。实验结果说明生成器 G 的输出分布与正常数据分布几乎相同。在表格3.1(b)与3.1(d)中，许多蓝点分布在正常流形的边界之外。由于KL散度系数相对较小，大多数蓝点分布在正常数据的流形区域内。这两个实验说明KL散度使得生成器 G 产生正常数据流形的邻域内的数据并且阻止 G 收敛到正常的数据流形。

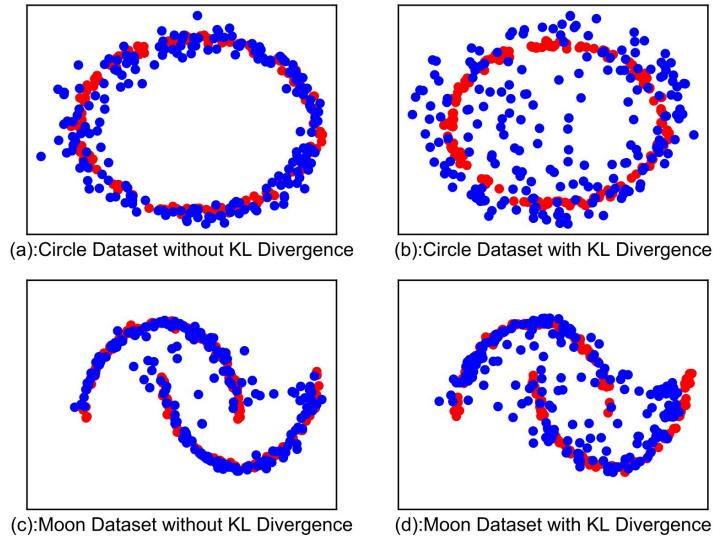


图 3.1: 生成器 G 在KL散度正则化下的性能对比。红色的点表示正常数据的流形，蓝色的点表示 G 的输出分布。图(a)和图(c)表示在没有KL散度正则化下，生成器 G 输出的分布。图(b)和图(d)表示在KL散度正则化下生成器 G 输出的分布。

3.4.2 公开数据集的实验结果

在这些实验中，训练集只包含正常数据。测试集中既包含正常数据，又包含异常数据。我们的方法包括极小似然GAN(MinLGAN)，集成MinLGAN(EMinLGAN-1) 和尺度化集成MinLGAN (EMinLGAN-2)。我们的对比方法包括GAN的基本方法(GAN)，OCSVM，IFOREST，VAE，AE。我们的方法基于Theano，我们的代码基于<https://github.com/openai/improved-gan>。OCSVM，IFOREST由LIBSVM软件包 [11]实施。异常值定义如下：

- GAN的异常值定义为判别器输出的相反数。
- OCSVM的异常值定义为到决策边界的有向距离。
- IFOREST的异常值定义为隔离一个样本所需超平面的个数。
- AE的异常值定义为样本重构误差。

表 3.1: Cifar10数据集的AUC值.

Normal	EMinLGAN-1	EMinLGAN-2	MinLGAN	GAN	IFOREST	OCSVM	VAE	AE
0	0.814	0.821	0.786	0.76	0.615	0.689	0.645	0.739
1	0.633	0.642	0.61	0.627	0.688	0.464	0.519	0.358
2	0.660	0.664	0.643	0.635	0.476	0.679	0.638	0.692
3	0.568	0.585	0.567	0.589	0.538	0.513	0.539	0.575
4	0.702	0.701	0.676	0.664	0.661	0.767	0.771	0.774
5	0.643	0.672	0.621	0.6	0.607	0.529	0.505	0.59
6	0.732	0.721	0.697	0.706	0.757	0.765	0.715	0.699
7	0.623	0.62	0.599	0.565	0.659	0.53	0.506	0.515
8	0.771	0.788	0.755	0.715	0.7	0.706	0.73	0.792
9	0.639	0.652	0.616	0.604	0.711	0.481	0.605	0.42
平均	0.679	0.687	0.657	0.647	0.641	0.613	0.617	0.615

- VAE的异常值定义为样本的重构概率。

在实验中，所有的异常点被定义为正类，正规数据被定义为负类。ROC曲线指标用来衡量算法的性能。不同方法的性能对比基于AUC值。

- ROC曲线：刻画在不同的阈值下，正确识别的正样本比例(TPR)和错误识别的正样本比例(FPR)。
- AUC(ROC)值：ROC曲线下方区域的面积。

3.4.2.1 Cifar10的实验结果

Cifar10数据集包含60000张尺寸为 32×32 的图片，包括50000张训练图片和10000张测试图片。在这里，我们展示10个子实验的结果。在每一个子实验中，10类中的一类被看作是正常数据，其余类别的数据是异常数据。所有的实验中共享相同的结构，学习率和正则化系数 a 。一个小的验证集确定方法的收敛性。我们对每一个子实验重复80次并且在每一次记录这些验证集中的最好结果。平均的AUC值在表3.1中展示。

表3.1说明在子实验0, 2, 4, 5, 7, 8, 9中，MinLGAN的性能比GAN高。在子实验1, 3, 6中，MinLGAN的性能比GAN低。这是因为我们所使用的学习率和正则化参数 a 在所有子实验中不变。尽管KL正则散度有助于产生更多的异常数据，当 a 选取过大时，GAN的动力会遭到破坏。对于子实验1和3，MinLGAN和GAN的性能都会下降。这是因为GAN的性能依赖于神经网络结构

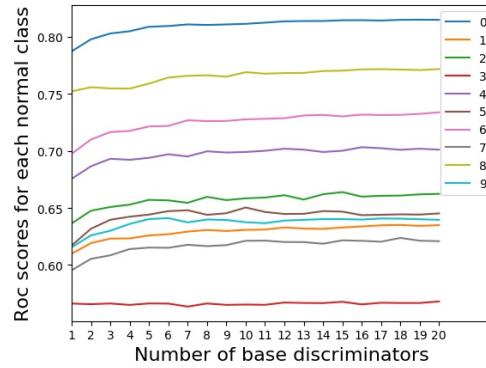


图 3.2: Cifar10数据集集成的AUC值. 当基判别器的个数增多时, 集成模型的AUC值逐渐增加。当集成个数达到10时, 集成的性能达到稳定。

的选取。在这两个子实验中, 适当修改神经网络结构能够提高GAN的识别性能。由于GAN训练的不稳定性, 在实验中会出现某一个结果明显比平均性能低, 这些坏的性能结果也包含在我们的表格中。

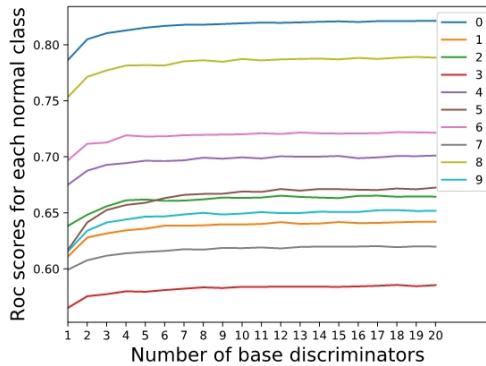


图 3.3: Cifar10数据集集成的AUC值. 当基判别器的个数增多时, 集成模型的AUC值逐渐增加。当集成个数达到5时, 集成的性能达到稳定。

图片3.2和图片3.3表明AUC值与基判别器个数的关系。实验结果表明对EMinLGAN-2, 当基判别器的个数超过5时, 集成后的AUC值会稳定。对于EMinLGAN-1, 使得AUC值稳定所需的基判别器最小个数是10。EMinLGAN-2的集成收敛速度高于EMinLGAN-1。当两种集成方式均收敛时, EMinLGAN-2的性能要优于EMinLGAN-1。这是因为对于每一个基判别器, 异常值的量纲

