计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191100141

# 高维大数据分析的无监督异常检测方法

邹承明[1,2,3]　陈　德[2]
1 交通物联网技术湖北省重点实验室　武汉 430070
2 武汉理工大学计算机科学与技术学院　武汉 430070
3 鹏城实验室　广东 深圳 518000
(zoucm@whut.edu.cn)

**摘　要**　高维数据的无监督异常检测是机器学习的重要挑战之一。虽然先前基于单一深度自动编码器和密度估计的方法已经取得了显著的进展,但是其仅通过一个深度自编码器来生成低维表示,这表明没有足够的信息来执行后续的密度估计任务。为了解决上述问题,文中提出了一种混合自动编码器高斯混合模型(Mixed Auto-encoding Gaussian Mixture Model,MAGMM)。MAGMM使用混合自动编码器来代替单一深度自动编码器生成串联的低维表示,因此它可以保存来自输入样本的特定集群的关键信息。此外,其利用分配网络来约束混合自动编码器,这样每个样本都可以分配给一个占主导地位的自动编码器。利用上述机制,MAGMM避免了陷入局部最优,降低了重构误差,从而可以促进密度估计任务的完成,提高高维数据异常检测的准确性。实验结果表明,该方法优于DAGMM,并在标准F1分数上提高了29%。

**关键词**:数据挖掘;无监督异常检测;降维;高斯混合模型;密度估计
**中图法分类号**　TP391

## Unsupervised Anomaly Detection Method for High-dimensional Big Data Analysis

ZOU Cheng-ming[1,2,3] and CHEN De[2]
1 Hubei Key Laboratory of Transportation Internet of Things Technology,Wuhan 430070,China
2 School of Computer Science and Technology,Wuhan University of Technology,Wuhan 430070,China
3 Peng Cheng Laboratory,Shenzhen,Guangdong 518000,China

**Abstract**　Unsupervised anomaly detection on high-dimensional data is one of the most significant challenges in machine learning. Although previous approaches based on single deep auto-encoder and density estimations have made significant progress, they generate low-dimensional representations as they use only a single deep auto-encoder, indicating that there is insufficient information to perform the subsequent density estimation task. To address the above challenge, a mixed auto-encoding gaussian mixture model (MAGMM) is proposed in this paper. MAGMM substitutes a single deep auto-encoder with a mixture of auto-encoders to generate concatenated low-dimensional representations, so that it can preserve key information from a specific cluster of the input sample. In addition, it utilizes an allocation network to constrain the mixture of auto-encoders, so that each sample can be assigned to a dominant auto-encoder. With the above mechanisms, MAGMM avoids from trapping into local optima and reduces the reconstruction errors, which can facilitate completing the density estimation tasks and improve the accuracy of high-dimensional data anomaly detection. Experimental results show that the proposed method performs better than DAGMM and achieves up to 29% improvement based on the standard F1 score.

**Keywords**　Data mining, Unsupervised anomaly detection, Dimensionality reduction, Gaussian mixture model, Density estimation

## 1　引言

异常检测可以发现有趣或者意外的模式,并揭示罕见但重要的现象。目前,其已被广泛应用于欺诈检测[1]、网络入侵检测[2]、医疗[3]、工业控制系统[4]等领域。然而在工业系统等领域中,数据通常是高维的、海量的、异构的、有噪声的、高需求的、实时的。因此,对高维数据的异常检测存在两个挑战,一个是"维数灾难"[5]问题,另一个是距离浓度[6]问题。这两个问题使得传统的数据分析算法不适用于高维数据领域。因此,高维数据分析中异常检测方法的研究成为当前的一个重要研究课题。

近年来,研究者对高维数据的异常检测方法进行了广泛

到稿日期:2019-11-19　返修日期:2020-04-02　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划(2018YFC0704300)
This work was supported by the National Key R&D Program of China(2018YFC0704300).
通信作者:陈德(chandler@whut.edu.cn)

的研究，具体方法包括深度聚类网络(Deep Clustering Network，DCN)[7]、两步法[8]、深度自编码高斯混合模型(Deep Autoencoding Gaussian Mixture Model，DAGMM)[9]等。上述工作在无监督异常检测方面取得了很大的进展，但仍存在一定的局限性。首先，这些工作大部分集中在预训练，而不是端到端训练。由于训练有素的自动编码器无法通过微调做出任何显著的改变，因此预训练很容易导致局部最优性能。例如，训练有素的自动编码器不知道后续的密度估计任务，在自动编码器中，异常检测所需的关键信息将被删除。其次，以往的研究大多没有考虑混合模型的优点，如 DAGMM 异常检测的精度主要依赖于单一深度自编码器产生的低维表示，而文献[9]没有考虑通过混合模型来进一步提升检测精度。最近，Ehsan 等[10]指出，与单个模型相比，这种混合物能够产生更好的预测。此外，Zhang 等[11]使用一种混合的自动编码器来识别和分离低维非线性流形的并集，并在数据集上产生了优于确定性深度聚类模型的性能。

针对上述局限性，本文提出了混合自动编码高斯混合模型(MAGMM)，这是一种用于无监督异常检测的新型混合结构。首先，在预训练方面，MAGMM 通过端到端的训练能够得到训练有素的自动编码器，这是因为预测网络使用的正则化极大地促进了混合自动编码器的性能提升，从而避免算法陷入局部最优。其次，在混合模型方面，用混合自动编码器代替 DAGMM 中使用的深度自动编码器，并添加分配网络来约束每个自动编码器。相比单个深度自编码器，该方法具有更低的重构误差和更高的异常检测精度。实验结果证明，本文模型能够有效地解决无监督异常检测方法中局部最优和低维表示信息不足的问题。

