Variational Autoencoder based Anomaly Detection using Reconstruction Probability

Jinwon An
jinwon@dm.snu.ac.kr

Sungzoon Cho
zoon@snu.ac.kr

December 27, 2015

目录 CONTENTS

一、摘要

1、摘要

作者提出了一种使用基于变分自动编码器的重建概率的异常检测方法。重建概率是考虑变量分布的可变性的概率度量。重建概率比重建误差更加原则性和客观的异常分数。基于自动编码器和基于主成分的方法都是使用重建误差。实验结果表明基于变分自动编码器的重建概率的异常检测方法要优于以上两种方法。

创新点：提出重建概率作为判断异常的依据，避免了选择特定的阈值和异构数据做差。

二、背景

1、正在解决的问题是什么？

异常或异常值是与其余数据显着不同的数据点。分析和检测异常揭示了有关数据生成过程特征的有用信息。

在许多异常检测方法中，利用异常检测技术找到原始数据的低维嵌入，找到这些较低维的嵌入后，将它们带回原始数据空间，这称为原始数据的重建。通过对比重建前后的数据，判断该数据是否异常。

我们期望获得数据的真实性质，而没有无用的特征和噪声。

2、其他人对这个问题做了什么?

基于主成分分析的方法:
将数据点的重构误差，即原始数据点与其低维重构之间的误差，用作异常分数来检测异常。

基于自动编码器的方法:
通过堆叠自动编码器，以分层方式应用降维，在更高的隐藏层中获得更抽象的特征，从而更好地重建数据。再判断数据点的重构误差进行异常检测。

3、异常检测

异常检测方法大致可分为统计异常检测方法、基于邻近度的异常检测方法和基于偏差的异常检测方法

统计异常检测假设数据是根据指定的概率分布建模的。如果从模型生成数据点的概率低于某个阈值，则将其定义为异常。该模型的优点是给出了概率作为判断异常的决策规则，具有客观性和理论上的合理性。

基于邻近度的异常检测假设异常数据与大多数数据隔离。用这种方法建立异常模型的方法有三种：基于聚类的、基于密度的和基于距离的。

3、异常检测

基于聚类的异常检测，识别数据中存在的密集区域或簇。评估数据点与每个聚类的关系以形成异常分数。如果到聚类质心的距离高于阈值或最近聚类的大小低于阈值，则数据点被定义为异常。

基于密度的异常检测将异常定义为位于数据稀疏区域的数据点。如果数据点所在的局部区域内的数据点数量低于阈值，则将其定义为异常。

基于距离的异常检测使用与给定数据点的相邻数据点相关的测量。计算数据点与最近K个点距离之和，如果距离之和大于某个阈值，则将其定义为异常。

3、异常检测

基于偏差的异常检测主要基于谱异常检测，它使用重建误差作为异常分数。使用主成分分析或自动编码器等降维方法重建数据。测量其原始数据点与重建之间的差异称为重建误差，该误差可用作异常分数。具有高重建误差的数据点被定义为异常。

4、自动编码器异常检测

自动编码器是一种通过无监督学习训练的神经网络，经过训练可以学习接近其原始输入的重建。自动编码器由编码器和解码器两部分组成。具有单个隐藏层的神经网络具有分别如等式（1）和等式（2）中的编码器和解码器。 W和b是神经网络的权重和偏差，σ是非线性变换函数。

$$h = \sigma(W_{xh}x + b_{xh}) \tag{1}$$

$$z = \sigma(W_{hx}h + b_{hx}) \tag{2}$$

$$\|x - z\| \tag{3}$$

原始输入向量 x 和重建 z 之间的差异称为重建误差，如方程（3）所示。自动编码器学习如何最小化这种重建误差。

4、自动编码器异常检测

f_θ 和 g_ϕ是自动编码器的多层神经网络。

Algorithm 1 Autoencoder training algorithm

INPUT: Dataset $x^{(1)}, \cdots, x^{(N)}$

OUTPUT: encoder f_ϕ, decoder g_θ

$\phi, \theta \leftarrow$ Initialize parameters

repeat

$E = \sum_{i=1}^{N} \|x^{(i)} - g_\theta(f_\phi(x^{(i)}))\|$ Calculate sum of reconstruction error

$\phi, \theta \leftarrow$ Update parameters using gradients of E(e.g. Stochastic Gradient Descent)

until convergence of parameters ϕ, θ

4、自动编码器异常检测

该算法仅使用具有正常实例的数据来训练自动编码器。训练后，自动编码器将很好地重建正常数据，但对于自动编码器未遇到的异常数据则无法重建。所以，当重建后的数据与原数据误差大于某个阈值，则定义为异常。