不一致。由EMinLGAN-1产生的异常值严重依赖于异常值量纲较大的基分类器的个数。

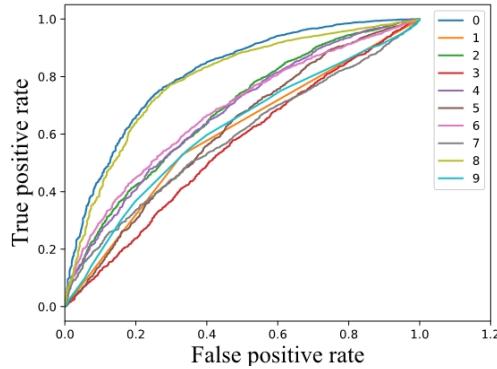


图 3.4: Cifar10数据集的ROC曲线.

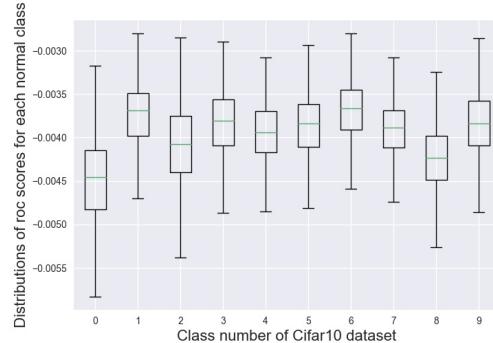


图 3.5: Cifar10数据集的异常值分布. 箱形图的顶端和低端分别代表第一和第三中位数。箱形图中的绿色线代表中位数。

图3.4描绘的是所有子实验的ROC曲线。图3.5是在子实验0中每一类别异常值分布的箱形图。类别2和8异常值分布在一定程度上与类别0重合。在我们的方法中，模型对不同未知类的识别能力是不同的。

3.4.2.2 UCI数据集的实验结果

我们选取一些UCI数据集来进一步展示我们方法的性能。这些数据集包括KDDCUP99, cover type和shuttle。KDD数据集包括5个主要类别。只有属于Normal类别的数据作为正常数据。其它类别的数据是异常数据。Shuttle数

表 3.2: UCI数据集的实验设置

名称	数据集	正常数据	异常数据
KDD-A	KDD	normal	attack
COV-A	cover type	class 1,3,5,6,7	2,4
COV-B	cover type	class 2	class 4
SHU-A	shuttle	class 1	class 2,3,4,5,6,7

据集包括9个属性，所有的属性都是数值型。类别1的样本在数据集中所占的比例是80%。Cover type数据集通过类别变量预测了7个森林覆盖类型。每一个数据包括54维属性。对于UCI实验，我们在表3.2中列出正常数据和异常数据的情况。验证集帮助各个方法的调参。我们采样80%个正常数据作为训练数据。其他的正常数据和异常数据被用作测试集。

在表3.3中，可以看出我们的方法在所有的数据集中取得了比较好的性能。对于KDD数据集，GAN的基本模型与EMinLGAN和EMinLGAN-1取得了相似的性能。OCSVM和IFOREST 方法除了COV-A外都取得了较好的性能。这是因为正常数据由若干个类别构成并且没有一个有效的特征提取方法。VAE和AE对所有的数据集都采用相同的网络结构。在我们的实验中，除了数据集SHU-A之外，重构概率没有重构误差稳定。

表 3.3: UCI数据集的AUC值

名称	KDD-A	COV-A	COV-B	SHU-A
EMinLGAN-1	0.993	0.811	0.975	0.988
MinLGAN	0.993	0.798	0.945	0.986
GAN	0.993	0.793	0.931	0.979
IFOREST	0.991	0.293	0.991	0.988
OCSVM	0.982	0.397	0.997	0.947
VAE	0.995	0.743	0.956	0.802
AE	0.937	0.735	0.998	0.978

第四章 基于对抗训练的开放集分类方法

分类问题是机器学习领域最基本的问题之一。经典的分类问题旨在利用属于若干个类别的数据训练分类器，从而使分类器在属于已知类别的测试数据中区分数据的类别。开放集识别问题是指分类器在区分已知类别的数据同时区分其他类别的数据。在开放集分类问题中，分类器需要分类和识别双重任务。与传统的分类问题相比，开放集分类问题的难点在于训练集中缺乏未知类别信息。与传统的异常点检测相比，开放集分类问题需要考虑识别未知类的机制与分类机制的相容性。因此，开放集分类问题与传统的分类问题和异常点检测任务联系紧密，但是具有更高的技术难度。

开放集分类问题的解决方法主要分两种。第一种是在已知类数据训练的分类器中设置未知类检测指标。这种方法对分类器训练无影响，对分类性能影响较小。由于分类器的特征空间的训练主要依靠分类目标函数，这种方法识别未知类的性能有限。第二种方法是正则化分类器的特征空间，从而增强分类器在某种指标下对未知类的识别能力。第二种方法的技术难度较高，比较典型的方法是利用对抗机器学习的方法来增强分类器的光滑性，从而提高分类器的性能。

在这一章，我们介绍对抗机器学习的代表性方法以及在开放集分类任务上的应用。此外，我们介绍在开放集分类任务上的研究进展。我们的实验说明神经网络可能把一些与原始样本非常不同的数据映射到特征空间的已知类区域中。因此，神经网络可能将一些未知类样本错误的判定为已知类。我们设计生成器来生成满足特定要求的样本。这些样本在原始空间和真实数据不同，但是在特征空间与真实数据相似。因此，我们利用特征匹配的目标函数来拟合真实数据在特征空间中的分布。利用二分类神经网络来区分真实样本和生成器输出的样本。实验结果说明我们的方法对增强分类器识别未知类的性能是有效的。

4.1 对抗机器学习：生成对抗样本

神经网络以强大的特征提取能力在应用领域中展现了优越的性能。与传统机器学习模型类似，神经网络在取得高性能的同时容易受到对抗样本的攻击。

换句话说，一些与原始样本相似的样本被神经网络分类器进行错误的识别，这些样本在对抗机器学习中叫做对抗样本。在近些年，神经网络的对抗样本问题主要分为攻击和防御两个方面。攻击者利用已知的环境设计神经网络的对抗样本生成方法。根据环境的不同，攻击者的攻击主要分为白箱攻击和黑箱攻击。白箱攻击假设攻击者已知神经网络的所有信息，包括参数值和网络结构。攻击者可以利用神经网络的已知信息生成对抗样本。黑箱攻击假设攻击者没有神经网络的任何信息，但是知道所生成的样本是否为神经网络的对抗样本。防御者的目的是探测对抗样本和减少攻击者产生对抗样本的可能性。因此，防御者对神经网络进行对抗训练和并且对神经网络建立关于对抗样本的探测机制。防御方法主要根据是否在神经网络中嵌入新的子网络进行划分。