本文的主要贡献如下：

(1)提出了一种用于无监督异常检测的深度学习的模型结构。该模型在串联的低维表示中执行密度估计任务，结合了混合模型和自编码器两者的优点。

(2)提出了一种混合自动编码器的方法，该方法可以产生比单一深度自编码器更低的重构误差，并最终提高异常检测的精度。

本文第 2 节简要回顾了异常检测方面的研究工作；第 3 节说明了所提方法的总体结构，并提供了所使用的基本概念；第 4 节对所提方法和其他最新的异常检测基线在公共基准数据集上的实验结果进行了分析；最后对全文进行总结。

## 2 相关工作

由于在现实中获取标签信息代价较高，异常检测通常是在无监督的情况下进行的。根据使用的标准，无监督异常检测方法[4,12-14]大致可以分为基于重构、基于密度、基于支撑域等异常检测方法。

### 2.1 基于重构的异常检测方法

基于重构的方法假设：与正常样本相比，异常样本是少数，因此很难在低维投影中进行精确样本重构。传统的方法包括主成分分析(Principal Components Analysis，PCA)[15]、核 PCA[16]、鲁棒的 PCA[8]，它们通过核映射数据到特征空间进行降维。近年来，许多研究建议使用自编码器和变分自编码器来获取重构误差进行异常分析[17-20]。然而，这些工作都是通过一对编码器和解码器在输出端重建输入数据而完成的，性能受到限制。因此，文献[9]利用重构误差和低维表示对异常进行分析，并证明了该方法比基于重构误差的异常检测方法具有更好的性能。但是，复杂的高维数据无法通过重构来检测异常值。

### 2.2 基于密度估计的异常检测方法

基于密度的方法用于密度估计任务和异常检测，具体方法包括一维高斯分布[21]、多元高斯模型[22]、高斯混合模型[5]。该方法的缺点是，由于数据依赖，混合部件的数量难以选择，需要更多的样本来克服“维数灾难”。为了解决由两步法引起的密度估计与降维分离的局部最优问题，文献[23]提出了降维与高斯混合模型的联合学习方法。

### 2.3 基于支撑域的异常检测方法

基于支撑域的方法有一类支持向量机(One Class Support Vector Machine，OCSVM)[24]和支持向量数据描述(Support Vector Data Description，SVDD)[25]，其假设正常样本和异常样本可以通过边界进行相应的区分。然而，随着数据维度的增大，基于支撑域的方法在性能上受到限制，并且对异常值非常敏感。因此，当训练数据中存在异常值时，该方法的检测效果会受到很大影响。

### 2.4 总结

最近，有很多关于如何与(Gaussian Mixture Model，GMM)同时进行降维的研究和讨论[9,22,26-27]。此外，近年来的研究表明，混合模型正受到越来越多的关注。文献[11]使用混合自动编码器进行聚类，并证明了其性能优于单一确定的聚类模型。文献[10]使用的混合模型可以基于数据来自动调整混合模型中自编码器的数量。

与现有方法不同，本文既考虑了联合训练，又考虑使用混合模型进行降维工作。此外，本文提出了混合自动编码器，可以进一步降低重建误差，从而更好地完成后续的密度估计任务。最重要的是，本文通过一个复合目标函数来联合最小化混合重构误差、样本熵、批处理熵、样本能量，以便同时解决传统的 GMM 会落入局部最优的问题和单个深度编码器得到的低维表示关键信息不足的问题。

## 3 混合自动编码器高斯混合模型

### 3.1 MAGMM 模型架构

图 1 所示为 MAGMM 的整体架构。

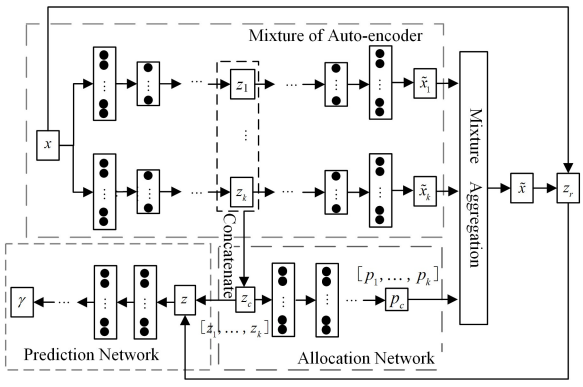



图 1 MAGMM 的整体架构

Fig. 1 Overall architecture of MAGMM

MAGMM 架构包含 4 个部分:1)混合自动编码器,由 $K$ 个自编码器组成,每个自编码器用于学习数据集中簇的低维流形;2)分配网络,接收 $K$ 个串联的低维空间$\boldsymbol{Z}_c$,并输出每个输入数据属于每个特定集群的聚类概率$\boldsymbol{P}_c$;3)混合聚集,其目的是得到输入样本 $\boldsymbol{X}$ 的聚集重构向量$\widetilde{\boldsymbol{X}}$,它是由每个自编码器的重构向量$\widetilde{\boldsymbol{X}}_i$和归一化后的聚类概率 $\boldsymbol{P}_i$ 计算得到的;4)预测网络,用于接收混合低维表示,由 $K$ 个串联的低维空间$\boldsymbol{Z}_c$和重构特征向量$\boldsymbol{Z}_r$得到,并预测输入样本的混合部件概率 $\boldsymbol{\gamma}$。

### 3.2 混合自动编码器

一般假设数据由多个低维非线性流形组成,我们的目标是将数据集中的每个样本分配到 $K$ 个集群。因此,使用单独的自动编码器对每个数据集群建模。通过混合自动编码器,其中每个自动编码器都应该确定一个适合于特定集群的非线性映射,我们可以将来自不同自动编码器的关键信息保存到串联的低维空间中,这些低维空间是由混合自动编码器中的每个自动编码器获得的。因此,串联的低维空间所保留的信息是由数据集中每个特定集群的低维空间组成。此外,每个自动编码器将产生一个单独的低维非线性空间和重构向量。给定一个样本 $\boldsymbol{X}$,混合自动编码器通过计算得到某个自动编码器的低维空间和重构向量如下:

$$\boldsymbol{Z}_i=\varepsilon(\boldsymbol{X},\theta_e),i=1,2,\cdots,K \tag{1}$$

$$\widetilde{\boldsymbol{X}}_i=\mathscr{D}(\mathscr{Z}_i,\theta_d),i=1,2,\cdots,K \tag{2}$$

