第46卷 第4期 2022年2月25日
DOI：10.7500/AEPS20201118006
电力系统自动化
Automation of Electric Power Systems
Vol. 46 No. 4 Feb. 25，2022

基于变分自编码高斯混合模型的发电企业串谋智能预警

华回春[1]，邓 彬[1]，刘 哲[2]，张立峰[1]
（1. 新能源电力系统国家重点实验室（华北电力大学），河北省保定市 071003；2. 国网上海市电力公司，上海市 200437）

摘要： 随着市场交易规模越来越大，交易数据量增加，结合数据进行串谋分析成为可能。为此，结合发电企业的串谋预警指标体系和无监督的变分自编码高斯混合模型（VAEGMM），实现了对发电企业串谋的智能预警。首先，提出了完善的串谋预警指标体系和详细的指标计算方法。其次，针对指标集具有高维且正负样本不均衡的数据特点，结合异常检测思想提出了VAEGMM。然后，详细阐述了VAEGMM的网络结构，并且重新构建了联合损失函数，使得该网络能够更好地学习得到原始数据的低维表达，从而有助于进行更准确的密度估计。最后，实例测算表明，与其他传统的无监督学习模型相比较，VAEGMM可以更加高效和准确地预警串谋风险。

关键词： 电力市场；发电企业；智能预警；串谋；变分自编码高斯混合模型

0 引言

在电力市场中，串谋是滥用市场力违规的主要表现形式之一[1]。通常，集中竞价中串谋企业通过平行报价和反向报价来获取更多的利润[2]。传统评估发电企业使用市场力进行串谋的方法主要有场前指标分析法[3-5]和事后检查法[6-8]。目前，中国电力市场力高度集中，赫芬达尔-赫希曼指数（Herfindahl-Hirschman index，HHI）明显高于1 800阈值，串谋更为常见[9]。因此，需要加强电力市场信用监管的顶层设计，实时监测发电企业的串谋行为，对信用风险提前做出预警。

目前，发电企业串谋的研究大致可以分为2类：一类是根据市场交易规则研究串谋发生的机理[10-12]；另一类是根据市场交易数据构建串谋指标进行预警[13-14]。前者使用博弈模型和回归模型等对电力市场交易进行建模，并通过企业的竞价策略来判断串谋风险。文献[15-16]还将强化学习应用于串谋风险的评估中。但是，由于电力市场的信息披露机制还不完善，该类方法建模所需的保密数据，如发电企业的边际成本，难以获得。后者则通过串谋指标特征直接反映串谋成员的异常报价行为并由专家进一步决策，更加适合监管机构进行串谋预警。随着电力市场逐步开放，市场交易规模不断增大，传统的专家决策已无法满足工作需要。因此，需要提出能够实时甄别发电企业串谋指标特征的智能预警模型，为相关监管机构提供有串谋嫌疑的企业名单。

文献[17]构建了发电企业的指标特征库，并采用模式识别来检测违规行为，但该类方法不能自我更新，无法实时适应市场的变化。文献[18]将串谋视为二分类问题，采用有监督学习算法训练分类器，从而区分串谋指标特征。但在实际操作中，带有标注的数据非常少，因而难以训练出有良好泛化能力的有监督学习模型，为此采用无监督学习模型是一种更好的选择。由于串谋实际上是电力市场中的异常行为，可以结合无监督学习的异常检测思想筛查出串谋的指标特征。

传统的异常检测方法有三大类，分别是基于误差重构的方法[19-22]、基于聚类分析的方法[23-24]以及基于分类的方法[25-27]。但是，传统异常检测方法难以处理电力市场交易的复杂高维数据，而无监督的深度联合学习网络在处理复杂数据的时候有优势。该网络一般由2个部分组成：表达网络和估计网络。在每一次迭代中，前者学习得到高维复杂数据的潜变量和特征，后者对其进行密度估计，将位于低密度区域的样本视为异常样本[28-29]。考虑到电力市场数据具有正负样本不均衡（正常样本远多于异常样本）的特点，因此提出了新的深度联合网络。该网络在表达网络中采用变分自编码器（variational autoencoder，VAE）降维，可以学习得到具有一定噪声的隐变量，保留更多的串谋样本信息，而且在每一次迭代中，包含的正类样本信息占比更多，对串谋样本的识别帮助更大[30-31]。本文在构建串谋指标特征

收稿日期：2020-11-18；修回日期：2021-10-18。
上网日期：2021-12-01。
国家电网公司科技项目（SGSHDK00HZJS2000254）。

的基础上，采用无监督学习模型，即变分自编码高斯混合模型（variational autoencoding Gaussian mixture model，VAEGMM），实现了发电企业串谋的智能预警。

1　集中竞价中发电企业串谋预警指标体系

1.1　串谋预警指标

在集中竞价中，只有2家或者2家以上的发电企业才有可能产生串谋行为，所以串谋预警指标需要充分反映任意2家企业在竞价过程中的报价特点。根据市场内常见的串谋形式，参考了现有的串谋指标，构建了较为完善的指标体系。它们的名称、含义和具体计算公式如下[13-14]。

1）申报电量市场份额均值$x_{ij}^{(1)}$，表示2家发电企业的申报电量在市场份额的占比，计算公式如式（1）所示。

$$x_{ij}^{(1)}=\frac{S_i+S_j}{2\sum_{i=1}^{n}S_i}\times 100\% \tag{1}$$

式中：S_i和S_j分别为市场中第i家和第j家发电企业在本次竞价中的申报电量；n为在本次竞价中参与的发电企业总数。

该指标衡量了2家发电企业在本次竞价的市场力大小，指标数值越大的2家发电企业越有能力影响整个市场的价格，说明串谋的可能性越大。

2）报价一致性指标$x_{ij}^{(2)}$，表示2家发电企业的申报价格差异，计算公式如式（2）所示。

$$x_{ij}^{(2)}=\frac{\sum_{a=1}^{3}(\rho_{ia}-\rho_{ja})^2}{\sum_{a=1}^{3}(\rho_{ia}-\bar{\rho}_a)(\rho_{ja}-\bar{\rho}_a)} \tag{2}$$

