基于增量SOM算法的虚拟机异常状态检测技术研究

周真
西南民族大学计算机科学与工程学院

摘　要：本文以云数据中心虚拟机运行状态为异常检测目标，针对由云环境下虚拟机部署运行高动态性所导致的无法批量获取运行状态训练数据，训练数据无标签等问题，提出基于增量SOM无监督学习方法的虚拟机运行状态建模及异常运行状态检测设别技术。文中对基于增量SOM算法的虚拟机运行状态建模和异常运行状检测设别过程进行了详细描述，对SOM网络关键训练参数的设定问题进行了讨论和研究，并给出优化设定方案。相关仿真实验和实际应用效果验证了本文提出的针对虚拟机运行状态的异常检测技术在云环境下的有效性。

关键词：云计算；虚拟机；运行状态；增量SOM；异常检测
【DOI】10.12252/j.issn.2096-6288.2023.04.217

一、研究背景

云计算已经成为当前及未来IT资源管理和使用的主要模式。为了进一步提升云数据中心资源利用率和管理效率，云服务提供商在云数据中心的构建中广泛采用虚拟化技术[1]。在虚拟化云环境下，用户应用系统主要通过虚拟机来承载和运行，虚拟机运行状态出现异常会直接影响应用使用体验。然而云环境的高动态、资源共享、大规模等特性又导致部署其中虚拟机容易出现各类运行异常，为了便于对运行异常虚拟机的及时接管和故障排除，就需要专门的异常检测和识别技术来准确发现运行状态存在异常的虚拟机[2-3]。

虽然目前针对各种类型的异常检测算法的相关研究工作较多，但大多数研究工作都聚焦在假设理想环境下对各类异常检测算法检测性能及计算复杂度的改进方面。这类研究虽然可通过仿真实验证明其有效性，但往往忽略具体应用场景特征对异常检测造成的影响，导致研究成果实际应用效果较差或者根本无法满足具体异常检测应用场景的需求。

目前面向云环境下虚拟机异常运行状态这一具体应用场景的专用异常检测技术还相对缺乏。针对这一情况，本文拟深入研究云环境下虚拟机部署运行特点，及其对虚拟机异常运行状态检测识别造成的影响，并在此基础上结合各类异常检测算法特点提出基于增量SOM无监督学习方法的虚拟机运行状态建模及异常运行状态检测设别技术。该技术是在充分考虑考虑具体应用场景特征基础上提出的，能够较好的适应和满足云环境下针对虚拟机异常运行状态进行检测识别的应用需求。

二、基于增量SOM（Incremental Self-Organiziong Maps）学习方法的虚拟机运行状态建模

由于虚拟机的封装特性，外界很难获取虚拟机内部的相关信息。因此只能通过虚拟机监控器（VMM）或虚拟机所驻留主机操作系统所采集的虚拟机性能指标集（System-Level Metrics）来全面反映虚拟机的运行状态。虚拟机性能指标集包含数大量各中类型的性能指标（数量通常在数千个），这些性能指标主要从不同角度和粒度反映虚拟机CPU、内存、I/O、网络等关键资源的使用和运行情况。

本文定义向量$M(m_1,m_2,m_3,\ldots,m_n)$来表示一个虚拟机的运行状态，其中$m_i$表示虚拟机性能指标集中的某个特定指标。$M\in R^n$是虚拟机性能指标集合样本空间的随机变量，在某时刻t对虚拟机性能指标集采样得到的样本$M(t)$（其中$t=1,2,3,\ldots$表示样本采集的时间序列号），都是针对随机变量M的一次随机实验结果。

要实现对虚拟机异常运行状态的检测和识别，首先需要采集大量虚拟机运行状态$M(t),t=1,2,3,\ldots$作为训练样本集，训练能够拟合虚拟机运行状态的模型$Y(w_1,w_2,w_3,\ldots,w_m)$（其中$w_i$表示模型参数），然后通过模型$Y(w_1,w_2,w_3,\ldots,w_m)$对虚拟机的待检测运行状态$M(t_d)$进行推理计算，并输出$M(t_d)$属于异常/正常的推理结果。