其中,$\boldsymbol{Z}_i$是混合自动编码器的第 $i$ 个低维空间,$\widetilde{\boldsymbol{X}}_i$是由低维空间$Z_i$得到的第 $i$ 个重构向量。一个自编码器由一个编码器($E$)和解码器($\mathscr{D}$)组成,$\theta_e$和$\theta_d$是自编码器中编码器 $\varepsilon$ 和解码器 $\mathscr{D}$ 的参数。给定输入样本 $\boldsymbol{X}$,编码器将 $\boldsymbol{X}$ 映射到它的低维空间,然后解码器将低维空间映射到重构向量,最后混合自动编码器将串联的低维空间反馈给后续的分配网络和预测网络。

### 3.3 分配网络

由于混合自动编码器是由多个具有相同结构的自编码器组成,因此很难根据特定的簇自动将输入样本划分到不同的自编码器中。为了解决这个问题,我们利用分配网络尽可能均匀地将输入样本分配到每个自动编码器,并最终把输入样本分配到一个占主导地位的自动编码器。分配网络基于串联的低维空间推导出每个输入样本的聚类标签。在给定输入样本 $X$ 的情况下,分配网络从混合自动编码器中接收串联的低维空间,并输出输入样本的聚类概率:

$$\boldsymbol{Z}_c=(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_k)=(\varepsilon_1(\boldsymbol{Z}),\cdots,\varepsilon_k(\boldsymbol{Z})) \tag{3}$$

$$\begin{aligned}\boldsymbol{P}_c&=(\boldsymbol{P}_1,\cdots,\boldsymbol{P}_k\mid\theta_m)\\&=\mathrm{softmax}\{\boldsymbol{X}\in clusterK\mid\boldsymbol{Z}_c,\theta_m\}\end{aligned} \tag{4}$$

其中,$\boldsymbol{Z}_c$是 $K$ 个自编码器中低维空间的串联;$\theta_m$是分配网络的参数;$\boldsymbol{P}_c$是一个 $K$ 维向量,表示输入样本 $\boldsymbol{X}$ 属于某一个簇 $K$ 的聚类概率。此外,分配网络将$\boldsymbol{P}_c$提供给后续的混合聚合和预测网络。

### 3.4 混合聚集

混合自动编码器中的每个解码器接收低维空间,并产生相应的重构向量。此外,分配网络基于串联的低维空间生成输入样本属于一个簇的聚类概率。因此,通过混合聚集,我们可以获得相对于输入样本 $\boldsymbol{X}$ 的唯一的重构向量。混合聚集由混合自动编码器中每个解码器的重构向量和分配网络获得的聚类概率一起生成。给定输入样本点$\boldsymbol{X}^{(i)}$,混合聚集计算得到的整合的重构向量如下:

$$\widetilde{\boldsymbol{X}}^{(i)}=\sum_{k=1}^{K}\boldsymbol{P}_k^{(i)}\widetilde{\boldsymbol{X}}_k^{(i)} \tag{5}$$

其中,$\boldsymbol{X}^{(i)}$是第 $i$ 个输入样本,$\widetilde{\boldsymbol{X}}_k{}^{(i)}$是样本 $i$ 的第 $K$ 个解码器的重构向量,$\boldsymbol{P}_k^{(i)}$是样本 $i$ 的聚类概率,$\widetilde{\boldsymbol{X}}^{(i)}$是样本 $i$ 的 $K$ 个自编码器整合的重构向量。

直觉上,$\widetilde{\boldsymbol{X}}^{(i)}$等于$\widetilde{\boldsymbol{X}}_k^{(i)}$时,式(5)会得到最好的结果,此时$\boldsymbol{P}^{(i)}$是一个热向量,$\boldsymbol{P}^{(i)}$等于 1。

为了使$\boldsymbol{P}^{(i)}$是一个热向量,本文添加了一个样本熵:

$$\boldsymbol{H}(\boldsymbol{S}_i)=-\sum_{i=1}^{K}\boldsymbol{P}_k^{(i)}\log\boldsymbol{P}_k^{(i)} \tag{6}$$

其中,$\boldsymbol{H}(\boldsymbol{S}_i)$表示第 $i$ 个样本的样本熵。式(6)是为了鼓励分配网络生成稀疏的聚类概率,以便最终可以将每个输入样本分配给主导的自动编码器。只有当$\boldsymbol{P}^{(i)}$是一个热向量时,式(6)才达到最小值 0。

当分配网络对所有输入样本都输出相同的概率值$P^{(i)}$时,混合自动编码器会得到局部最小值。为了解决这一问题,给定一个批次共 $N$ 个样本,我们添加一个批处理熵来鼓励所有自动编码器等概率调用。

$$\boldsymbol{H}(B)=-\sum_{i=1}^{K}\overline{\boldsymbol{P}}_k^{(i)}\log\overline{\boldsymbol{P}}_k^{(i)},\overline{\boldsymbol{P}}_k^{(i)}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{P}^{(i)} \tag{7}$$

其中,$\overline{\boldsymbol{P}}^{(i)}$是样本 $i$ 在整个批次样本上的平均聚类概率,$\boldsymbol{H}(B)$是整个批次的批处理熵。式(7)的最大值为 $\log(K)$,此时$\overline{\boldsymbol{P}}^{(i)}=\frac{1}{k}$。这表明每个样本被分配给混合自动编码器的概率是相等的。