Algorithm 2 Autoencoder based anomaly detection algorithm

INPUT: Normal dataset X, Anomalous dataset $x^{(i)}$ $i = 1, \cdots, N$, threshold α

OUTPUT: reconstruction error $\|x - \hat{x}\|$

$\phi, \theta \leftarrow$ train a autoencoder using the normal dataset X

for i=1 **to** N **do**

$reconstruction\ error(i) = \|x^{(i)} - g_\theta(f_\phi(x^{(i)}))\|$

if $reconstruction\ error(i) > \alpha$ **then**

$x^{(i)}$ is an anomaly

else

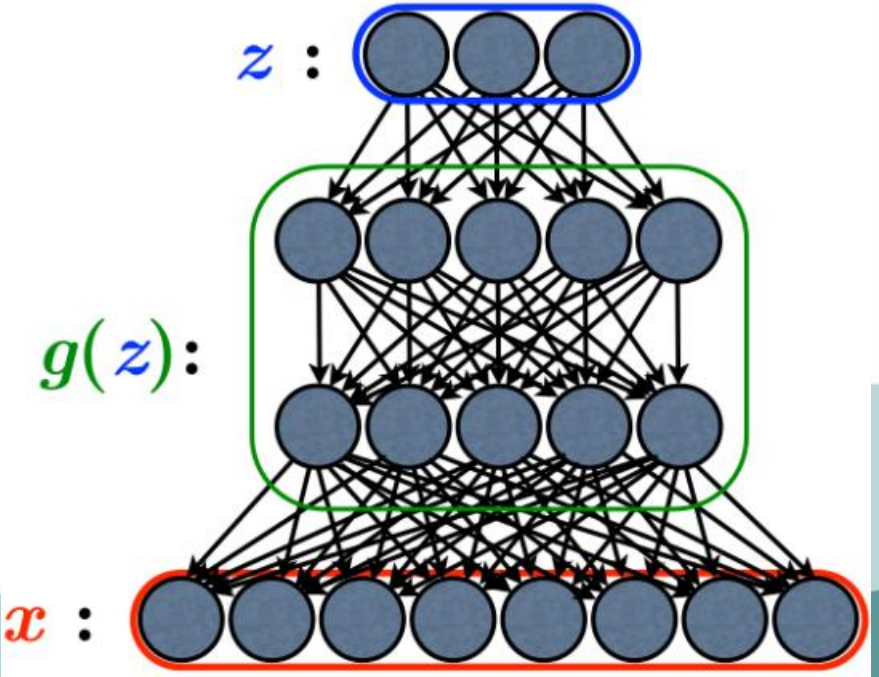
$x^{(i)}$ is not an anomaly

end if

end for

5、变分自动编码器

变分自动编码器（VAE）是一种有向概率图形模型（DPGM），其后验由神经网络近似，形成类似自动编码器的架构。
有向图模型使用有向非循环图（Directed Acyclic Graph，DAG）来描述变量之间的关系。如果两个节点之间有连边，表示对应的两个变量为因果关系，即不存在其他变量使得这两个节点对应的变量条件独立。