云数据中心虚拟机部署调度的高动态性特点会对虚拟机运行状态建模造成两方面的不利影响：（1）无法即时采集获取目标虚拟机的批量运行状态数据作为训练样本集合；（2）采集获取虚拟机运行状态训练样本通常是无标签的。针对上述情况，本文采用轻量级的无监督增量SOM学习方法来实现针对虚拟机的运行状态建模和推理检测，以适应在高动态云环境下针对虚拟机异常运行状态的实时检测。

SOM学习方法通常构建一个$N\times N$晶格状神经元阵列（lattice type of the array of neurons），也被称为SOM网络，其中每个神经元n_{ij}，$i,j=1,2,3,\ldots,N$都关联一个对应的权重向量$W_{ij}(w_1,w_2,w_3,\ldots,w_m)$，且权重向量维数与训练样本数据维数相同。SOM学习方法按照一定策略基于训练样本对阵列中神经元所关联的权重向量进行多轮迭代修改调整，直到整个神经元阵列的权重向量收敛（即SOM网络收敛）。具体到本文应用场景，基于SOM学习方法训练获得的SOM网络中各神经元权重会拟合和映射目标虚拟机的常见（即正常）运行状态。如果目标虚

拟机的某待检测运行状态$M(t_d)$与SOM网络中所有神经元权重$W_{ij}(w_1,w_2,w_3,...,w_m)$都存在较大距离（即不相似），则认为目标虚拟机的运行状态$M(t_d)$不在正常范围内，并判断其运行状态存在异常。

三、SOM网络的迭代训练

针对目标虚拟机，其SOM网络的迭代拟合训练过程如下图1所示：

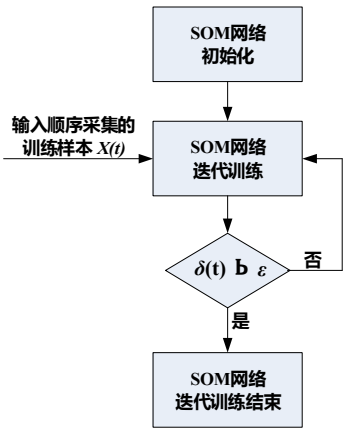

图1 SOM网络迭代训练过程

在开始迭代训练前，需要为SOM网络中神经元的权重向量设置初始化取值$W_{ij}(0),\ i,j=1,2,3,...,N$。本文采用随机方设定方式实现权重向量的初始化。

在迭代训练过程中，针对t时刻采集训练样本$M(t)$，在当前状态下的SOM网络中寻找一个神经元作为本轮迭代训练邻域的中心点n_C，其选择依据如公式（1）所示：

$$n_C=\begin{cases}\arg\min\limits_{(i,j)}\left\{\left\|M(t)-W_{ij}(0)\right\|\right\}, & t=1\\ \arg\min\limits_{(i,j)}\left\{\left\|M(t)-W_{ij}(t-1)\right\|\right\}, & t=2,3,...\end{cases} \quad (1)$$

根据式（1）可知，在基于样本$M(t)$进行训练时，会选择SOM网络中权重值与$M(t)$具有最小欧式距离的神经元作为练邻域的中心点n_C。基于n_C可以进一步定义训练邻域，训练领域限制了基于样本$M(t)$对SOM网络进行训练时受影响的神经元范围，其具体定义如公式（2）所示：

$$H_{n_C}^{(i,j)}=\alpha(t)\cdot\exp\left(-\frac{\left\|n_C(i,j)-(i,j)\right\|^2}{2\sigma^2(t)}\right) \quad (2)$$