### 3.5 预测网络

考虑到联合训练混合自动编码器和 GMM 的性能受限于一个过于简单的 GMM,其无法处理复杂数据结构的密度估计任务,本文提出一个预测网络来生成输入样本的混合部件概率,并将其用于计算 GMM 的参数,从而不需要使用期望最大化(Expectation Maximization,EM)[28] 等交替算法。此外,与传统方法仅利用重构误差进行异常分析不同,本文通过混合低维表示来执行异常分析,不仅考虑了串联的低维空间,而且考虑了重构特征向量。给定样本 $i$ 的混合低维表示$Z_i$和 GMM 中混合部件的数量 $G$,预测网络接收混合低维表示并预测样本的混合部件概率。

$$\boldsymbol{Z}_r^{(i)}=f(\boldsymbol{X}^{(i)},\widetilde{\boldsymbol{X}}^{(i)}) \tag{8}$$

$$\boldsymbol{Z}_i=[\boldsymbol{Z}_c^{(i)},\boldsymbol{Z}_r^{(i)}] \tag{9}$$

$$\boldsymbol{\gamma}_i=h(\boldsymbol{Z}_i,\boldsymbol{\theta}_p,G) \tag{10}$$

其中,$\boldsymbol{Z}_r^{(i)}$是由样本 $i$ 的重构误差得到的重构特征向量;$f(\cdot)$表示用于计算混合重构特征向量的函数,包括绝对欧氏距离、相对欧氏距离、余弦相似度等;$\boldsymbol{Z}_i$是混合低维表示,包括第 $i$ 个串联的低维空间$\boldsymbol{Z}_c^{(i)}$和第 $i$ 个混合重构特征向量$\boldsymbol{Z}_r^{(i)}$;$\boldsymbol{\gamma}_i$是由多层神经网络$h(\cdot)$学习到的第 $i$ 个混合部件概率,其最后一层是 softmax 层,参数是$\theta_p$。给定训练时一个 batch size 的样本个数 $N$ 和对应的学习到的混合部件概率 $\boldsymbol{\gamma}$,$\forall 1\leqslant g\leqslant G$,可以计算 GMM 的 3 个参数如下:

$$\boldsymbol{N}_g=\sum_{i=1}^{N}\boldsymbol{\gamma}_{ig} \tag{11}$$

$$\boldsymbol{\mu}_g=\frac{1}{\boldsymbol{N}_g}\sum_{i=1}^{N}\boldsymbol{\gamma}_{ig}\boldsymbol{Z}_i \tag{12}$$

$$\boldsymbol{\Sigma}_g=\frac{1}{\boldsymbol{N}_g}\sum_{i=1}^{N}\boldsymbol{\gamma}_{ig}(\boldsymbol{Z}_i-\boldsymbol{\mu}_g)(\boldsymbol{Z}_i-\boldsymbol{\mu}_g)^{\mathrm{T}} \tag{13}$$

$$\boldsymbol{\alpha}_g=\frac{\boldsymbol{N}_g}{N} \tag{14}$$

其中，$\boldsymbol{\gamma}_{ig}$ 表示基于混合低维表示 $Z$ 的第 $g$ 个混合部件生成的第 $i$ 个样本的概率，$\boldsymbol{\mu}_g$ 和 $\boldsymbol{\Sigma}_g$ 是 *GMM* 中混合部件 $g$ 的均值和协方差，$\boldsymbol{\alpha}_g$ 是 GMM 中第 $g$ 个混合部件的权重因子的混合参数。

通过学习得到的 GMM 参数，可以进一步计算如下概率密度函数：

$$\begin{aligned}\boldsymbol{P}(\boldsymbol{Z}_i\mid\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\boldsymbol{\alpha}_g)&=\sum_{g=1}^{G}\boldsymbol{\alpha}_g\mathcal{N}(\boldsymbol{Z}_i\mid\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\\&=\sum_{g=1}^{G}\boldsymbol{\alpha}_g\frac{1}{\sqrt{(2\pi)^D\mid\boldsymbol{\Sigma}\mid}}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{Z}_i-\boldsymbol{\mu}_g)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z}_i-\boldsymbol{\mu}_g)}\end{aligned} \tag{15}$$

其中，$D$ 是混合低维表示 $\boldsymbol{Z}$ 的维数，$\boldsymbol{\Sigma}\in\mathbb{R}^{D\times D}$，$\mid\boldsymbol{\Sigma}\mid$ 表示矩阵 $\boldsymbol{\Sigma}$ 的行列式，$\boldsymbol{\Sigma}^{-1}$ 表示矩阵 $\boldsymbol{\Sigma}$ 的逆矩阵。通过学习得到的概率密度 $\boldsymbol{P}(\boldsymbol{Z}_i)$，可以进一步计算输入样本的能量为：

$$\begin{aligned}\boldsymbol{E}(\boldsymbol{Z})&=-L(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\boldsymbol{\alpha}_g)\\&=-\sum_{i=1}^{N}\log\boldsymbol{P}(\boldsymbol{Z}_i\mid\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\boldsymbol{\alpha}_g)\end{aligned} \tag{16}$$

我们使用最大似然函数来计算参数 $\boldsymbol{\mu}_g$，$\boldsymbol{\Sigma}_g$，$\boldsymbol{\alpha}_g$ 的值，$\boldsymbol{E}(\boldsymbol{Z})$ 表示一个批次中 $N$ 个样本的样本能量和。

在测试阶段，因为 GMM 的参数已经被训练完成，我们可以直接预测样本能量，并推断出比预先设定的阈值更高的样本点是离群值。

### 3.6 复合目标函数

$\theta=(\theta_{e1},\theta_{d1},\cdots,\theta_{ek},\theta_{dk},\theta_m,\theta_p)$ 是混合自动编码器、分配网络和预测网络的参数。给定一个批次 $N$ 个样本，最小化复合目标函数如下：

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta})=&\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{L}(\boldsymbol{X}_i,\widetilde{\boldsymbol{X}}_i)+\alpha\boldsymbol{H}(\boldsymbol{S}_i))+\frac{\lambda_1}{N}\boldsymbol{E}(\boldsymbol{Z})+\\&\lambda_2\boldsymbol{P}(\boldsymbol{\Sigma})-\beta\boldsymbol{H}(B)\end{aligned} \tag{17}$$