对抗机器学习和异常点检测方法具有密切的联系。异常点检测方法通过检测特征空间的异常数据来辅助防御者探测对抗样本。对抗机器学习可以增强网络的光滑性和泛化能力，有助于将正常样本的特征区域变得更加紧致。

在 [22] 中，作者论证了神经网络中对抗样本的存在性。作者在文章中解释了神经网络对抗样本存在的原因。作者指出对抗样本的存在性与神经网络的局部线性性质关系密切。给定真实样本 x ，令 $\tilde{x} = x + \eta$ ，其中 $\eta < \epsilon$ 。由

$$f(\tilde{x}) \approx f(x) + \nabla f(x)^T \eta \quad (4.1)$$

可知，当 $\eta = \nabla f(x)$ 时， $|f(\tilde{x}) - f(x)|$ 较大。因此，作者提出针对神经网络的对抗样本生成方法：

$$\eta = \epsilon \text{sign}(\nabla_x J(\theta, x, y)). \quad (4.2)$$

基于文献 [22]，已有的工作从多种角度改进对抗样本的生成方式。下面，我们介绍一些重要的对抗样本生成方法。

箱形约束的L-BFGS方法：在 [74] 中，作者提出基于箱形约束的白箱攻击方法(L-BFGS)。L-BFGS方法不仅可以产生使得分类器错误分类的对抗样本，而且可以指定对抗样本的错误分类。L-BFGS 方法的目标函数如下：

$$\min_{\rho} \|\rho\|_2, \quad s.t. \quad C(I_c + \rho) = l; \quad I_c + \rho \in [0, 1]^m, \quad (4.3)$$

其中， C 为目标分类器， I_c 为原始图片， l 为将对抗样本错分的类别。由于优化问题含有分类器 C ，L-BFGS方法需要利用分类器内部的结构和参数。因此，

L-BFGS方法是白箱攻击方法。利用L-BFGS方法产生的对抗样本比其他的方法更容易识别。

快速梯度符号方法(FGSM): 在 [22]中，作者提出基于分类器梯度的攻击方法。与L-BFGS的对抗样本构造方法不同，FGSM方法通过梯度来显式构造对抗样本。对抗样本的形式如下：

$$\rho = \epsilon \text{sign}(\nabla J(\theta, I_c, l)), \quad \rho = \epsilon \frac{\nabla J(\theta, I_c, l)}{\|\nabla J(\theta, I_c, l)\|_2}, \quad (4.4)$$

其中， J 为目标分类器， I_c 为原始图片， l 为将对抗样本错分的类别。由于异常样本的构造需要分类器的梯度，FGSM方法属于白箱攻击方法。

基本迭代法(BIM): 在 [37]中，作者指出基于单步扰动图片的方法所产生的对抗样本性能较弱。作者提出基于迭代的多步方法产生对抗样本，从而提升了对抗样本的攻击性能。BIM方法的迭代公式如下：

$$I_\rho^{i+1} = \text{Clip}_\epsilon\{I_\rho^i + \alpha \text{sign}(\nabla J(\theta, I_c, l))\}, \quad (4.5)$$

其中 I_ρ^i 表示第*i*次被扰动的图片， $\text{Clip}_\epsilon\{\cdot\}$ 表示切割图片的尺寸函数。

基于雅可比的显著映射攻击(JSMA): 在 [56]中，作者提出基于 l_0 的稀疏扰动攻击方法。作者每一次扰动一个像素并且查看像素对分类结果的影响，从而通过扰动样本的少量像素来生成对抗样本。像素对神经网络分类结果的影响主要通过输出的梯度构造的显著性映射来构造。

单像素攻击(One-pixel): 在 [73]中，作者通过改变单个像素生成对抗样本的方法。与基于梯度的方法不同的是，作者通过进化算法随机改变样本的一个像素来生成子一代扰动。最后一代产生的样本被用作对抗样本的扰动方向。

Carlini和Wagner攻击(CW): 在 [9]中，作者提出能渗透对抗蒸馏策略的对抗扰动。CW攻击可以分别产生关于 l_0, l_1, l_2 范数的小扰动。

DeepFool攻击: 在 [49]中，作者提出基于线性规划的迭代方法生成对抗样本。在基于线性规划的对抗样本生成方法中，作者将其分类边界的区域线性化，从而将一般的优化问题转化成线性规划问题。通过求解线性优化问题的扰动累加到原始图像上。作者说明DeepFool攻击在产生相同对抗效果的样本所需要的扰动比其他的攻击产生的扰动更小。

经典的对抗样本生成方法对不同的原始图片产生不同的对抗扰动。在 [48]中，作者提出一个新的对抗扰动生成方法，使得该对抗扰动对数据集中的

所有图片产生对抗样本，叫做一致对抗扰动。一致对抗扰动具有严格的数学定义。给定原始图片的扰动 ρ ， ρ 是一致对抗扰动当且仅当下面的条件成立：

$$P(C(I_c) \neq C(I_c + \rho)) \geq \theta, \text{ s.t. } \|\rho\|_p \leq \xi, \quad (4.6)$$

其中 I_c 是原始图片， $C(I_c)$ 是分类器对原始图片的分类， $\theta \in (0, 1]$ 和 ξ 是给定的常数。

在 [66]中，作者提出两种黑箱攻击算法UPSET和ANGRI。这两种方法生成对抗样本的网络结构相似。一个微小的差别是UPSET的攻击将样本错分为某一类别，ANGRI 攻击将样本错分即可。与其它攻击方法不同的是，UPSET和ANGRI算法产生的对抗样本可以同时攻击多个分类器。

Houdini攻击：在 [14]中，作者提出对训练样本加权的方法来提升一致对抗扰动对抗样本的效果。令 $g_\theta(x, y)$ 为分类器输出的第*i*个分量， $y_\theta(x) = \arg \max_y g_\theta(x, y)$ ， $l(x)$ 为训练目标损失函数。作者假设对抗样本的生成方式如下：

$$\tilde{x} = \arg \max l(y_\theta(\tilde{x}), y), \text{ s.t. } \tilde{x} : \|\tilde{x} - x\| \leq \epsilon. \quad (4.7)$$

Houdini攻击基于给定的对抗样本生成方法，它可以进一步提升该方法的攻击效果。该算法旨在考虑模型对原始训练样本的预测误差。如果训练样本的预测误差较大，则该方法对训练样本设置较小的权值。Houdini攻击的目标函数如下：

$$\bar{l}_H(\theta, x, y) = P_{\gamma \sim N(0, 1)}[g_\theta(x, y) - g_\theta(x, \hat{y}) < \gamma] \cdot l(\hat{y}, y), \quad (4.8)$$