式中：ρ_{ia}和ρ_{ja}分别为第i家和第j家发电企业在本次竞价中的第a段报价；$\bar{\rho}_a$为在本次竞价中所有发电企业第a段报价的均值。

该指标衡量了2家发电企业在同一时间段、同比例地改变申报电价的可能性大小。数值越小，表明存在串谋的嫌疑越大。

3）报量一致性指标$x_{ij}^{(3)}$，表示2家发电企业的申报电量差异，计算公式如式（3）所示。

$$x_{ij}^{(3)}=\frac{\sum_{a=1}^{3}(S_{ia}-S_{ja})^2}{\sum_{a=1}^{3}(S_{ia}-\bar{S}_a)(S_{ja}-\bar{S}_a)} \tag{3}$$

式中：S_{ia}和S_{ja}分别为第i家和第j家发电企业在本次竞价中的第a段申报电量；$\bar{S}_a$为在本次竞价中所有发电企业第a段申报电量的均值。

该指标衡量了2家发电企业在同一时间段、同比例地改变申报电量的可能性大小。数值越小，说明存在串谋的嫌疑越大。

4）报价曲线差异面积比率$x_{ij}^{(4)}$，反映了2家发电企业的平行竞价程度，计算公式如式（4）所示。

$$x_{ij}^{(4)}=\frac{\int_0^{S'}\left|f_i-f_j\right|\mathrm{d}s}{S'} \tag{4}$$

式中：f_i和f_j分别为第i家和第j家发电企业在本次竞价中的报价曲线函数；S'为2家发电企业申报电量的较小值；s为积分变量。

该指标衡量了2家发电企业在同一时间段、同比例地改变申报电量和电价的可能性大小。数值越小，说明存在串谋的嫌疑越大。

5）报价安全度均值$x_{ij}^{(5)}$，用于衡量发电企业之间报价与历史平均报价的偏离程度，计算公式如式（5）所示。

$$x_{ij}^{(5)}=\frac{\bar{\rho}_i+\bar{\rho}_j-2E}{2E}\times 100\% \tag{5}$$

式中：$\bar{\rho}_i$和$\bar{\rho}_j$分别为第i家和第j家发电企业在本次竞价中的加权平均申报价格；E为市场边际价格的期望值，可以用历史交易的边际价格计算得到。

该指标衡量了当前报价与历史报价的偏差，偏差越大，说明它们的报价越有可能脱离自身的发电能力，存在与其他企业串通、抬高市场出清价格的嫌疑。

6）报价相对比均值$x_{ij}^{(6)}$，表示2家发电企业的报价与本次集中竞价平均价格的区别，计算公式如式（6）所示。

$$x_{ij}^{(6)}=\frac{1}{2}\frac{\bar{\rho}_i+\bar{\rho}_j}{\frac{1}{n}\left(\sum_{i=1}^{n}\bar{\rho}_i\right)}\times 100\% \tag{6}$$

该指标衡量了当前报价与市场整体报价水平的偏差。偏差越大，说明发电企业试图报高价以抬高市场价格的可能性越大。

上述指标体系主要反映了串谋企业的市场力大小、平行报价和异常报价的行为特征。在供大于求的情况下，报价一致性、报量一致性和报价曲线差异面积比率指标能够反映发电企业通过平行报价来抢占市场份额的行为。在供求关系相对平衡的情况下，报价安全度均值和报价相对比均值这2个指标能够反映发电企业通过物理缩减和经济缩减等异常报价行为制造市场供不应求的情况，从而使其他成员获得提价的空间。尽管参与市场的不同类型发电企业成本有较大差异，但是上述指标体系依然有效。

1.2　串谋指标集测算方法

同时对多场集中竞价进行串谋预警时，需要对不同场次的竞价数据进行标注再糅合，直接糅合可能会使无监督学习模型无法正确判断串谋行为。例如，场次A和B都发生了不同类型的串谋，那么场次A和B中正常企业的竞价数据会因为各自的串谋类型而呈现出不同的数据特点，直接糅合多场次数据会让VAEGMM认为这些正常企业也存在串谋的嫌疑。因此，本文提出了新的标注指标，具体定义如下。

集中竞价场次$x_{ij}^{(7)}$，代表2家发电企业集中竞价的场次，如式(7)所示。

$$x_{ij}^{(7)}=l \qquad l=1,2,3,\cdots \tag{7}$$

式中：l为第i家和第j家发电企业参与集中竞价的场次。

该指标的选取可按照时间周期选取，如周、月、年等，或者按照变量选取，如同水平电价、进入市场的发电企业等。

假设某地同一时期的电力市场一共有L场集中竞价，参与某一场次竞价的发电企业有m家，那么串谋预警的指标集计算过程如下。

首先根据原始的报价数据，计算某一场次竞价中2家发电企业之间第k个指标的数据矩阵$\boldsymbol{R}^{(k)}$如式(8)所示。

$$\boldsymbol{R}^{(k)}=\begin{bmatrix} x_{11}^{(k)} & x_{12}^{(k)} & \cdots & x_{1m}^{(k)} \\ x_{21}^{(k)} & x_{22}^{(k)} & \cdots & x_{2m}^{(k)} \\ \vdots & \vdots & & \vdots \\ x_{m1}^{(k)} & x_{m2}^{(k)} & \cdots & x_{mm}^{(k)} \end{bmatrix} \tag{8}$$

将矩阵元素按照除对角线外的上三角形依次平铺，得到列向量$\boldsymbol{x}^{(k)}$，即为指标集的一列指标特征。

$$\boldsymbol{x}^{(k)}=[\,x_{12}^{(k)}\ \ x_{13}^{(k)}\ \cdots\ x_{1m}^{(k)}\ \ x_{23}^{(k)}\ \ x_{24}^{(k)}\ \cdots\ x_{2m}^{(k)}\ \cdots\ x_{m-1,m}^{(k)}\,]^{\mathrm{T}} \tag{9}$$

然后，按照式(1)至式(7)分别计算出7个指标特征，按列组合即可得到某一场次l集中竞价的指标集$\boldsymbol{X}_l$。

$$\boldsymbol{X}_l=[\,\boldsymbol{x}^{(1)}\ \ \boldsymbol{x}^{(2)}\ \ \cdots\ \ \boldsymbol{x}^{(7)}] \tag{10}$$