5、变分自动编码器

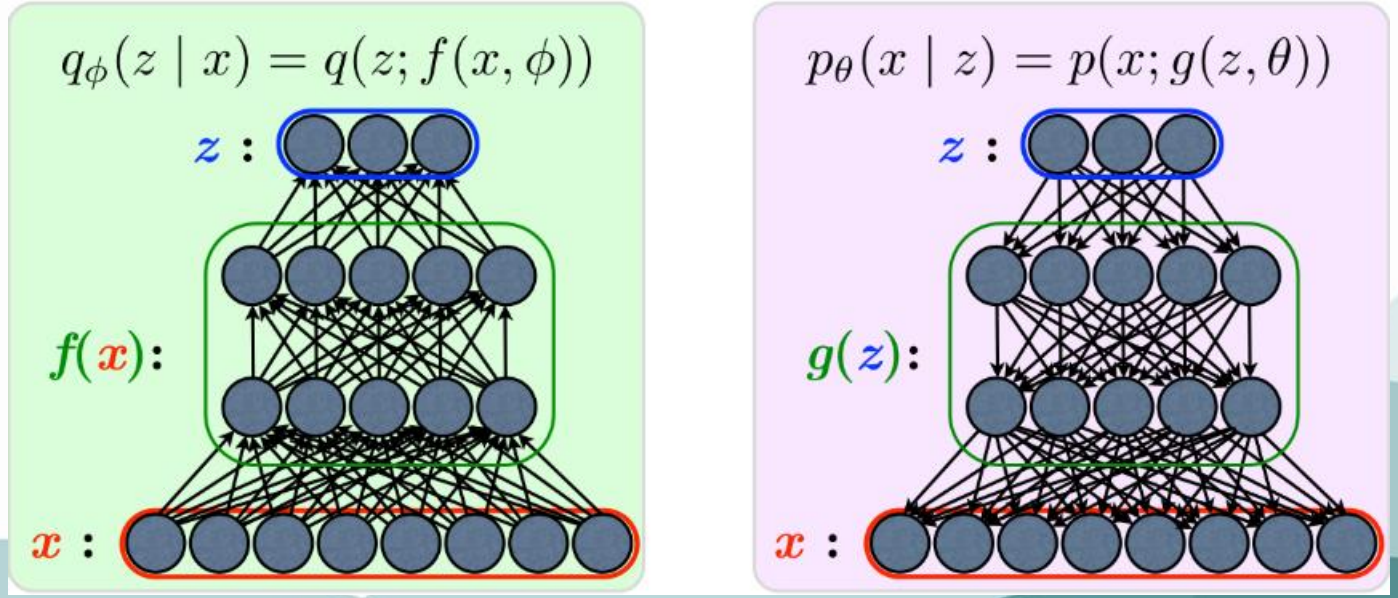
VAE利用两个神经网络建立两个概率密度分布模型：
一个用于原始输入数据的变分推断，生成隐变量的变分概率分布，称为推断网络。另一个根据生成的隐变量变分概率分布，还原生成原始数据的近似概率分布，称为生成网络。
想要推断网络求出隐变量 z 的变分概率分布 $p_\theta(z)$，但 $p_\theta(z)$ 难求，则利用近似后验分布 $q_\phi(z|x)$ 代替 $p_\theta(z)$。故最小化KL散度。
各个数据点的边缘似然之和 $\log p_\theta(x^{(1)},\cdots,x^{(N)}) = \sum_{i=1}^{N} \log p_\theta(x^{(i)})$

$$\log p_\theta(x^{(i)}) = D_{KL}(q_\phi(z|x)||p_\theta(z)) + \mathcal{L}(\theta,\phi;x^{(i)})$$

$$\begin{aligned}\log p_\theta(x^{(i)}) &\geq \mathcal{L}(\theta,\phi;x^{(i)}) \\ &= E_{q_\phi(z|x^{(i)})}[-\log q_\phi(z|x) + \log p_\theta(x|z)] \\ &= -D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)) + E_{q_\phi(z|x^{(i)})}[\log p_\theta(x|z)]\end{aligned}$$