\hat{y} 为数据的真实类别。

基于对抗变换网络的攻击：经典的对抗样本生成算法主要基于模型的梯度来显式构造对抗样本。在 [4]中，作者利用对抗变换网络生成对抗样本。假设对抗变换网络为

$$g_{f,\theta}(x) : x \in X \rightarrow x'. \quad (4.9)$$

训练对抗变换网络的目标函数如下：

$$\arg \max \sum_{x_i \in X} \beta L_x(g_{f,\theta}(x_i), x_i) + L_y(f(g_{f,\theta}(x_i)), f(x_i)). \quad (4.10)$$

该目标函数的第一项旨在减小对抗样本和原始样本之间的距离，第二项旨在增加生成样本和原始样本在分类器输出的距离。

对抗样本的防御主要分为三种类别：

- 修改训练目标函数以及输入样本。
- 修改网络结构。
- 添加新的网络。

修改分类器训练目标函数的防御方法，旨在对分类器的训练目标函数添加正则项，从而提高分类器抵御对抗样本的性能。这类方法又叫作暴力对抗训练。在 [22] 中，作者提出基于抵御FGSM产生的对抗样本的方法。作者利用由FGSM产生的对抗样本训练分类器，使得对抗样本与原样本的输出类别一致。在 [65] 中，作者提出在神经网络不同特征层中扰动来产生对抗样本的训练方法。作者将神经网络的每一层分别作为输出层来求解原始图像在该层输出的对抗样本，利用这些对抗样本进行暴力对抗训练，从而提升模型对对抗样本的防御性能。半监督学习是机器学习的重要分支，对抗训练不仅可以增加分类器抵御对抗样本的性能，还能作为半监督学习的目标函数，使得分类器在少量标记样本的情形仍然能够取得优越的分类性能。在 [47] 中，作者提出虚拟对抗训练(VAT)的正则化函数。虚拟对抗函数旨在光滑化神经网络的输出分布。基于虚拟对抗训练正则化的半监督算法在半监督算法中具有最优的分类性能。

数据压缩技术通过改变输入样本来提高分类器防御对抗样本的性能。在 [17] 中，作者指出 JPG 图像的压缩可以抵御 FGSM 攻击和 DeepFool 攻击产生的对抗样本。在 [52] 中，作者指出离散余弦变换(DCT)可以增强分类器抵御对抗样本的能力。在 [42] 中，作者提出基于 PCA 压缩的防御方法。作者同时提出如果数据压缩技术使用不当，会导致输入数据失去重要的分类信息，从而影响分类性能。

一些基于修改网络结构的防御方法已经被提出。在 [24] 中，作者提出利用深度收缩网络(DCN)来提高神经网络对对抗样本的稳健性。作者在训练分类器的目标函数中加入了模型雅可比矩阵范数的正则化。由于对抗样本距离原始样本距离很小，所以对抗样本的存在性与雅可比矩阵的性质关系密切。作者通过极小化雅可比矩阵的二范数来降低分类器对对抗样本的敏感性。作者指出，在去噪自编码器(denoising Autoencoder)中，雅可比矩阵的正则化等价于对原始数据随机扰动的数据增广。这一观点应用到分类器防御对抗样本的问题之中。

在 [51] 中，作者提出添加新的神经网络(DLN)作为输入图片的转换器。当输入图片为对抗样本时，转换器可以输出一个样本，使得该样本可以被分类

器正确分类。假设原始样本为 x , 对抗样本为 x' , DLN网络记作 D , 分类器记作 C 。DLN网络的目标函数如下:

$$a\overline{sim}(x, D(x')) + \overline{opsim}(\text{Cat}(y_x), C_p(D(x'))), \quad (4.11)$$

其中 \overline{sim} 表示相似度, \overline{opsim} 表示距离。

数据蒸馏旨在将复杂大型网络的知识和信息转移到小规模并且结构简单的网络之中。在知识蒸馏的训练过程中, 复杂网络的输出作为知识训练小规模网络。在 [57]中, 作者指出知识蒸馏的过程可以增加小网络对原始图片小扰动的抵御能力。由于数据蒸馏会对小网络产生不稳定性, 一些稳定化数据蒸馏的方法在 [55]中被提出。尽管数据蒸馏方法可以有效避免FGSM等基本攻击方法, 它不能防御CW攻击。

在 [36] [50]中, 作者论证了神经网络的高度非线性特征有助于促进抵御对抗样本的性能。基于这个观点, 作者设计了深度联合记忆网络。深度联合训练网络的设计主要基于神经营回路, 突触的计算方式是高度非线性并且饱和的。

在 [14]中, 作者提出Parseval网络来抵御对抗攻击。Parseval网络的核心思想是借用分层的正则化来控制神经网络, 使得神经网络具有全局Lipschitz连续性。作者将神经网络的每一层当做一个函数, 整体神经网络作这些函数的复合函数。神经网络对原始图片小扰动的稳健性由层函数的Lipschitz连续性来决定。Parseval网络训练的目标函数旨在控制权值矩阵的谱范数。

在 [19]中, 作者提出DeepCloak网络来抵御对抗攻击。DeepCloak网络在神经网络的中间层添加遮掩层。它能提取原始样本和对抗样本在输出特征层中的差异特征。由于神经网络中最大的权值对应于神经网络的最敏感特征, 因此这些最大的权值被遮掩层置为0。

一些防御的方法旨在设立机制来探测对抗样本, 在 [40]中, 作者假设对抗样本在神经网络的RELU激活层中具有不同的特征。作者使用径向基的SVM分类器来区分原始样本和异常样本的特征。在 [45]中, 作者提议将子网络添加到已知网络的中间层中进行而分类训练。子网络可以用来识别对抗样本和原始数据的差别。在 [23]中, 作者设计新的损失函数进行类增广训练, 从而探测对抗样本。

4.2 对抗机器学习和异常点检测方法

在对抗机器学习中，原始样本定义为已知类数据集中的样本。给定原始样本，它的对抗样本定义为人类无法区分它和原始样本，但是分类器却错误分类的样本。如果原始样本和对抗样本在分类器输出的结果不同，对抗样本在神经网络中间特征层与原始样本存在显著差别。因此，对抗样本在神经网络中间层的输出可能为原始数据在中间层分布的异常点。在对抗机器学习中，防御方利用异常点检测法来探测对抗样本。在开放识别问题中，一些工作利用对抗训练来提高分类器对未知类的识别能力。

在 [46] 中，作者提出基于异常点检测法来识别对抗样本。给定训练样本 $X = \{X^c, c = 1, \dots, K\}$ ，测试样本 x ， L 层深度神经网络DNN以及第 c 类训练样本 $X^c = \{x_1^c, x_2^c, \dots, x_{N_c}^c\}$ 。作者首先利用训练样本 X 训练DNN，并且计算 X_c 在第 l 层的输出为 $Z^c = \{z_1^c, z_2^c, \dots, z_{N_c}^c\}$ 。对所有的 Z_c ，计算条件概率密度函数 $f(z|c)$ 。给定测试样本 x ，一个经典的检测方法如下：

- 通过 x 的后验概率密度函数确定 x 的类别 $c^* = \arg \max_{c \in 1, \dots, K} P(C = c|x)$ 。
- 计算 x 在 DNN 第 l 层的输出 z 。
- 计算 z 的条件概率密度函数值 $f(z|c)$ 。当 $f(z|c)$ 小于某一个阈值时，则 x 为对抗样本。