最后，将L场次的指标集$\boldsymbol{X}_l$按行组合得到用于网络训练的指标集$\boldsymbol{X}$。

$$\boldsymbol{X}=[\,\boldsymbol{X}_1\ \ \boldsymbol{X}_2\ \ \cdots\ \ \boldsymbol{X}_L\,]^{\mathrm{T}} \tag{11}$$

值得注意的是，在训练模型前，需要对标注指标(集中竞价场次)进行独热编码来消除量纲[32]，因此，指标集$\boldsymbol{X}$的维度会随场次增加呈线性增长。另外，指标集$\boldsymbol{X}$的每个样本都是由任意2家发电企业的竞价数据计算得到的指标结果，且串谋行为是一种少数违法行为。因此，指标集$\boldsymbol{X}$具有维数高且正负样本不均衡的数据特点(正常样本远多于串谋样本)。由于串谋样本的指标特征与大部分样本的指标特征相差较大，在样本空间中，串谋样本点表现为离群点，即异常样本点。针对此特点，本文提出了一种新的深度联合学习网络来实时甄别指标集$\boldsymbol{X}$中的异常样本。

2　串谋预警模型建立

2.1　网络结构

如图1所示，VAEGMM的网络结构由2个部分网络构成：表达网络和估计网络。其中，表达网络通过VAE对网络的输入进行降维，同时得到潜变量Z_l和重构概率Z_r。然后，将这2个特征整合起来作为估计网络的输入，使用高斯混合模型(Gaussian mixture model，GMM)计算得到每个样本在低维空间的密度估计。图2中：$\boldsymbol{X}'$为重构样本；$\boldsymbol{Z}=[\,Z_l, Z_r]$为整合变量；$\mu(\boldsymbol{X})$与$\sigma(\boldsymbol{X})$分别为均值和方差函数；$\hat{\boldsymbol{\pi}}$为整个网络的输出。

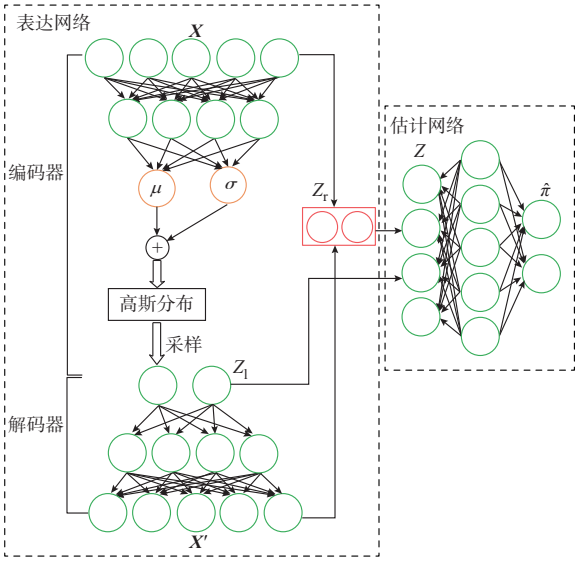

图1　VAEGMM网络结构
Fig. 1　Network structure of VAEGMM

2.2　表达网络

VAEGMM的表达网络使用VAE对指标集$\boldsymbol{X}$的指标特征进行降维重组，形成一个更可辨的低维样本空间。VAE包含编码器和解码器2个部分。

如图2所示，编码器的目标是学习潜变量的近似后验分布$q(Z_l|\boldsymbol{X})\sim N'(\mu'(\boldsymbol{X}),\sigma'(\boldsymbol{X}))$，其中，$N'(\mu'(\boldsymbol{X}),\sigma'(\boldsymbol{X}))$为分布函数，$\mu'(\boldsymbol{X})$与$\sigma'(\boldsymbol{X})$分别为重构均值和方差函数。首先，$\mu'(\boldsymbol{X})$与$\sigma'(\boldsymbol{X})$需要通过网络学习得到；然后，从后验分布$q(Z_l|\boldsymbol{X})$中采样得到原始样本的潜变量$Z_l$。由于$\sigma'(\boldsymbol{X})$不为0，$Z_l$

带有一定的噪声，体现了潜在变量空间的可变性。也就是说在每一次迭代中，相较于其他降维，VAEGMM所学习得到的潜变量都具有更加丰富的正类信息，从而有助于估计网络将负类样本从低密度区域中识别出来。

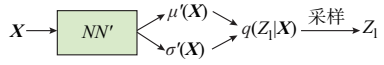

图2　VAE编码器
Fig. 2　Encoder of VAE

如图3所示，解码器通过近似后验分布$p(\boldsymbol{X}'|Z)$采样，对潜变量Z_{l}进行重建得到重构样本$\boldsymbol{X}'$。其中，$\boldsymbol{X}'|Z_{\mathrm{l}}\sim N(\mu(\boldsymbol{X}),\sigma(\boldsymbol{X}))$，$N(\mu(\boldsymbol{X}),\sigma(\boldsymbol{X}))$为分布函数，函数$\mu(\boldsymbol{X})$与$\sigma(\boldsymbol{X})$也需要通过网络学习得到。通过函数$g(\boldsymbol{X},\boldsymbol{X}')$的计算可以得到重构样本和输入样本的重构概率$Z_{\mathrm{r}}$。该特征不同于普通降维网络的重构误差，它不仅将重构样本与原始输入之间的差异考虑在内，而且还考虑了由近似后验分布$p(\boldsymbol{X}'|Z_{\mathrm{l}})$的方差$\sigma(\boldsymbol{X})$来重建$\boldsymbol{X}'$的可变性。该特征反映了不同样本的方差灵敏度，灵敏度高的样本能通过高方差重构被视为正常样本，从而降低重构概率。

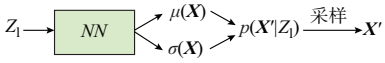

图3　VAE解码器
Fig. 3　Decoder of VAE

2.3　估计网络

通过表达网络的计算，VAEGMM将原始变量的潜变量Z_{l}和与重构样本的重建概率Z_{r}整合起来馈入估计网络中。由于大量随机变量的累计分布收敛于高斯混合分布，VAEGMM选取GMM作为估计网络。

$$\hat{\boldsymbol{\pi}}=\mathrm{soft\,max}(h(\boldsymbol{Z},\boldsymbol{\theta}_{\mathrm{m}})) \tag{12}$$