该复合目标函数包括 5 个部分；

(1) $L(\boldsymbol{X}_i,\widetilde{\boldsymbol{X}}_i)$ 是重构误差。直觉上，当 $\boldsymbol{P}^{(i)}$ 是热向量时，重构误差会达到最小值。当混合自动编码器的重构误差达到最小时，混合低维表示能更好地保留输入样本的关键信息。

(2) $\boldsymbol{H}(\boldsymbol{S}_i)$ 是样本熵。通过最小化样本熵，得到热向量 $\boldsymbol{P}^{(i)}$。

(3) $\boldsymbol{E}(\boldsymbol{Z})$ 为一个批次 $N$ 个样本的样本能量之和。通过最小化样本能量，可以得到输入样本的最大能量。

(4) $\boldsymbol{P}(\boldsymbol{\Sigma})$ 用来避免 GMM 的平凡解，即防止协方差矩阵对角线元素变成 0。我们在协方差矩阵的对角线上加上一个较小值 $\boldsymbol{P}(\boldsymbol{\Sigma})=\sum_{k=1}^{K}\sum_{i=1}^{D}1/\boldsymbol{\Sigma}_k^{(ii)}$，其中 $\boldsymbol{\Sigma}\in\mathbb{R}^{D\times D}$。

(5) $\boldsymbol{H}(B)$ 是批处理熵。通过最小化批处理熵，可以平均地将每个样本分配给所有的自动编码器。

$\lambda_1$，$\lambda_2$，$\alpha$，$\beta$ 是复合目标函数的参数，这 4 个参数会显著影响密度估计任务的质量。

直观地，样本熵、批处理熵应该被首先考虑，以便平等地使用所有的自编码器，同时避免自编码器对每个输入样本进行同等优化。接着，需要联合最小化样本熵，以确保每个输入样本最终只分配给一个自编码器。进一步，通过最小化重构误差来获得包含输入样本中大部分关键信息的混合低维表示。最后，最小化协方差矩阵，以避免 GMM 的平凡解。

## 4 实验结果和分析

### 4.1 实验数据

本文使用来自 UCI 资料库和 ODDS 资料库的几个真实的公共基准数据集。表 1 列出了这些数据集的关键统计信息。此外，这些数据集中的每个属性列都会被规范化到[0,1]范围内。

表 1 公共基准数据集的统计信息

Table 1 Statistics of public benchmark data sets

| Dataset | Instances | Dimensions | Outliers |
|---|---|---|---|
| Kddcup99 | 494021 | 120 | 97278(20%) |
| Thyroid | 3772 | 6 | 93(2.5%) |
| Arrhythmia | 452 | 274 | 66(15%) |

Kddcup99 数据集是来自 UCI 资料库的网络入侵数据集。该数据集中的样本属性有 41 个维度，其中 34 个是连续属性，7 个是分类属性。对于分类属性，我们进一步使用 one-hot 编码对其进行编码，最终得到 120 维的数据集。在这个任务中，标记为“normal”的 20%的数据点作为离群点，其余标记为“attack”的数据点作为内联点(正常数据)。

甲状腺数据集是 ODDS 资料库中的一个分类数据集。它有 15 个分类属性和 6 个连续属性。对于异常检测，我们只使用了 6 个连续属性。在这个任务中，hyperfunction 类被视为离群点，其他两个类被视为内联点，因为 hyperfunction 显然是一个少数类。

心律失常数据集是 ODDS 资料库中的一个 280 维的多类分类数据集。删除 6 个分类属性后，还剩下 274 个属性。最小的类，即类 3,4,5,7,8,9,14,15 被合并为离群点，其余的类被合并为内联点。

### 4.2 基准方法

本节比较了本文所提方法与其他几种基准方法，即传统的和最先进的深度学习方法。OCSVM[24]是一种通过构造决策边界来进行异常检测的常用方法。我们使用径向基函数(RBF)核进行实验。DCN[28]是目前最先进的聚类方法。异常检测是通过判断样本与聚类中心的距离来完成的，离聚类中心越远的样本越有可能是离群点。深度结构能量基模型(Deep Structured Energy Based Model，DSEBM-r)[19]是以重构误差为异常检测准则的最先进的无监督学习方法。DAGMM[9]是最先进的用于无监督异常检测的深度学习方法，其利用深度自编码器和 GMM 来执行密度估计任务，并将能量高于预设阈值的样本作为离群点。

### 4.3 模型配置

在实验中，选择相对欧氏距离和余弦相似度作为最终的混

合重构特征向量。给定样本 $X$ 和聚合重构向量 $\widetilde{\boldsymbol{X}}$，相对欧氏距离表示为 $\frac{\|X-\widetilde{\boldsymbol{X}}\|_2}{\|X\|_2}$，余弦相似度表示为 $\frac{X\cdot\widetilde{\boldsymbol{X}}}{\|X\|_2\|\widetilde{\boldsymbol{X}}\|_2}$。

在 3 个数据集上 MAGMM 模型的结构细节总结如下。

在 Kddcup99 数据集上，混合自动编码器包括 4 个结构相同但初始权值不同的自编码器，预测网络采用 6 维混合低维表示，其中 4 维来自于混合自动编码器的串联的低维空间，另外 2 维由重构误差计算得到的重构特征向量得到，并输出混合分量概率，该概率由一个带有 4 个混合分量的 GMM 组成。实验中设置 $\lambda_1=0.01$，$\lambda_2=0.0005$，$\alpha=0.06$，$\beta=0.06$。MAGMM 模型结构如图 2 所示。