作者提出新的模型来检测对抗样本。当测试样本 x 关于不同类别的 $f(z|c)$ 分布与 softmax 的输出分布不一致时，则称该样本为对抗样本。样本的一致性利用 KL 散度来度量。假设 $\hat{c}_s = \arg \max_{c \in \{1, \dots, K\} - c^*} f(z|c)$ 。作者构造二类概率密度函数

$$P \equiv \{P_{c^*}, P_{\hat{c}_s}\} = \{p_0 P_{c^*}, p_0 P_{\hat{c}_s}\}, \quad (4.12)$$

其中 p_0 是标准化常数： $p_0 = (f(z|c^*) + f(z|\hat{c}_s))^{-1}$ 。同时，作者构造二类概率密度函数

$$Q \equiv \{Q_{c^*}, Q_{\hat{c}_s}\} = \{q_0 P[c^*|x], q_0 P[\hat{c}_s|x]\}, \quad (4.13)$$

其中 q_0 是标准化常数： $q_0 = (P[c^*|x] + P[\hat{c}_s|x])^{-1}$ 。 P 和 Q 的 KL 散度定义为

$$D_{KL}(P\|Q) = \sum_{c \in \{c^*, \hat{c}_s\}} P_c \log\left(\frac{P_c}{Q_c}\right). \quad (4.14)$$

当 $D_{KL}(P\|Q)$ 较大时，则判定 x 为对抗样本。

4.3 基于对抗机器学习的开放集分类问题

在 [43] 中，作者提出基于距离的置信度以及对抗训练来增强分类器的未知类识别能力的方法。利用基于距离的置信度判别未知类的基本假设是在测试样本 x 的邻域内，某一类别的点所占的比例越大，到这一类别的点的距离越小，则测试样本 x 属于该类别的可能性越大。令 $f(x)$ 为 x 在神经网络中间层的输出， $A(x) = \{x_{\text{train}}^j\}_{j=1}^k$ 为样本 x 在训练集中的 k 近邻域，令 $\{y^j\}_{j=1}^k$ 为 $A(x)$ 中的点对应的类别坐标。则将 x 的类别判定为 \hat{y} 的置信度 $D(x)$ 定义如下：

$$D(x) = \frac{\sum_{j=1, y^j=\hat{y}}^k e^{-\|f(x)-f(x_{\text{train}}^j)\|_2}}{\sum_{j=1}^k e^{-\|f(x)-f(x_{\text{train}}^j)\|_2}}. \quad (4.15)$$

作者提出两个正则化方法来提高分类器的性能。第一个方法的原理是减小同类样本在特征空间分布中的距离，增加不同类别的样本在特征空间中的距离。第一个方法的目标函数如下：

$$L(x, y) = L_{\text{class}}(X, Y) + \alpha L_{\text{dist}}(X, Y) \quad L_{\text{dist}}(X, Y) = \frac{1}{P} \sum_{p=1}^P L_{\text{dist}}(x^{p_1}, x^{p_2}), \quad (4.16)$$

这里， $L_{\text{dist}}(x^i, x^j)$ 定义为

$$\begin{cases} \|f(x^i) - f(x^j)\|_2 & \text{if } y^i = y^j. \\ \max\{0, (m - \|f(x^i) - f(x^j)\|_2)\} & \text{if } y^i \neq y^j. \end{cases} \quad (4.17)$$

FGSM 方法是经典的对抗样本生成方法。FGSM 生成的对抗样本具有解析表达式。文献的第二个方法利用 FGSM 方法对分类器做对抗训练。作者首先利用 FGSM 方法产生对抗样本，然后将对抗样本设定为和原始样本相同的标签进行训练。

4.4 基于提高分类器可分性的新算法

在 [43] 中，作者利用对抗机器学习结合基于距离的度量来增强分类器的分类效果和识别异常样本的能力。接下来，我们说明对抗机器学习技术有效的原因为。

给定利用已知类数据集训练的分类器，在未知类识别的任务中，会犯两类错误。第一类错误是将已知类数据样本映射到其它类别的特征区域之中。在这

个情况下，分类器会将已知类样本错判定为未知类的样本。第二个错误是将未知类数据映射到已知类的特征区域之中。在这个情况下，分类器将未知类样本错分为已知类样本。

对抗机器学习中的对抗训练旨在通过将对抗样本和原始样本在分类器的输出相同来增加分类器的光滑性，其中光滑性是指对输入样本进行小扰动，分类器的输出变化很小。当分类器光滑性增加时，分类器犯第一个错误的概率将减少。与原始图片相似的样本会被分类器映射到已知类的特征区域之中。

现在，我们通过实验来说明分类器犯第二类错误的可能性以及设计减小分类器犯第二类错误的方法。给定基于SVHN 数据集进行训练的分类器 c 和其中100张图片 $\{x_i\}_{i=1}^{100}$ 。利用优化方法计算100个近似随机的图片 $\{x'_i\}_{i=1}^{100}$ ，使得 x_i 和 x'_i 在分类器特征层的输出相同。令分类器 c 的特征层输出为 f ，则优化方法的目标函数如下：

$$\min_{x'_i} \sum_i \|f(x_i) - f(x'_i)\|. \quad (4.18)$$

在图4.1中，左边的图片是给定的100张训练图片，右边的图片是求解上述优化问题的结果。右边近似随机的图片与训练集图片在特征空间中的输出相同。

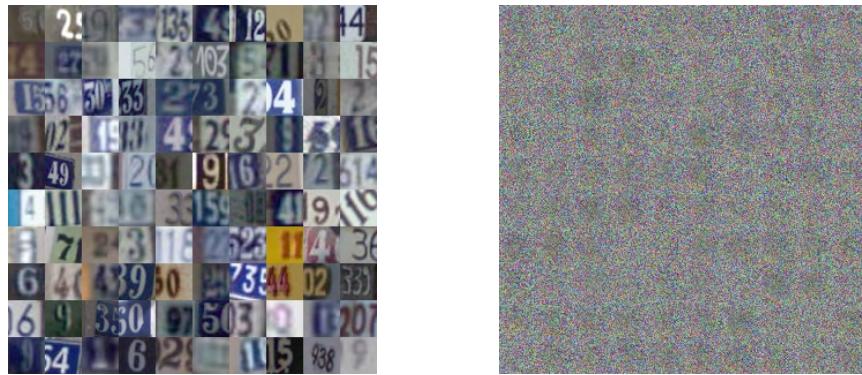


图 4.1: SVHN数据集和新型对抗样本

由于第二类错误的存在，分类器可能将未知类别的图片以高置信度错分成某些已知类别。分类器犯第二类错误的现象是合理的，其主要原因是分类器的输入针对的是全空间。只有对全空间的样本都进行标注，才能令分类器避免犯第二类错误。由于分类器缺乏针对原始数据以外的数据信息进行训练，分类器针对这些数据缺乏足够的辨识能力。考虑极端的情形，分类器只有一个类别，