式中：$\hat{\boldsymbol{\pi}}$为整个网络的输出，其元素取1或0，是原始变量的整合变量$\boldsymbol{Z}$的密度估计；$\mathrm{soft\,max}(\cdot)$为激活函数；$\boldsymbol{\theta}_{\mathrm{m}}$为估计网络的参数向量；$h(\boldsymbol{Z},\boldsymbol{\theta}_{\mathrm{m}})$为隐含层输出。

假设$\hat{\boldsymbol{\pi}}$有K维特征，那么可以得到GMM的参数[28]：

$$\hat{\varphi}_k=\sum_{w=1}^{W}\frac{\hat{\pi}_{wk}}{W} \tag{13}$$

$$\hat{\boldsymbol{\mu}}_k=\frac{\sum_{w=1}^{W}\hat{\pi}_{wk}\boldsymbol{Z}_w}{\sum_{w=1}^{W}\hat{\pi}_{wk}} \tag{14}$$

$$\hat{\boldsymbol{\Sigma}}_k=\frac{\sum_{w=1}^{W}\hat{\pi}_{wk}(\boldsymbol{Z}_w-\hat{\boldsymbol{\mu}}_k)(\boldsymbol{Z}_w-\hat{\boldsymbol{\mu}}_k)^{\mathrm{T}}}{\sum_{w=1}^{W}\hat{\pi}_{wk}} \tag{15}$$

式中：$\hat{\varphi}_k$、$\hat{\boldsymbol{\mu}}_k$、$\hat{\boldsymbol{\Sigma}}_k$分别为第$k$维特征的加权概率、期望和方差；$W$为样本总数；$\hat{\pi}_{wk}$为第$k$个维度整个网络第$w$个样本的输出；$\boldsymbol{Z}_w$为第$w$个样本的整合变量。

进一步，可以推出样本能量$E(\boldsymbol{Z})$的定义如下：

$$E(\boldsymbol{Z})=-\ln\left(\sum_{k=1}^{K}\hat{\varphi}_k\frac{\exp\left(-\frac{1}{2}(\boldsymbol{Z}-\hat{\boldsymbol{\mu}}_k)^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_k^{-1}(\boldsymbol{Z}-\hat{\boldsymbol{\mu}}_k)\right)}{\sqrt{2\pi\left|\hat{\boldsymbol{\Sigma}}_k\right|}}\right) \tag{16}$$

一般，通过GMM的估计，将拥有高能量的样本视为异常样本。网络输出$\hat{\pi}$是每个样本服从于整个数据集近似样本分布的可能性。它也是两两发电企业的竞价数据与市场总体水平的偏离程度，可认为是串谋嫌疑度。

2.4　联合损失函数

VAEGMM是一种深度联合学习模型，通过联合损失函数同时优化表达网络和估计网络的参数。估计网络的损失函数主要由样本能量$E(\boldsymbol{Z})$构成，而表达网络的优化目标是使所有样本的边际和$p(\boldsymbol{X})$最大[30]：

$$\max p(\boldsymbol{X})=\frac{p(\boldsymbol{X}Z_{\mathrm{l}})}{p(Z_{\mathrm{l}}|\boldsymbol{X})}=\frac{p(\boldsymbol{X}Z_{\mathrm{l}})}{p(Z_{\mathrm{l}}|\boldsymbol{X})}\frac{q(Z_{\mathrm{l}}|\boldsymbol{X})}{q(Z_{\mathrm{l}}|\boldsymbol{X})} \tag{17}$$

式中：$p(\boldsymbol{X}Z_{\mathrm{l}})$为$\boldsymbol{X}$和$Z_{\mathrm{l}}$的联合先验分布；$p(Z_{\mathrm{l}}|\boldsymbol{X})$为潜变量$Z_{\mathrm{l}}$的真实条件分布。

两边取对数再积分得，

$$\ln p(\boldsymbol{X})\int_{Z_{\mathrm{l}}}q(Z_{\mathrm{l}}|\boldsymbol{X})\mathrm{d}Z_{\mathrm{l}}=\int_{Z_{\mathrm{l}}}\ln\left(\frac{p(\boldsymbol{X}Z_{\mathrm{l}})}{q(Z_{\mathrm{l}}|\boldsymbol{X})}\right)q(Z_{\mathrm{l}})\mathrm{d}Z_{\mathrm{l}}+G_{\mathrm{KL}}(q(Z_{\mathrm{l}}|\boldsymbol{X})\|p(Z_{\mathrm{l}}|\boldsymbol{X})) \tag{18}$$

$$G_{\mathrm{ELBO}}=\int_{Z_{\mathrm{l}}}\ln\left(\frac{p(Z_{\mathrm{l}})p(\boldsymbol{X}|Z_{\mathrm{l}})}{q(Z_{\mathrm{l}}|\boldsymbol{X})}\right)q(Z_{\mathrm{l}}|\boldsymbol{X})\mathrm{d}Z_{\mathrm{l}}=E_{q(Z_{\mathrm{l}}|\boldsymbol{X})}(\ln p(\boldsymbol{X}|Z_{\mathrm{l}}))-G_{\mathrm{KL}}(q(Z_{\mathrm{l}}|\boldsymbol{X})\|p(Z_{\mathrm{l}}|\boldsymbol{X})) \tag{19}$$

式中：G_{ELBO}为变分下界（evidence lower bound，ELBO）；$G_{\mathrm{KL}}(q(Z_{\mathrm{l}}|\boldsymbol{X})\|p(Z_{\mathrm{l}}|\boldsymbol{X})$为KL（Kullback-Leibler）散度。

式(18)中，等式等号右边由G_{ELBO}和潜变量Z_{l}的真实条件分布$p(Z_{\mathrm{l}}|\boldsymbol{X})$与近似后验分布$q(Z_{\mathrm{l}}|\boldsymbol{X})$