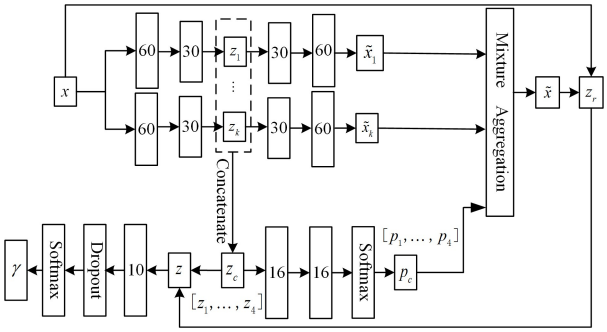



图 2 Kddcup99 数据集上 MAGMM 的结构

Fig. 2 Structure of MAGMM on Kddcup99 dataset

在甲状腺数据集上，混合自动编码器提供了 2 个自编码器，其结构为 FC(6,12,tanh)-FC(12,4,tanh)-FC(4,1,none)-FC(1,4,none)-FC(4,12,tanh)-FC(12,6,none)。分配网络的结构为 FC(2,10,tanh)-FC(10,10,tanh)-FC(10,2,softmax)。预测网络采用了 4 维混合低维表示，其中 2 维来自于自动编码器的串联的低维空间，另外 2 维是由重构误差计算得到的重构特征向量得到，并输出混合分量概率，该概率由一个带有 2 个混合分量的 GMM 组成，其结构为 FC(4,10,tanh)-Dropout(0.5)-FC(10,2,softmax)。实验中设置 $\lambda_1=0.05$，$\lambda_2=0.0005$，$\alpha=1$，$\beta=0.06$。

在心律失常数据集上，混合自动编码器提供了 6 个自编码器，其结构为 FC(274,10,tanh)-FC(10,1,none) FC(1,10,tanh)-FC(10,274,none)。分配网络的结构为 FC(6,16,tanh)-FC(16,16,tanh)-FC(16,6,softmax)。预测网络采用 8 维混合低维表示，其中 6 维来自于混合自动编码器的串联的低维空间，另外 2 维由重构误差计算得到的重构特征向量得到，并输出混合分量概率，该概率由一个带有 7 个混合分量的 GMM 组成，其结构为 FC(8,10,tanh)-Dropout(0.5)-FC(10,7,softmax)。实验中设置 $\lambda_1=0.05$，$\lambda_2=0.00005$，$\alpha=0.1$，$\beta=0.1$。

其中 FC($a$,$b$,$f$)表示输入为 $a$ 个神经元，输出为 $b$ 个神经元且激活函数为 $f$ 的全连接层。None 表示不使用激活函数，Dropout($p$)表示在训练阶段，每个批次都会随机移除每层中 $p$ 百分比的神经元。

### 4.4 检测结果

根据表 1，3 个数据集有不同的异常值比率，因此根据不同的数据集，我们选择相对应的阈值来识别测试集中的异常样本。Kddcup99、甲状腺和心律失常数据集上的阈值分别取 80，97.5 和 85 百位数，即整个数据集中样本能量的前 20%，2.5%和 15%的样本将被认定为异常值。

表 2 列出了 MAGMM 和其他异常检测基准方法的性能指标，其中最好的结果以黑体突出显示。

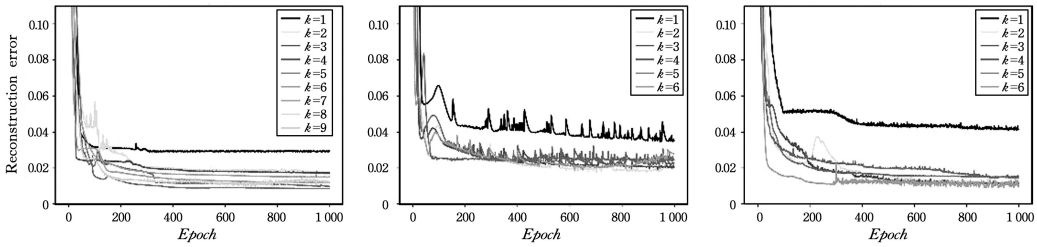
表 2 MAGMM 与当前最新的方法相比，精度、召回率和 F1 得分的结果(对于每个度量，最好的结果以黑体突出显示)

Table 2 Compared with state-of-the-art methods, precision, recall and F1 score results

| Dataset | | OCSVM | DSEBM-r | DSEBM-e | DCN | DAGMM | MAGMM |
|---|---|---|---|---|---|---|---|
| Kddcup99 | precision | 0.7457 | 0.1972 | 0.7369 | 0.7696 | 0.9297 | **0.9318** |
| | recall | 0.8523 | 0.2001 | 0.7477 | 0.7829 | **0.9442** | 0.8531 |
| | $F_1$ | 0.7954 | 0.1987 | 0.7423 | 0.7762 | **0.9369** | 0.8907 |
| Thyroid | precision | 0.3639 | 0.0404 | 0.1319 | 0.3319 | 0.4766 | **0.8250** |
| | recall | 0.4239 | 0.0403 | 0.1319 | 0.3169 | 0.4834 | **0.7204** |
| | $F_1$ | 0.3887 | 0.0403 | 0.1319 | 0.3251 | 0.4782 | **0.7691** |
| Arrhythmia | precision | 0.5397 | 0.1515 | 0.4667 | 0.3758 | 0.4909 | **0.6863** |
| | recall | 0.4082 | 0.1513 | 0.4565 | 0.3907 | 0.5078 | **0.5903** |
| | $F_1$ | 0.4581 | 0.1510 | 0.4601 | 0.3815 | 0.4983 | **0.6347** |