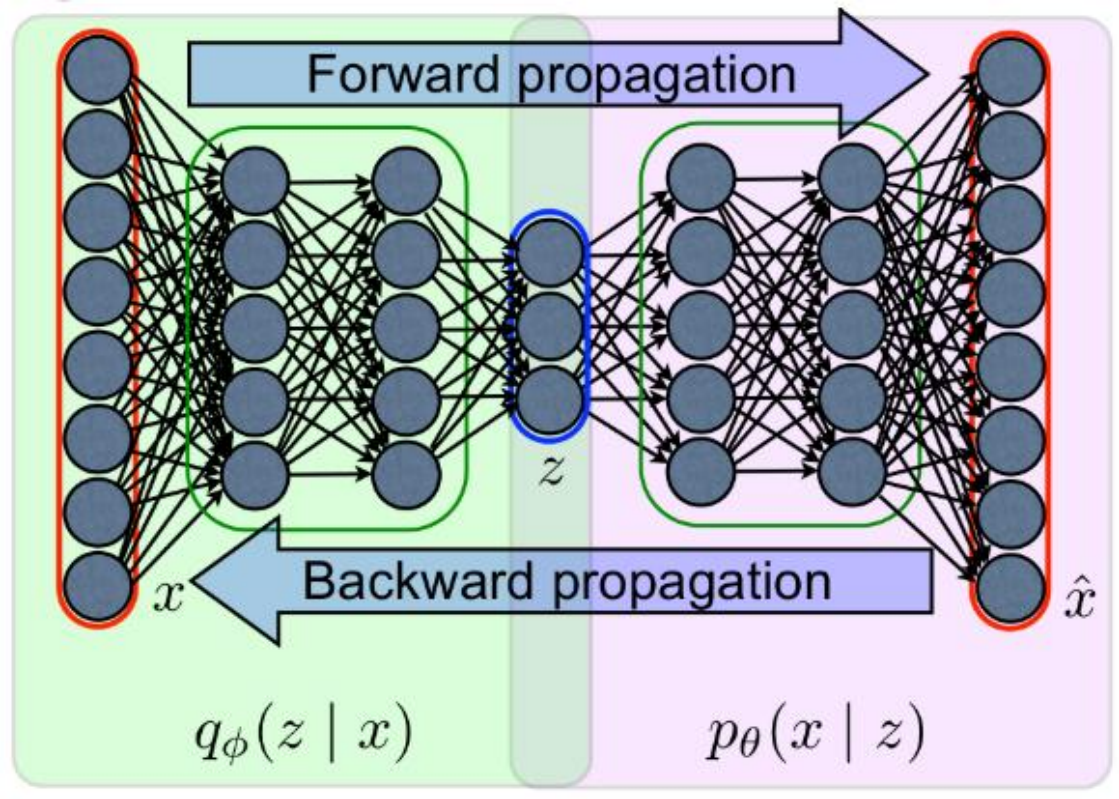
5、变分自动编码器

VAE 使用神经网络对近似后验$q_\phi(z|x)$的参数进行建模。$q_\phi(z|x)$是编码器，$p_\theta(x|z)$是解码器。 VAE 对分布参数而不是值本身进行建模，编码器中的 $f(x,\phi)$ 输出近似后验 $q_\phi(z|x)$的参数，并从 $q(z;f(x,\phi))$ 采样得到潜在变量 z 的实际值。为了得到重建 $\hat{x}$，给定样本 z，$p_\theta(x|z)$的参数通过 $g(z,\theta)$ 获得，其中重建 $\hat{x}$ 是从 $p_\theta(x;g(z,\theta))$采样的。



5、变分自动编码器

VAE 和自动编码器之间的主要区别在于，VAE 是一种随机生成模型，可以给出校准的概率，而自动编码器是一种确定性判别模型，没有概率基础。



5、变分自动编码器

随机变量 $z \sim q_\phi(z|x)$ 通过确定性变换 $h_\phi(\epsilon, x)$ 重新参数化，其中 ϵ 来自标准正态分布。

$$\tilde{z} = h_\phi(\epsilon, x) \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, 1)$$

Algorithm 3 Variational autoencoder training algorithm

INPUT: Dataset $x^{(1)}, \cdots, x^{(N)}$

OUTPUT: probabilistic encoder f_ϕ, probabilistic decoder g_θ

$\phi, \theta \leftarrow$ Initialize parameters

repeat

for i=1 **to** N **do**

Draw L samples from $\epsilon \sim \mathcal{N}(0, 1)$

$z^{(i,l)} = h_\phi(\epsilon^{(i)}, x^{(i)}) \quad i = 1, \cdots, N$

end for

$E = \sum_{i=1}^{N} -D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)) + \frac{1}{L}\sum_{l=1}^{L}(\log p_\theta(x^{(i)}|z^{(i,l)}))$

$\phi, \theta \leftarrow$ Update parameters using gradients of E (e.g. Stochastic Gradient Descent)

until convergence of parameters ϕ, θ

三、算法

1、变分自动编码器

我们提出了一种异常检测方法，该方法使用 VAE 来根据我们称为重建概率的概率来计算异常分数。
仅使用正常实例的数据来训练VAE。概率编码器 f_ϕ 和解码器 g_θ 都分别参数化潜在变量空间和原始输入变量空间中的各向同性正态分布。
计算从分布生成原始数据的概率。平均概率用作异常分数，称为重建概率。
具有高重建概率的数据点被归类为异常。

1、变分自动编码器

Algorithm 4 Variational autoencoder based anomaly detection algorithm

INPUT: Normal dataset X, Anomalous dataset $x^{(i)}$ $i = 1, \cdots, N$, threshold α