的KL散度2个部分组成。由于后一项为正,那么网络的优化目标简化成最大化变分下界G_{ELBO}。式(19)中,G_{ELBO}的前一项为重构样本的能量函数,反映了与原始样本的差异,在损失函数中使用两者的距离代替;后一项为潜变量Z_1的真实分布$p(Z_1)$与近似后验分布$q(Z_1|\boldsymbol{X})$的KL散度。假设潜变量$Z_{1,w}$的真实分布$p(Z_{1,w})$服从标准正态分布,则

$$G_{\mathrm{KL}}(q(Z_1|\boldsymbol{X})\|p(Z_1))=$$
$$G_{\mathrm{KL}}(N'(\mu'(\boldsymbol{X}),\sigma'(\boldsymbol{X})^2\|N(0,1))=$$
$$\frac{1}{2}[-\ln(\sigma'(\boldsymbol{X})^2)+\mu'(\boldsymbol{X})^2-(1+\sigma'(\boldsymbol{X})^2)] \tag{20}$$

综上,VAEGMM网络的联合损失函数J的定义为:

$$J=\frac{1}{W}\sum_{w=1}^{W}L(\boldsymbol{X}_w,\boldsymbol{X}'_w)+\frac{1}{W}\sum_{w=1}^{W}G_{\mathrm{KL}}(q(Z_{1,w}|\boldsymbol{X}_w)\|p(Z_{1,w}))+\frac{\lambda_1}{W}\sum_{w=1}^{W}E(Z_w) \tag{21}$$

式中:λ_1为估计网络的样本能量$E(Z)$在联合损失函数J中的权重;$L(\boldsymbol{X}_w,\boldsymbol{X}'_w)$为重构损失函数。

3 串谋预警步骤

集中竞价发电企业串谋预警的具体步骤如下:

步骤1:采用提出的串谋指标体系和指标集计算方法对原始数据进行测算,得到指标集$\boldsymbol{X}$。

步骤2:对$\boldsymbol{X}$进行归一化处理,同时对指标特征集中竞价场次进行独热编码处理,将数据集维度扩增L维,消除量纲。

步骤3:将$\boldsymbol{X}$划分成训练集、验证集和测试集。

步骤4:采用VAEGMM模型进行训练,得到每个样本在低维空间中的密度估计值$\hat{\boldsymbol{\pi}}$,可认为是串谋嫌疑度。

步骤5:将$\hat{\boldsymbol{\pi}}$逆映射回式(8)的矩阵当中,通过横纵坐标获得两两企业的串谋嫌疑度。倘若企业1分别与企业2和企业3有串谋嫌疑,即可认为企业1、2、3为一个串谋联盟。

步骤6:将串谋嫌疑大的企业名单提交给相关监管机构进一步调查,包括检查交易申报的计算机MAC地址与网络IP地址、问询笔录和核查账簿等。

4 实例计算与分析

为了验证VAEGMM用于串谋预警的有效性,采用某省电力市场集中竞价中4场3阶段式报价数据作为原始数据。经过串谋指标集测算,数据集$\boldsymbol{X}$包含3 797个样本,其中负类样本(串谋样本)有363个,占总体的9.56%,具有正负样本不均衡的特点。$\boldsymbol{X}$的特征为1.1节中的7个指标,其中第7个竞价场次指标经过了独热编码处理。

表1展示了指标集$\boldsymbol{X}$的10个样本,其中标签为1表示串谋样本,标签为0表示正常样本。可以看到,前3个串谋样本的第2个指标数值远小于正常样本,表示这些串谋企业的报价一致性很高。显然,这3个样本在样本空间中远离正常样本,被视为离群点,也就是异常样本。

表1 $\boldsymbol{X}$部分数据集
Table 1 Part data set of $\boldsymbol{X}$

样本	指标1	指标2	指标3	指标4	指标5	指标6	指标7	标签
1	0.010 2	0.153 0	0.142 2	0.994 8	0.345 7	1.001 5	[0 0 1 0]	1
2	0.005 6	0.905 5	0.106 9	1.649 0	0.341 5	0.998 4	[0 0 1 0]	1
3	0.010 9	0.948 4	0.314 4	1.729 9	0.343 1	0.999 6	[0 0 1 0]	1
4	0.014 9	6.733 3	3.182 2	7.465 5	0.345 8	1.001 6	[0 0 1 0]	0
5	0.014 3	6.112 1	0.031 2	3.192 1	0.346 3	1.000 0	[0 0 0 1]	0
6	0.050 8	4.461 1	10.966 3	2.323 5	0.349 0	1.001 2	[0 1 0 0]	0
7	0.012 4	2.860 9	1.225 9	6.608 6	0.345 8	1.001 7	[0 0 1 0]	0
8	0.006 2	3.870 1	0.324 2	1.471 3	0.341 6	0.996 5	[0 0 0 1]	0
9	0.043 6	3.584 3	12.428 6	3.418 2	0.343 3	0.997 8	[0 0 0 1]	0
10	0.014 0	6.305 6	3.189 2	4.259 3	0.342 6	0.997 2	[0 0 0 1]	0

4.1 网络结构和参数设置

本次训练中,VAEGMM的网络结构设置如表2所示。表2中,FC表示该神经网络层为全连接层;L2(0.001)表示权重为0.001的L2正则化;Sampling表示采样层,从高斯分布$N(Z_{\mathrm{mean}},\mathrm{e}^{0.5Z_{\mathrm{var}}}\varepsilon)$中采样得到潜变量$Z_1$,其中$\varepsilon$为服从$N(0,1)$的伪随机数,$Z_{\mathrm{mean}}$和$Z_{\mathrm{var}}$分别为$Z$的均值和方差。

另外,为了考虑每个样本降维前后的重构概率,

本算例采用表达网络输入和输出之间的相对欧氏距离d_E和相对余弦d_C相似度，计算公式分别如式(22)和式(23)所示。

$$d_E=\frac{\|\boldsymbol{X}-\boldsymbol{X}'\|_2}{\|\boldsymbol{X}\|_2} \tag{22}$$

$$d_C=\frac{\boldsymbol{X}\boldsymbol{X}'}{\|\boldsymbol{X}\|_2\|\boldsymbol{X}'\|_2} \tag{23}$$