可以看到，在大多数情况下，本文方法比其他方法表现得更好。特别是在甲状腺和心律失常两个数据集上，MAGMM 在 F1 得分上分别比 DAGMM 提高了 29%和 13%，验证了该方法的有效性。OCSVM 的性能会受到数据维数的影响，过多的无关特征会影响 OCSVM 的建模能力。DSEBM-r 在所有数据集上的性能都非常糟糕，这说明基于重构误差的异常检测方法中没有包含足够的信息来执行异常检测任务。DSEBM-e、DAGMM 和 MAGMM 均使用样本能量进行异常检测，而 DSEBM-e 的性能不如其他两种方法。因为 DSEBM-e 的能量建模能力无法与其他两种方法相比。DCN 的性能会受到预先训练好的自动编码器的限制。由于已经训练好的自动编码器不知道后续的密度估计任务，因此训练好的自动编码器可能会去除异常检测的关键信息。DAGMM 在 Kddcup99 数据集上检测精度结果排名第一，而在其他两个数据集上检测结果不佳。一种可能的原因是，如果数据集的样本量与 Kddcup99 数据集的一样大，那么仅由一个深度自编码器得到的低维潜在表示就包含了足够的关键信息，可以进行异常检测；但是当数据集的样本数量不大时，混合自动编码器可以获得比单一深度自编码器更低的重构误差，因而会促进后续密度估计任务的完成，得到更好的性能表现。图 3 显示

了 Kddcup99、甲状腺、心律失常数据集上 $K$ 值变化时的重构误差。结果表明，当数据集中的样本数量不超过 Kddcup99 数据集时，混合自动编码器产生的低维潜在表示将获得更多的关键信息，从而比单一深度自编码器产生更低的重构误差。




(a) Kddcupp99 (b) Thyroid (c) Arrhythmia

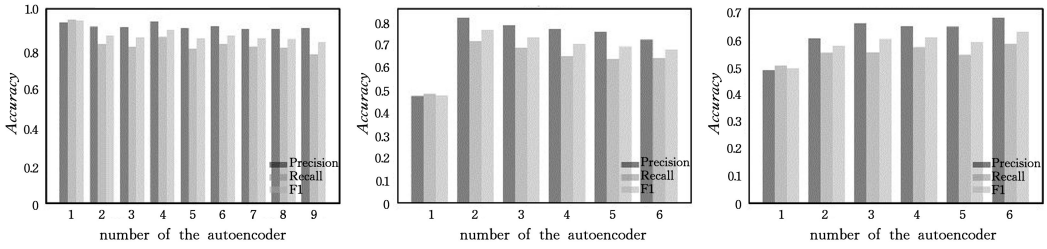
图 3 Kddcup99、甲状腺、心律失常数据集上混合自动编码器的产生的重构误差随 $K$ 值变化的情况

Fig. 3 Reconstruction errors derive from the mixed auto-encoders vary with $K$ on Kddcup99, Thyroid and Arrhythmia datasets

### 4.5 参数敏感性

本节对模型的关键参数的敏感性进行了测试。

自编码器的数量：本文分析了参数对于混合自动编码器中自编码器数量的敏感性。我们通过改变自动编码器的数量来展示本文方法的效果。从图 4 可以看出，MAGMM 对自动编码器的数量不是很敏感，其通过在甲状腺和心律失常数据集上使用混合自动编码器，得到了相对较大的性能提高。




(a) Kddcupp99 (b) Thyroid (c) Arrhythmia

图 4 随着混合自动编码器中 $K$ 值的改变，3 个数据集上精度、召回率、F1 得分的结果

Fig. 4 Results for precision, recall and F1 scores on three data sets as $K$ value changes in mixed auto-encoder

复合目标函数参数：本文分析复合目标函数中参数 $\lambda_1$，$\lambda_2$，$\alpha$，$\beta$ 的敏感性。如式(17)所示，复合目标函数包括 5 个部分：由混合聚集得到的混合重构误差函数、由分配网络得到的样本熵、由预测网络得到的样本能量、由协方差矩阵得到的惩罚函数、由分配网络得到的批次熵函数。这 5 部分的系数率可以表示为 $1:\lambda_1:\lambda_2:\alpha:\beta$。对于 $\lambda_1$，一个较大的值可能会减少反向传播时混合自动编码器的重构误差的重要程度，这样就无法获得预期的输入样本的低维表示，而一个较小的值可能不利于密度估计任务。对于 $\lambda_2$，一个较大的值会导致协方差比重过大，从而导致样本能量太大而把正常样本错归为异常，而一个较小的值可能不足以解决平凡解的问题。对于 $\alpha$，一个较大的值会导致所有输入样本被分配到同一个自编码器中，这样我们会获得局部最小值，而一个较小的值会导致每个输入样本最终被分配给一个主导的自编码器。对于 $\beta$，一个较大的值会导致输入样本不能分配给一个主导的自编码器，而一个较小的值可能会使目标函数失去对自动编码器的约束力，这会鼓励输入样本平等地使用所有的自编码器。

在实验中我们发现，系数率为 $1:0.05:0.0005:1:0.06$ 时，在甲状腺数据集上通常会得到满意的结果。为了测试所提方法对系数率的敏感性，我们将基数从 1 调整到 5。表 3 列出了在甲状腺数据集上，1 000 次迭代后的精度、召回率和 F1 得分的平均值。可以看出，$\lambda_1$，$\lambda_2$，$\alpha$，$\beta$ 在系数率基数的变化上不敏感。

表 3 甲状腺数据集上 $\lambda_1$，$\lambda_2$，$\alpha$ 和 $\beta$ 的固定比率为 $1:0.05:0.0005:1:0.06$ 的 MAGMM 的敏感性

Table 3 Sensitivity of MAGMM for $\lambda_1$, $\lambda_2$, $\alpha$ and $\beta$ with a fixed ratio of $1:0.05:0.0005:1:0.06$ on thyroid dataset

| Base | precision | recall | $F_1$ |
|---|---|---|---|
| 1 | 0.8245 | 0.7215 | 0.7696 |
| 2 | 0.8252 | 0.7208 | 0.7692 |
| 3 | 0.8221 | 0.7204 | 0.7679 |
| 4 | 0.8286 | 0.7184 | 0.7696 |
| 5 | 0.8204 | 0.7241 | 0.7692 |