那么分类器在训练的过程中，分类器对应的函数会收敛到一个常数。在这个情况下，分类器会将所有的未知类样本错分为已知类别。

我们观察到当分类器犯第二类错误的概率较大时，特征空间中的已知类别区域的原像范围过大。因此我们训练生成器，使得生成器生成在原始空间中距离原始数据集较远，但是在特征空间中被分类器映射到已知类别的区域中的样本。这些样本的存在说明分类器区别未知类别的可分性较弱。利用这些样本训练分类器，可以减少分类器犯第二类错误的概率。我们提出的目标函数的考虑如下：

- 生成器输出的样本在分类器特征层的输出分布与原始样本在分类器特征层的输出分布相同。
- 构造二分类器，使得二分类器可以区分真实数据和生成器输出的数据。

在我们提出的方法中，生成器生成的样本和GAN的生成器生成的样本具有下面两个不同之处：

- GAN的生成器在原始空间中拟合训练数据，我们提出的生成器在特征空间中拟合训练数据在特征空间的输出分布。
- 在GAN的目标函数中，生成器的输出收敛到真实的数据分布。在我们提出的目标函数中，生成器的输出不一定收敛到真实的数据分布。

给定分类器 $F(x)$ ，生成器 $G(x)$ ，二分类器 $L(x)$ 。令分类器通过softmax层的输出为 $F(x)$ ，在softmax前一层的输出为 $Z(x)$ 。识别未知类的异常值 T 定义为 $\sum_{i=1}^n \exp(Z_i(x))$ ，其中 n 为已知类的个数。给定训练集 S 的样本 x 。 $T(x)$ 的值越高，则样本属于已知类数据的可能性越大。

在我们的训练方法中，分类器，生成器和二分类器交替训练。在每一次迭代中，分类器，生成器和二分类器的训练依次进行，其具体步骤如下：

1. 训练分类器 $F(x)$ ：对已知类数据进行标准的分类任务，将生成器输出的样本作为未知类进行对抗训练。其目标函数定义如下：

$$\sum_{(x,y) \in S} l(F(x), y) + \sum_i \int_{z \sim p(z)} \exp(Z_i(G(z))). \quad (4.19)$$

2. 训练生成器 G ：设计生成器的两个目标函数，使得生成器的输出满足下面两个条件：

- 生成器输出的样本被分类器判定为已知类。
- 生成器输出的样本在原始空间中和原始数据保持一定的距离，容易被区分。

对第一个条件，我们采用GAN中的feature matching目标函数来拟合生成器的输出和原始样本在分类器特征空间中的分布：

$$\|\mathbb{E}_{x \sim p_{data}} f(x) - \mathbb{E}_{z \sim p(z)} f(G(z))\|, \quad (4.20)$$

其中 f 为分类器 F 在特征层的输出。对于第二个条件，我们利用二分类器 $L(x)$ 来训练生成器 G 。假设原始样本的输出标签为 $(1, 0)$ ，生成器输出样本的标签为 $(0, 1)$ ，则生成器 G 的训练目标函数如下：

$$\int_{z \sim p(z)} l(L(G(z)), (0, 1)), \quad (4.21)$$

其中 l 为分类损失函数。

二分类器 $L(x)$ 的目的是训练生成器，使得生成器输出的样本与原始样本保持一定的距离。二分类器的目标函数如下：

$$\int_{x \sim p_{data}} l(L(x), (1, 0)) + \int_{z \in p(z)} l(L(G(z)), (0, 1)). \quad (4.22)$$

4.5 实验结果

我们的训练模型在SVHN数据集的实验结果包括对已知类的分类结果和对未知类的识别结果。我们将训练集中的9类样本作为已知类进行训练，将剩余的1类样本作为未知类。对于分类精度，我们记录测试集的对应的9类样本的分类正确率。对于识别未知类的精度，我们将测试集中的9类样本作为已知类，其余的1类样本作为未知类，利用AUC指标来度量分类器识别未知类的性能。假设分类器的logit输出(softmax输出的前一层)为 Z ，softmax的输出为 F 。在实验中，我们列举的比较方法的基本思想是利用已知类别的训练数据训练分类器。利用统计量和异常值检测方法在分类器的特征空间中检测未知类数据，其具体形式如下：

- 熵: Entropy(F)。当测试样本输出分布的熵增加时, 其分布趋近于均匀分布并且它的最高置信度会降低。基于熵的异常值, 熵越大, 则样本属于未知类别的可能性就越高。
- 最大值: $\max Z_i$ 。测试样本的 Z_i 输出越大, 测试样本属于已知类别的可能性越高。
- logit和: $\sum \exp(Z_i)$ 。原理和最大值相同, 考虑属于已知类别的可能性之和。
- KDE: 利用KDE拟合神经网络特征空间中的已知类别数据的特征分布, 利用拟合分布的似然度来检测未知类。
- KNN: 在神经网络的特征层中进行基于KNN的异常点检测方法。测试样本与已知类分布的距离越大, 则属于未知类别的可能性越大。

我们的分类正确率和识别未知类的精度是分别独立测量的。首先分类正确率和未识别精度的最高值不一定同时达到, 所以需要考虑折中方案。用于识别未知类的结果(AUC值)在表4.1中列出。

表 4.1: SVHN数据集识别未知类的AUC值

Anomaly	logit和	极大值	熵	KDE	KNN	FM	FM-2
0	0.88	0.882	0.887	0.865	0.87	0.912	0.914
1	0.955	0.953	0.946	0.923	0.914	0.938	0.931
2	0.941	0.943	0.934	0.936	0.923	0.939	0.947
3	0.891	0.889	0.885	0.871	0.884	0.887	0.891
4	0.942	0.941	0.946	0.915	0.927	0.937	0.932
5	0.871	0.872	0.865	0.853	0.844	0.897	0.906
6	0.843	0.845	0.842	0.824	0.839	0.864	0.868
7	0.901	0.896	0.881	0.871	0.872	0.895	0.892
8	0.883	0.884	0.886	0.859	0.874	0.892	0.896
9	0.896	0.893	0.898	0.868	0.849	0.907	0.917
平均	0.901	0.898	0.897	0.878	0.879	0.906	0.909

对于分类精度, Baseline方法定义为softmax分类器直接训练得到的分类

器。我们的方法和Baseline方法的分类精度浮动范围为0.3个点，最后的平均数值几乎相同。分类精度的实验结果(分类正确率)在表4.2中列出。

表 4.2: SVHN数据集的分类正确率

Anomaly	Baseline	FM	FM-2
0	0.948	0.947	0.949
1	0.951	0.953	0.952
2	0.953	0.952	0.954
3	0.947	0.949	0.947
4	0.952	0.951	0.953
5	0.949	0.95	0.95
6	0.947	0.947	0.946
7	0.953	0.953	0.954
8	0.952	0.953	0.953
9	0.955	0.956	0.955
平均	0.951	0.952	0.952