表2　表达网络设置
Table 2　Settings of express network

网络类型	输入	权重	激活函数	正则化	输出
FC	$\boldsymbol{X}$	(10,5)	tanh	L2(0.001)	$\boldsymbol{X}_1$
FC	$\boldsymbol{X}_1$	(5,2)	无	无	$\boldsymbol{Z}_{mean}$
FC	$\boldsymbol{X}_1$	(5,2)	Softplus	无	$\boldsymbol{Z}_{var}$
Sampling	$\boldsymbol{Z}_{mean},\boldsymbol{Z}_{var}$	无	无	无	$\boldsymbol{Z}_1$
FC	$\boldsymbol{Z}_1$	(2,5)	tanh	L2(0.001)	$\boldsymbol{X}_2$
FC	$\boldsymbol{X}_2$	(5,10)	Sigmoid	无	$\boldsymbol{X}'$

表3中，为了防止估计网络发生过拟合现象，在第2层中添加Dropout层，即在每次迭代时，该层神经元节点以0.5的概率关闭[33]。估计网络通过softmax激活函数输出每个样本在低维空间的密度估计$\hat{\boldsymbol{\pi}}$，并由式(22)分离出异常样本。最后，本次训练设置批量为512，优化器为Adam，λ_1=0.01，学习率为0.000 1，迭代次数为5 000。

表3　估计网络设置
Table 3　Settings of estimation network

网络类型	输入	权重	激活函数	正则化	输出
FC	$\boldsymbol{Z}_1,d_E,d_C$	(4,10)	tanh	无	$\boldsymbol{X}_3$
Dropout	$\boldsymbol{X}_3$	0.5	无	无	$\boldsymbol{X}_3$
FC	$\boldsymbol{X}_3$	(10,2)	softmax	无	$\hat{\boldsymbol{\pi}}$

4.2　串谋预警效果分析

根据3.1节中的网络结构和相关参数设置，该模型的训练集和验证集误差如图4所示。在经过2 500次迭代后，训练误差和验证误差已经下降到一个非常低的水平，并且在后续迭代中也保持稳定，没有出现过拟合现象。这表明VAEGMM具有收敛快、精度高的优点。

为了体现VAEGMM的串谋预警效率，与其他无监督智能方法进行了对比，包括基于树模型的孤立森林、单类支持向量机(one class support vector machine，OC-SVM)方法、基于先降维再聚类思想的主成分分析(principal component analysis，PCA)和K均值聚类(PCA+KMeans)方法、基于先聚类再进行密度估计的主成分分析和GMM(PCA+GMM)方法、基于密度的噪声应用空间聚类(density-based spatial clustering of applications with noise，DBSCAN)方法、基于距离的局部异常因子(local outlier factor，LOF)方法、基于深度自编码器(deep autoencoder，DA)和高斯混合模型的深度联合(DAGMM)方法。为了计算准确率，本文设置了阈值$\lambda=0.8\hat{\pi}_{max,w}$来判断串谋，即$\hat{\pi}_w>0.8$为串谋样本，否则为正常。其中$\hat{\pi}_w$和$\hat{\pi}_{max,w}$分别为第$w$个样本的网络输出及其最大值，采取的评价体系为：准确率、召回率与F_1指数，定义如下。

$$A_{cc}=\frac{\alpha_s+\alpha_{us}}{\alpha_s+\alpha_{us}+\gamma_{us}+\gamma_s} \tag{24}$$

$$\eta_{SR}=\frac{\alpha_s}{\alpha_s+\gamma_{us}} \tag{25}$$

$$\eta_{UR}=\frac{\alpha_{us}}{\alpha_{us}+\gamma_s} \tag{26}$$

$$O_{mean}=\frac{2\eta_{SR}\eta_{UR}}{\eta_{SR}+\eta_{UR}} \tag{27}$$

式中：α_s和γ_{us}分别为正常样本被正确或错误预测的数目；α_{us}和γ_s分别为异常样本被正确或错误预测的数目；A_{cc}、η_{SR}、η_{UR}、O_{mean}分别为准确率、召回率、精确度、F_1指数。

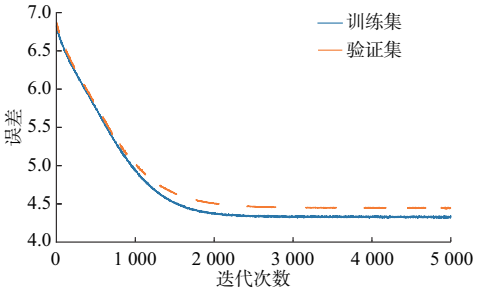

图4　VAEGMM的训练误差和验证误差
Fig. 4　Training and validation error of VAEGMM

表4　不同方法的串谋预警效率
Table 4　Early-warning efficiency of collusion with different methods

方法	准确率/%	召回率/%	F1指数/%
孤立森林	61.84	64.90	75.68
OC-SVM	75.00	76.60	84.86
PCA+KMeans	80.61	89.48	89.26
PCA+GMM(α_{us}=0)	85.08	94.44	91.94
DBSCAN	81.45	82.41	88.93
LOF	83.15	90.94	90.67
DAGMM	82.73	89.00	90.45
VAEGMM	84.00	90.40	91.23

表4的结果表明，VAEGMM在电力市场集中竞价的串谋预警准确率高于孤立森林、OC-SVM和PCA+Kmeans方法，分别高出了22.16%、9%、3.36%。对于DBSCAN方法，VAEGMM的召回率更高，对正常样本的识别更加敏感。虽然LOF在本算例的表现很接近VAEGMM，但是LOF对于复杂的高维数据处理比较困难。虽然PCA+GMM方法的3个指标都优于VAEGMM，但它并没有识别出异常样本(α_{us}=0)。因为在正负样本不均衡的情况下，简单的线性降维模型PCA会把异常样本的信息当作噪声删除，导致GMM对异常样本不敏感，无法分离出串谋样本，反而只留下正类样本的信息，能够帮助GMM识别出更多的正类样本，具有高准确率。因此，PCA+GMM方法在实际应用中没有实用价值。对比同样是深度联合网络的DAGMM，VAEGMM在3个指标的表现上都更优异，分别高出1.27%、1.4%、0.78%，表明VAEGMM的表达网络比DAGMM更加能够学习得到有助于密度估计的低维表示。