OUTPUT: reconstruction probability $p_\theta(x|\hat{x})$

$\phi, \theta \leftarrow$ train a variational autoencoder using the normal dataset X

for i=1 **to** N **do**

$\boldsymbol{\mu}_{\boldsymbol{z}^{(i)}}, \boldsymbol{\sigma}_{\boldsymbol{z}^{(i)}} = f_\theta(z|x^{(i)})$

draw L samples from $z \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{z}^{(i)}}, \boldsymbol{\sigma}_{\boldsymbol{z}^{(i)}})$

for l=1 **to** L **do**

$\boldsymbol{\mu}_{\hat{\boldsymbol{x}}^{(i,l)}}, \boldsymbol{\sigma}_{\hat{\boldsymbol{x}}^{(i,l)}} = g_\phi(x|z^{(i,l)})$

end for

$reconstruction\ probability(i) = \frac{1}{L}\sum_{l=1}^{L} p_\theta(x^{(i)}|\boldsymbol{\mu}_{\hat{\boldsymbol{x}}^{(i,l)}}, \boldsymbol{\sigma}_{\hat{\boldsymbol{x}}^{(i,l)}})$

if $reconstruction\ probability(i) < \alpha$ **then**

$x^{(i)}$ is an anomaly

else

$x^{(i)}$ is not an anomaly

end if

end for

四、评估

1、数据集

用于异常检测的数据集是 MNIST 数据集 和 KDD cup 1999 网络入侵数据集（KDD）。其中， KDD cup 数据集由五个主要类别组成，即 DoS、R2L、U2R、 Probe 和 Normal。前 4 类为异常，Normal 类为正常。
训练数据由80%的正常数据组成，测试数据由剩余的20%的正常数据和所有异常数据组成。

对于每个异常类别，正常数据以两种不同的方式定义。第一种是将普通数据定义为仅具有普通类的数据。另一种是将除指定异常类之外的所有数据定义为正常数据。

Table 1: KDD dataset

class	number of instances
Normal	972,770
DOS	3,914,580
probe	41,070
R2L	11,260
U2R	520

2、评估

使用受试者工作特征曲线下面积（AUC ROC）和平均精度或精度召回曲线曲线下面积（AUC PRC）和f1分数进行评估。 f1 分数是通过从验证数据集 f1 分数确定二元决策的阈值来获得的。

3、MNIST 数据集

VAE 在大多数情况下都优于其他模型。对于所有型号，当数字 1、7 和 9 为异常数字时，性能较低。

Table 2: MNIST AUC ROC performance

Anomaly digit	VAE	AE	PCA	kPCA
0	0.917	0.825	0.785	0.694
1	0.136	0.135	0.205	0.231
2	0.921	0.874	0.798	0.801
3	0.781	0.761	0.632	0.638
4	0.808	0.727	0.682	0.702
5	0.862	0.792	0.627	0.598
6	0.848	0.812	0.733	0.720
7	0.596	0.508	0.512	0.560
8	0.895	0.869	0.493	0.502
9	0.545	0.548	0.41	0.420

Table 3: MNIST VAE evaluation

Anomaly digit	AUC PRC	f1 score
0	0.517	0.537
1	0.063	0.205
2	0.644	0.598
3	0.251	0.332
4	0.337	0.381
5	0.325	0.427
6	0.432	0.433
7	0.148	0.212
8	0.499	0.49
9	0.104	0.21

4、KDD数据集

除异常类别为 DoS 的情况外，异常异常方法显示出更好的性能。

Table 4: KDD AUC ROC (only normal)

Anomaly	VAE	AE	PCA	kPCA
DoS	0.795	0.727	0.585	0.590
R2L	0.777	0.773	0.705	0.712
U2R	0.782	0.781	0.698	0712
Probe	0.944	0.946	0.832	0.821

Table 5: KDD AUC ROC (except anomaly)

Anomaly	VAE	AE	PCA	kPCA
DoS	0.744	0.685	0.785	0.780
R2L	0.786	0.782	0.502	0.514
U2R	0.921	0.806	0.717	0.760
Probe	0.970	0.968	0.647	0.645

Table 6: KDD VAE (only normal)

Anomaly	AUC ROC	AUC PRC	f1 score
DoS	0.795	0.944	0.981
R2L	0.777	0.17	0.406
U2R	0.782	0.084	0.324
Probe	0.944	0.751	0.791

Table 7: KDD VAE (except anomaly)

Anomaly	AUC ROC	AUC PRC	f1 score
DoS	0.744	0.935	0.979
R2L	0.786	0.135	0.389
U2R	0.921	0.0096	0.347
Probe	0.970	0.706	0.720

五、结论

1、结论

变分自动编码器的重建概率的异常检测方法。通过考虑变异性的概念，重构概率结合了变分自动编码器的概率特征。
重建概率作为一种概率度量，使其成为比自动编码器和基于 PCA 的方法的重建误差更加客观和有原则的异常分数。
实验结果表明，所提出的方法优于基于自动编码器和基于 PCA 的方法。由于其生成特性，还可以导出数据的重建来分析异常的根本原因。