在开放集分类问题中，我们提出新的方法来增强分类器识别未知类的能力。虽然在已有的实验中，我们的方法有效，但是仍然有提升空间。首先考虑生成器的选取方面。常用的生成器结构包括autoencoder和decoder。基于decoder结构的生成器主要用在GAN中，其生成样本的多样性强，但是收敛性弱。基于autoencoder的生成器收敛性强，但是由于一个输入只能输出一个样本，因此多样性较弱。分类器做对抗训练的目标函数值得深入研究。一些工作利用极小化边缘的目标函数将输出分布拟合为均匀分布，它们的性能和我们所提出的方法的优劣性有待深入的研究。旨在在原始空间中区分原始样本和生成器输出的样本的目标函数效果不明显。当生成器生成的样本保留原始样本的风格信息时，才能产生强对抗样本。因此，如何生成与原始样本风格一致的强对抗样本是值得研究的问题。

第五章 总结和展望

在这个出站报告中，我们主要介绍两个工作。第一个工作是基于GAN的单类分类器来解决异常点检验问题。第二个工作利用对抗学习的思想来正则化分类器的特征空间，从而改进分类器识别未知类别的性能。

尽管这两个工作的有效性在当前的实验中被论证，这些工作仍然有更多的探索和挖掘空间。对于第一个工作，我们的方法在标准的数据集中展现了高性能。与其它的异常点检验方法相比，由于GAN训练的随机性，GAN的性能稳定性低于其他的方法。如何稳定化GAN的性能并且生成更有效的异常样本是将来研究的主题。同时，研究在数据规模较大的情况下如何利用动态样本生成的方法来设计高效的异常点检验方法至今没有被充分研究。

对于第二个工作，我们的方法和性能还有待进一步的探索和提高。首先我们设计的对抗样本缺乏理论解释和保证，对抗样本对分类器影响的机制和原理没有被研究。方法的有效性还需要在更多的数据集中验证。如何设计高效的目标函数和生成器来提高对抗样本的质量是未来比较重要的研究方向。

参考文献

- [1] J. An and S. Cho. Variational autoencoder based anomaly detection using reconstruction probability. *SNU Data Mining Center, Tech. Rep.*, 2015.
- [2] S. Ando. Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection. In *Seventh IEEE International Conference on Data Mining*, pages 13–22. IEEE, 2007.
- [3] C. Aytekin, X. Ni, F. Cricri, and E. Aksu. Clustering and unsupervised anomaly detection with l₂ normalized deep auto-encoder representations. *arXiv preprint arXiv:1802.00187*, 2018.
- [4] S. Baluja and I. Fischer. Adversarial transformation networks: Learning to generate adversarial examples. *arXiv preprint arXiv:1703.09387*, 2017.
- [5] A. Bendale and T. Boult. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1893–1902, 2015.
- [6] A. Bendale and T. E. Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.
- [7] X. Bin, Y. Zhao, and B. Shen. Abnormal subspace sparse pca for anomaly detection and interpretation. *arXiv preprint arXiv:1605.04644*, 2016.
- [8] C. Campbell and K. P. Bennett. A linear programming approach to novelty detection. In *Advances in neural information processing systems*, pages 395–401, 2001.
- [9] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57. IEEE, 2017.

- [10] R. Chalapathy, A. K. Menon, and S. Chawla. Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*, 2018.
- [11] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology*, 2(3):27, 2011.
- [12] V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, and B. Maglaris. Hierarchical anomaly detection in distributed large-scale sensor networks. In *IEEE Symposium on Computers and Communications*, pages 761–767. IEEE, 2006.
- [13] A. Chaudhuri, D. Kakde, C. Sadek, L. Gonzalez, and S. Kong. The mean and median criterion for automatic kernel bandwidth selection for support vector data description. *CoRR*, abs/1708.05106, 2017.
- [14] M. Cisse, Y. Adi, N. Neverova, and J. Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017.
- [15] D. A. Clifton, P. R. Bannister, and L. Tarassenko. Learning shape for jet engine novelty detection. In *International Symposium on Neural Networks*, pages 828–835. Springer, 2006.
- [16] K. Do, T. Tran, and S. Venkatesh. Multilevel anomaly detection for mixed data. *arXiv preprint arXiv:1610.06249*, 2016.
- [17] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.
- [18] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie. High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition*, 58:121–134, 2016.
- [19] J. Gao, B. Wang, Z. Lin, W. Xu, and Y. Qi. Deepcloak: Masking deep neural network models for robustness against adversarial samples. *arXiv preprint arXiv:1702.06763*, 2017.

- [20] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017.
- [21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [22] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [23] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.
- [24] S. Gu and L. Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.
- [25] M. Hassen and P. K. Chan. Learning a neural-network-based representation for open set recognition. *arXiv preprint arXiv:1802.04365*, 2018.
- [26] V. Hautamaki, I. Karkkainen, and P. Franti. Outlier detection using k-nearest neighbour graph. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 430–433. IEEE, 2004.
- [27] K. A. Heller, K. M. Svore, A. D. Keromytis, and S. J. Stolfo. One class support vector machines for detecting anomalous windows registry accesses. In *Proc. of the workshop on Data Mining for Computer Security*, volume 9, 2003.
- [28] T. Ina, A. Hashimoto, M. Iiyama, H. Kasahara, M. Mori, and M. Minoh. Outlier cluster formation in spectral clustering. *arXiv preprint arXiv:1703.01028*, 2017.
- [29] L. P. Jain, W. J. Scheirer, and T. E. Boult. Multi-class open set recognition using probability of inclusion. In *European Conference on Computer Vision*, pages 393–409. Springer, 2014.

- [30] R. A. Jarvis and E. A. Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on computers*, 100(11):1025–1034, 1973.
- [31] H. Jiang, H. Wang, W. Hu, D. Kakde, and A. Chaudhuri. Fast incremental svdd learning algorithm with the gaussian kernel. *arXiv preprint arXiv:1709.00139*, 2017.
- [32] P. R. M. Júnior, T. E. Boult, J. Wainer, and A. Rocha. Specialized support vector machines for open-set recognition. *arXiv preprint arXiv:1606.03802*, 2016.
- [33] P. R. M. Júnior, R. M. de Souza, R. d. O. Werneck, B. V. Stein, D. V. Pazinato, W. R. de Almeida, O. A. Penatti, R. d. S. Torres, and A. Rocha. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106(3):359–386, 2017.
- [34] V. Jyothisna, V. R. Prasad, and K. M. Prasad. A review of anomaly based intrusion detection systems. *International Journal of Computer Applications*, 28(7):26–35, 2011.
- [35] D. Kakde, A. Chaudhuri, S. Kong, M. Jahja, H. Jiang, and J. Silva. Peak criterion for choosing gaussian kernel bandwidth in support vector data description. In *IEEE International Conference on Prognostics and Health Management*, pages 32–39. IEEE, 2017.
- [36] D. Krotov and J. J. Hopfield. Dense associative memory is robust to adversarial inputs. *arXiv preprint arXiv:1701.00939*, 2017.
- [37] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.
- [38] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