综上，深度联合学习网络VAEGMM能够高效地将串谋样本识别出来，与其他无监督学习方法相比，更加契合电力市场的串谋数据特点，具有更高的准确率。

4.3　与其他方法比较分析

为了说明本文方法的优越性，将从不同维度对各类发电企业的串谋预警方法进行评价。

如表5所示，与现有的其他方法[10-14,17-18]相比，本文方法不需要很强的先验知识，能够根据市场的交易数据不断自我更新，从而可以对串谋行为实时监测。在信息披露机制不完善的情况下，本文方法只需要发电企业交易后的竞价数据，不需要保密的机组数据，也不需要标注数据，更加适合于第三方监管机构使用。

表5　不同方法比较结果
Table 5　Comparison results of different methods

方法	所需数据	数据标签	实时性能	自我更新能力
专家决策[13-14]	竞价数据	无	坏	无
模式识别[17]	竞价数据	无	好	无
有监督学习[18]	竞价数据	有	好	有
本文方法	竞价数据	无	好	有
博弈论方法[10-12]	边际成本、竞价数据	无	好	有

如表6所示，经同一实例数据计算，基于有监督学习方法[13]的串谋预警准确率达到91.04%，比本文方法高出了7.04%。但是在实际操作中，人工标注的串谋样本非常少，训练一个有监督学习模型是不现实的。另外，若测试集中发电企业通过新的方式进行串谋，有监督学习模型则无法将该类样本识别出来。因此，本文基于无监督学习的串谋智能预警方法更加符合实际。

表6　不同方法预警效果比较
Table 6　Comparison of early-warning effects with different methods

方法	准确率/%	召回率/%	F_1指数/%
有监督学习[18]	91.04	96.84	95.22
本文方法	84.00	90.40	91.23

5　结语

为了完善中国电力市场集中竞价的监管体系，结合串谋指标体系和智能算法，提出了一套电力市场发电企业之间的串谋智能预警方法，用于辅助专家决策。得到的主要结论如下：

1）构建了较为完善的发电企业串谋指标体系，提出了集中竞价场次标注指标，提升了方法的预警性能。

2）由于电力市场串谋行为样本数据的标签难以获取，因此选择了无监督学习模型。考虑到指标集具有高维度和正负样本不均衡特点，提出了深度联合学习网络VAEGMM，实现了电力市场发电企业串谋行为的智能预警。

3）与其他方法相比较，VAEGMM每次迭代都进行1次降维，最大化保留了原始数据信息，对数据空间的调整也更加灵活，在实际应用中具有更高的准确率。

理论上，本文所提出的VAEGMM是基于电力市场中的数据特点开发出来的。它既可以预警发电侧的串谋行为，也可以适用于其他电力市场主体，这取决于串谋指标体系如何构建。因此，未来的研究可以着眼于VAEGMM对其他电力市场主体的串谋预警研究。

在本文审稿过程中，审稿专家与作者的讨论见附录A。

附录见本刊网络版(http://www.aeps-info.com/aeps/ch/index.aspx)，扫英文摘要后二维码可以阅读网络全文。

参考文献

[1] 董礼，王胜华，华回春，等. 中国现货电力市场中发电企业滥用市场力违规识别[J/OL]. 中国电机工程学报：1-11[2021-11-11]. https://doi.org/10.13334/j.0258-8013.pcsee.201625.
DONG Li, WANG Shenghua, HUA Huichun, et al.

Identification of market power abuse in spot market of Chinese electric market [J]. Proceedings of the CSEE: 1-11 [2021-11-11]. https://doi.org/10.13334/j.0258-8013.pcsee.201625.

[2] 王文举，范合君.企业价格串谋识别的博弈分析及模拟[J].商业研究，2010(5)：49-52.
WANG Wenju, FAN Hejun. Analysis and simulation of identification of price collusion [J]. Commercial Research, 2010(5): 49-52.

[3] EU Commission, DG Competition. Report on energy sector inquiry, SEC (2006) 1724, 10.1.2007: Brussels [R/OL]. [2021-11-11]. http://ec. europa. eu/public_opinion/archives/eb_special_en.htm.

[4] KANAGALA A, SAHNI M, SHARMA S, et al. A probabilistic approach of Hirschman-Herfindahl Index (HHI) to determine possibility of market power acquisition [C]// Power Systems Conference and Exposition, October 10-13, 2004, New York, USA: 1277-1282.

[5] BATAILLE M, BODNAR O, STEINMETZ A, et al. Screening instruments for monitoring market power—the return on withholding capacity index (RWC) [J]. Energy Economics, 2019, 81: 227-237.

[6] NEUHOFF K, BARQUIN J, BOOTS M G, et al. Network-constrained Cournot models of liberalized electricity markets: the devil is in the details [J]. Energy Economics, 2005, 27(3): 495-525.

[7] HOBBS B F, HELMAN U. Complementarity-based equilibrium modeling for electric power markets [M]. New York, USA: Wiley, 2004.

[8] IÑÓN J, HOBBS B F. Generation adequacy, market regulation and demand elasticity in the electricity industry: a stochastic long run equilibrium analysis of capacity markets [EB/OL]. [2021-11-11]. http://iaee. org/documents/washington/Ben_Hobbs2.pdf.

[9] XIE W, DING Q, TU M F, et al. Market power monitoring framework and measures in electricity market [C]// 2019 IEEE Sustainable Power and Energy Conference (iSPEC), November 21-23, 2019, Beijing, China: 129-133.

[10] 孙波，邓瑞林，谢敬东，等.基于排序多元Logit模型的卡特尔类机组串谋竞价识别[J].电力系统自动化，2021，45(6)：109-115.
SUN Bo, DENG Ruilin, XIE Jingdong, et al. Collusion bidding identification of Cartel-type generators based on ordered logit model [J]. Automation of Electric Power Systems, 2021, 45(6): 109-115.

[11] 蒋玮，吴杰，冯伟，等.日前电力市场不完全信息条件下的电力供需双边博弈模型[J].电力系统自动化，2019，43(2)：18-24.
JIANG Wei, WU Jie, FENG Wei, et al. Bilateral game model of power supply and demand sides with incomplete information in day-ahead electricity market [J]. Automation of Electric Power Systems, 2019, 43(2): 18-24.

[12] ESMAEILI ALIABADI D, KAYA M, ŞAHIN G. Determining collusion opportunities in deregulated electricity markets [J]. Electric Power Systems Research, 2016, 141: 432-441.