**结束语** 本文提出了一种用于无监督异常检测的混合自动编码高斯混合模型(MAGMM)，其利用混合自动编码器将样本投射到串联的低维表示中，保留了数据集中每个簇足够的关键信息，从而解决了仅使用单一深度自编码器的缺点。为了使每个样本被平均分配给每个自编码器并最终分配给一个占据主要地位的自编码器，本文提出了分配网络。在此基础上，我们提出了预测网络模型来执行密度估计任务，其接收串联的低维表示和重构误差并输出混合分量概率。MAGMM 通过联合训练混合自动编码器、分配网络和预测网络，可以避免局部最优，降低重构误差，提高表示能力，从而间接促进密度估计任务的完成。在几个公共基准数据集上的实验结果表明，该方法显著优于 DAGMM，其 F1 得分相比基准方法的 F1 得分提高了 29%。虽然混合自动编码器可以取得比单一自编码器更低的重构误差，但是自编码器个数过多会导致协方差矩阵的计算复杂度上升，并且容易导致矩阵行列

式趋于无穷小,从而导致出现平凡解的问题。在未来的工作中,我们将进一步讨论混合自动编码器与高斯混合模型集成的复杂度的优化问题。

## 参 考 文 献


[1] HUANG D,MU D,YANG L,et al. CoDetect:financial fraud detection with anomaly feature detection[J]. IEEE Access,2018,6:19161-19174.

[2] VIEGAS E,SANTIN A,BESSANI A,et al. BigFlow:Real-time and reliable anomaly-based intrusion detection for high-speed networks[J]. Future Generation Computer Systems,2019,93:473-485.

[3] SANEJA B,RANI R. An efficient approach for outlier detection in big sensor data of health care[J]. International Journal of Communication Systems,2017,30(17):e3352.

[4] CHEN Z,HUANG Y,ZOU H. Anomaly Detection of Industrial Control System Based on Outlier Mining[J]. Computer Science,2014,41(5):178-181.

[5] ZIMEK A,SCHUBERT E,KRIEGEL H P. A survey on unsupervised outlier detection in high dimensional numerical data[J]. Statistical Analysis and Data Mining:The ASA Data Science Journal,2012,5(5):363-387.

[6] RADOVANOVI M,NANOPOULOS A,IVANOVI M. Reverse nearest neighbors in unsupervised distance-based outlier detection[J]. IEEE Transactions on Knowledge and Data Engineering,2014,27(5):1369-1382.

[7] YANG B,FU X,SIDIROPOULOS N D,et al. Towards k-means-friendly spaces:Simultaneous deep learning and clustering[C]// Proceedings of the 34th International Conference on Machine Learning. 2017:3861-3870.

[8] CANDÈS E J,LI X,MA Y,et al. Robust principal component analysis? [J]. Journal of the ACM,2011,58(3):1-37.

[9] ZONG B,SONG Q,MIN M R,et al. Deep autoencoding gaussian mixture model for unsupervised anomaly detection[C]// International Conference on Learning Representations. 2018:781-795.

[10] EHSAN A M,DICK A,VAN D H A. Infinite variational autoencoder for semi-supervised learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:5888-5897.

[11] ZHANG D,SUN Y,ERIKSSON B,et al. Deep unsupervised clustering using mixture of autoencoders[J]. arXiv:1712.07788,2017.

[12] CHANDOLA V,BANERJEE A,KUMAR V. Anomaly detection:A survey[J]. ACM Computing Surveys (CSUR),2009,41(3):15.1-15.58.

[13] AGGARWAL C C. Outlier analysis[C]// Data mining. Springer,Cham,2015:237-263.

[14] WU J F,JIN Y D,TANG P. Survey on Monitoring Techniques for Data Abnormalities[J]. Computer Science,2017,44(Z11):24-28.

[15] JOLLIFFE I. Principal component analysis[M]. Berlin Heidelberg:Springer,2011.

[16] SCHÖLKOPF B,SMOLA A,Müller K R. Kernel principal component analysis[C]// International conference on artificial neural networks. Berlin,Heidelberg:Springer,1997:583-588.

[17] XIA Y,CAO X,WEN F,et al. Learning discriminative reconstructions for unsupervised outlier removal[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:1511-1519.

[18] AN J,CHO S. Variational autoencoder based anomaly detection using reconstruction probability[J]. Special Lecture on IE,2015,2(1):216-234.

[19] ZHAI S,CHENG Y,LU W,et al. Deep structured energy based models for anomaly detection[J]. arXiv:1605.07717,2016.

[20] ZHOU C,PAFFENROTH R C. Anomaly detection with robust deep autoencoders[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2017:665-674.

[21] DUDA R O,HART P E,STORK D G. Pattern classification[M]. John Wiley & Sons,2012.

[22] BISHOP C M. Neural networks for pattern recognition[M]. Oxford University Press,1995.

[23] YANG X,HUANG K,GOULERMAS J Y,et al. Joint learning of unsupervised dimensionality reduction and gaussian mixture model[J]. Neural Processing Letters,2017,45(3):791-806.

[24] SCHÖLKOPF B,PLATT J C,SHAWE T J,et al. Estimating the support of a high-dimensional distribution[J]. Neural computation,2001,13(7):1443-1471.

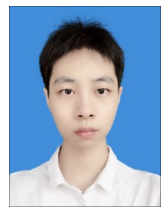
[25] TAX D M J,DUIN R P W. Support vector data description[J]. Machine learning,2004,54(1):45-66.

[26] YANG X,HUANG K,ZHANG R. Unsupervised dimensionality reduction for gaussian mixture model[C]// International Conference on Neural Information Processing. Springer,Cham,2014:84-92.

[27] TÜSKE Z,TAHIR M A,SCHLÜTER R,et al. Integrating Gaussian mixtures into deep neural networks:Softmax layer with hidden variables[C]// 2015 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). IEEE,2015:4285-4289.

[28] HUBER P J. Robust statistics[M]. Berlin,Heidelberg:Springer,2011.


**ZOU Cheng-ming**, born in 1975, Ph.D, professor, is a member of China Computer Federation. His main research interests include computer vision, embedded system, software theory and method.



**CHEN De**, born in 1995, postgraduate. His main research interests include deep learning, data mining and so on.