- [39] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):3, 2012.
- [40] J. Lu, T. Issaranon, and D. Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. *arXiv preprint arXiv:1704.00103*, 2017.
- [41] M. Lucic, O. Bachem, and A. Krause. Linear-time outlier detection via sensitivity. *arXiv preprint arXiv:1605.00519*, 2016.
- [42] Y. Luo, X. Boix, G. Roig, T. Poggio, and Q. Zhao. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292*, 2015.
- [43] A. Mandelbaum and D. Weinshall. Distance-based confidence score for neural network classifiers. *arXiv preprint arXiv:1709.09844*, 2017.
- [44] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2624–2637, 2013.
- [45] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.
- [46] D. J. Miller, Y. Wang, and G. Kesisidis. When not to classify: Anomaly detection of attacks (ada) on dnn classifiers at test time. *arXiv preprint arXiv:1712.06646*, 2017.
- [47] T. Miyato, A. M. Dai, and I. Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.
- [48] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. *CoRR*, abs/1610.08401, 2016.
- [49] S. M. Moosavi Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

- [50] A. Nayebi and S. Ganguli. Biologically inspired protection of deep networks from adversarial attacks. *arXiv preprint arXiv:1703.09202*, 2017.
- [51] L. Nguyen and A. Sinha. A learning approach to secure learning. *CoRR*, abs/1709.04447, 2017.
- [52] A. Nitin Bhagoji, D. Cullina, C. Sitawarin, and P. Mittal. Enhancing robustness of machine learning systems via data transformations. *arXiv preprint arXiv:1704.02654*, 2017.
- [53] T. Ortner, P. Filzmoser, M. Zaharieva, S. Brodinova, and C. Breiteneder. Local projections for high-dimensional outlier detection. *arXiv preprint arXiv:1708.01550*, 2017.
- [54] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. In *19th International Conference on Data Engineering*, pages 315–326. IEEE, 2003.
- [55] N. Papernot and P. McDaniel. Extending defensive distillation. *arXiv preprint arXiv:1705.05264*, 2017.
- [56] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *IEEE Symposium on Security and Privacy*, pages 372–387. IEEE, 2016.
- [57] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, pages 582–597. IEEE, 2016.
- [58] S. Peredriy, D. Kakde, and A. Chaudhuri. Kernel bandwidth selection for svdd: The sampling peak criterion method for large data. In *IEEE International Conference on Big Data*, pages 3540–3549. IEEE, 2017.
- [59] P. Perera and V. M. Patel. Learning deep features for one-class classification. *arXiv preprint arXiv:1801.05365*, 2018.

- [60] D. Pokrajac, A. Lazarevic, and L. J. Latecki. Incremental local outlier detection for data streams. In *IEEE Symposium on Computational Intelligence and Data Mining*, pages 504–515. IEEE, 2007.
- [61] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult. The extreme value machine. *IEEE transactions on pattern analysis and machine intelligence*, 2017.
- [62] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli. Adversarially learned one-class classifier for novelty detection. *arXiv preprint arXiv:1802.09088*, 2018.
- [63] M. Salehi, C. Leckie, J. C. Bezdek, T. Vaithianathan, and X. Zhang. Fast memory efficient local outlier detection in data streams. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3246–3260, 2016.
- [64] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.
- [65] S. Sankaranarayanan, A. Jain, R. Chellappa, and S. N. Lim. Regularizing deep networks using efficient layerwise adversarial training. *arXiv preprint arXiv:1705.07819*, 2017.
- [66] S. Sarkar, A. Bansal, U. Mahbub, and R. Chellappa. Upset and angri: Breaking high performance image classifiers. *arXiv preprint arXiv:1707.01159*, 2017.
- [67] W. J. Scheirer, L. P. Jain, and T. E. Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2317–2324, 2014.
- [68] W. J. Scheirer, A. Rocha, R. J. Micheals, and T. E. Boult. Meta-recognition: The theory and practice of recognition score analysis. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1689–1695, 2011.

- [69] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.
- [70] M. Schneider, W. Ertel, and F. Ramos. Expected similarity estimation for large-scale batch and streaming anomaly detection. *Machine Learning*, 105(3):305–333, 2016.
- [71] M. Schreyer, T. Sattarov, D. Borth, A. Dengel, and B. Reimer. Detection of anomalies in large scale accounting data using deep autoencoder networks. *arXiv preprint arXiv:1709.05254*, 2017.
- [72] A. Siffer, P.-A. Fouque, A. Termier, and C. Largouet. Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1067–1075. ACM, 2017.
- [73] J. Su, D. V. Vargas, and S. Kouichi. One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864*, 2017.
- [74] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [75] Y. Yan, L. Cao, C. Kulhman, and E. Rundensteiner. Distributed local outlier detection in big data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1225–1234. ACM, 2017.
- [76] Y. Yan, L. Cao, and E. A. Rundensteiner. Scalable top-n local outlier detection. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1235–1244. ACM, 2017.

- [77] Q. Yang and F. Li. Support vector machine for intrusion detection based on lsi feature selection. In *Sixth World Congress on Intelligent Control and Automation*, volume 1, pages 4113–4117. IEEE, 2006.
- [78] C. You, D. P. Robinson, and R. Vidal. Provable self-representation based outlier detection in a union of subspaces. *arXiv preprint arXiv:1704.03925*, 2017.
- [79] Y. Yu, W.-Y. Qu, N. Li, and Z. Guo. Open-category classification by adversarial sample generation. *arXiv preprint arXiv:1705.08722*, 2017.
- [80] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang. Deep structured energy based models for anomaly detection. *arXiv preprint arXiv:1605.07717*, 2016.
- [81] H. Zhang and V. M. Patel. Sparse representation-based open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1690–1696, 2017.
- [82] P. Zhou and J. Feng. Outlier-robust tensor pca. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–9, 2017.
- [83] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.
- [84] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.

发表文章目录

- [1] Chu Wang, Yan-Ming Zhang, Cheng-Lin Liu. Anomaly Detection via Minimum Likelihood Generative Adversarial Networks. ICPR2018.

简 历

王础, 男, 辽宁省, 1989年出生. E-mail: chu.wang@ia.ac.cn

教育状况

- 2007.9-2011.7 大连理工大学, 数学学院, 信息与计算科学, 学士.
- 2011.9-2016.7 中国科学院数学与系统科学研究院, 系统所, 应用数学, 博士生, 导师: 支丽红研究员
- 2016.7-2018.7 中国科学院自动化研究所. 机器学习, 博士后, 合作导师: 刘成林研究员

研究领域

异常点检测, 对抗机器学习

致 谢

转眼间,我已经在中国科学院自动化所度过了两年的时光.这两年来,有辛苦的工作,有成功的喜悦,有失败的沮丧,有太多的人和事值得记忆.在出站报告完成之际,我感觉非常充实,感慨良多.首先,我感谢刘成林研究员.刘老师严谨的科研态度,敏锐的眼光和勇于探索的精神让我在学术上找到了正确的研究方向.他的教诲和鞭策深深的激励着我,让我勇于克服科研中所遇到的困难,不断的取得进步.刘老师积极乐观的生活态度和宽广豁达的胸襟让我学会了做人的道理.刘老师为我树立了一生的典范,让我终身受益.

衷心的感谢殷飞老师,张燕明老师,张煦尧老师,杨沛沛老师.没有各位老师的努力和教导,就没有PAL组现在的成就.