[13] 史述红，刘敦楠，胡会文，等.市场开放下电力交易全过程违规行为识别探究[J].价格理论与实践，2018(8)：51-54.
SHI Shuhong, LIU Dunnan, HU Huiwen, et al. Identification of violations in the whole process of electricity market transactions [J]. Price: Theory & Practice, 2018(8): 51-54.

[14] 刘敦楠，李瑞庆，陈雪青，等.电力市场监管指标及市场评价体系[J].电力系统自动化，2004，28(9)：16-21.
LIU Dunnan, LI Ruiqing, CHEN Xueqing, et al. Surveillance indices and evaluating system of electricity market [J]. Automation of Electric Power Systems, 2004, 28(9): 16-21.

[15] TELLIDOU A C, BAKIRTZIS A G. Agent-based analysis of capacity withholding and tacit collusion in electricity markets [J]. IEEE Transactions on Power Systems, 2007, 22(4): 1735-1742.

[16] LIANG Y C, GUO C L, DING Z H, et al. Agent-based modeling in electricity market using deep deterministic policy gradient algorithm [J]. IEEE Transactions on Power Systems, 2020, 35(6): 4180-4192.

[17] 刘敦楠，张潜，李霄彤，等.基于云模型与模糊Petri网的电力市场潜在危害行为识别[J].电力系统自动化，2019，43(2)：25-33.
LIU Dunnan, ZHANG Qian, LI Xiaotong, et al. Identification of potential harmful behaviors in electricity market based on cloud model and fuzzy Petri net [J]. Automation of Electric Power Systems, 2019, 43(2): 25-33.

[18] 张海生，曹喆，杨昌海，等.基于AdaBoost-DT算法的电力市场串谋行为识别研究[J].电力工程技术，2020，39(2)：152-158.
ZHANG Haisheng, CAO Zhe, YANG Changhai, et al. Collusive behavior recognition in electricity market based on AdaBoost-DT algorithm [J]. Electric Power Engineering Technology, 2020, 39(2): 152-158.

[19] KIM K I, FRANZ M O, SCHOLKOPF B. Iterative kernel principal component analysis for image modeling [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(9): 1351-1366.

[20] HUBER P J. International encyclopedia of statistical science [M]. Berlin, Heidelberg, Germany: Springer, 2011: 1248-1251.

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

[22] ZHOU C, PAFFENROTH R C. Anomaly detection with robust deep autoencoders [C]// 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2017, Halifax, Canada: 665-674.

[23] 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.

[24] KIM J, SCOTT C. Robust kernel density estimation [J]. The Journal of Machine Learning Research, 2012, 13(1): 2529-2565.

[25] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection [J]. ACM Computing Surveys, 2009, 41(3): 1-58.

[26] CHEN Y Q, ZHOU X S, HUANG T S. One-class SVM for learning in image retrieval [C]// 2001 International Conference on Image Processing, October 7-10, 2001, Thessaloniki, Greece: 34-37.

[27] SONG Q, HU W J, XIE W F. Robust support vector machine with bullet hole image classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2002, 32(4): 440-448.
[28] ZONG B, SONG Q, MIN M R, et al. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection [EB/OL]. [2021-11-11]. https://sites.cs.ucsb.edu/~bzong/doc/iclr18-dagmm.pdf.
[29] FAN H Y, ZHANG F B, WANG R D, et al. Correlation-aware deep generative model for unsupervised anomaly detection [C]// 24th Pacific-Asia Conference, PAKDD 2020, May 11-14, 2020, Singapore, Singapore: 688-700.
[30] DOERSCH C. Tutorial on variational autoencoders[EB/OL]. [2021-11-11]. https://arxiv.org/pdf/1606.05908.pdf.
[31] AN J, CHO S. Variational autoencoder based anomaly detection using reconstruction probability[EB/OL]. [2021-10-10]. http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf.
[32] 梁杰,陈嘉豪,张雪芹,等.基于独热编码和卷积神经网络的异常检测[J].清华大学学报(自然科学版),2019,59(7):523-529.
LIANG Jie, CHEN Jiahao, ZHANG Xueqin, et al. One-hot encoding and convolutional neural network based anomaly detection [J]. Journal of Tsinghua University (Science and Technology), 2019, 59(7): 523-529.
[33] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

华回春(1980—),男,博士,副教授,主要研究方向:电能质量、电力市场。E-mail:huahuichun@126.com
邓　彬(1996—),男,通信作者,硕士研究生,主要研究方向:电力市场、深度学习。E-mail:dengbin13992@126.com
刘　哲(1992—),男,硕士,高级工程师,主要研究方向:能源互联网、电能质量。E-mail:liuzheacyy@163.com

(编辑　顾晓荣)

Intelligent Early-warning of Collusion Between Power Generation Enterprises Based on Variational Autoencoding Gaussian Mixture Model

HUA Huichun[1], *DENG Bin*[1], *LIU Zhe*[2], *ZHANG Lifeng*[1]
(1. State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources (North China Electric Power University), Baoding 071003, China;
2. State Grid Shanghai Municipal Electric Power Company, Shanghai 200437, China)

Abstract: As the scale of market transactions becomes larger and the amount of transaction data increases, it becomes possible to conduct collusion analysis with data. Therefore, combining with the collusion early-warning indicator system of the power generation enterprises and the unsupervised variational autoencoding Gaussian mixture model (VAEGMM), the intelligent early-warning of the collusion between power generation enterprises is realized. Firstly, a complete indicator system for the collusion early-warning and a detailed indicator calculation method are proposed. Secondly, in view of the high-dimensional data characteristics of the index set and the imbalance of positive and negative samples, the VAEGMM is proposed based on the idea of anomaly detection. Then, the network structure of VAEGMM is described in detail, and the joint loss function is reconstructed, making the network better learn the low dimensional expression of the original data. Thus it is helpful for more accurate density estimation. Finally, the actual case study shows that compared with other traditional unsupervised learning models, VAEGMM can warn the risk of collusion more efficiently and accurately.

This work is supported by State Grid Corporation of China (No. SGSHDK00HZJS2000254).

Key words: electricity market; power generation enterprise; intelligent early-warning; collusion; variational autoencoding Gaussian mixture model (VAEGMM)