为了保证SOM网络迭代训练的收敛性，训练邻$H_{n_C}^{(i,j)}$域函数（即smoothing kernel）应该是$\left\|n_C(i,j)-(i,j)\right\|$和迭代训练次数t的单调递减函数，其中$n_C(i,j)$是$n_C$在SOM网络中的坐标，$(i,j)$是SOM网络中神经元$n_{ij}$的坐标。根据公式（2）可知，为了更好的平滑效果本文采用高斯函数（Gaussian function）作为平滑核，其中$\sigma(t)$用于调整训练邻域半径，其大小决定训练样本$M(t)$对SOM网络拟合训练时的影响范围，$\alpha(t)$为学习速率因子，其大小决定了训练时SOM网络对训练样本$M(t)$的拟合程度。$\alpha(t)$和$\sigma(t)$都应该是迭代训练次数t的单调递减函数，以保证SOM网络训练的收敛性。

确定训练邻域后，可基于公式（3）根据训练样本$M(t)$对SOM网络中处在训练邻域内的神经元权重值进行训练调整。

$$W_{ij}(t)=W_{ij}(t-1)+H_{n_C}^{(i,j)}\cdot\left[M(t)-W_{ij}(t-1)\right] \quad (3)$$

根据按时间序列采集的训练样本$M(t),t=1,2,3,...$，上述训练过程可反复迭代进行直到SOM网络收敛。SOM网络的收敛条件如下面的公式（4）所示：

$$\delta(t)=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|W_{ij}(t)-W_{ij}(t-1)\right\|\le\varepsilon \quad (4)$$

其中ε表示一个足够小的实数，$W_{ij}(t-1),\ W_{ij}(t),1\le i,j\le N$，分别表示经过第$t-1$和$t$个样本训练后SOM网络中各神经元的权重值。公式（4）含义是当相邻两轮迭代训练后的SOM网络中各神经元权重值之间的平均偏差小于给定的阈值ε，就认为各神经元权重值趋于稳定（即SOM网络已收敛），可以结束迭代训练过程，反之需要更多的训练样本继续迭代训练。

四、SOM网络迭代训练参数设定

SOM网络的迭代拟合训练过程可以分为：初步有序化阶段（roughly ordering phase）和收敛阶段（convergence phase）。一般经过1000个左右的训练样本训练后，SOM网络就可达到初步有序化，在本文应用场景下初步有序化的SOM网络可基本拟合虚拟机的常见运行状态，并用于虚拟机异常运行状态的检测识别。而SOM网络的收敛阶段是在初步有序化的SOM网络基础上基于更多训练样本对SOM网络进行局部调整和精化，最终得到拟合程度更高的模型，但该过程需要较多训练样本和迭代训练次数，耗费较长时间。

为了兼顾SOM网络训练拟合速度和避免SOM网络陷入亚稳定状态，在上述两个训练阶段中对训$H_{n_C}^{(i,j)}$练邻域函数中的参数$\alpha(t)$和$\sigma(t)$的设定有不同的要求。具体设定方法如下：

（一）训练参数$\alpha(t)$设定

$$\alpha(t)=\begin{cases}\exp\left(-\frac{t-1}{t}\right), & t=1,2,3,...,1000\\ 0.2\cdot\exp\left(-\frac{t-1}{t}\right), & t>1000\end{cases} \quad (5)$$

根据公式（5）可知，当迭代训练次数$t<1000$时（即初步有序化阶段）$\alpha(t)$取值被设定在[0.2,1]区间，取较大的值。其目的是在训练初始阶段让SOM网络充分拟合每个训练样本，提升训练速度。而当迭代训练次数$t>1000$时（即收敛阶段）$\alpha(t)$取值被设定在[0,0.2]区间，取较小的值。其目的是在SOM网络中权重值已经趋于基本稳定情况下限制单个训练样本对让SOM网络产生的影响，避免SOM网络中神经元权重值出现大幅度震荡

抖动现象。

（二）训练参数$\sigma(t)$设定

根据公式（2）可知，在$\alpha(t)$确定的情况下的$\sigma(t)$取值会直接影响训练邻域的大小。和前述$\alpha(t)$设定情况类似，在训练初始阶段为让了SOM网络充分拟合每个训练样本，提升训练速度，$\sigma(t)$取值要让训练邻域有较大的覆盖范围。典型的针对规模$N\times N$的SOM网络其应该覆盖SOM网络直径一半的范围（即N/2）。而在精化训练阶段，为了避免SOM网络中神经元权重值出现大幅度震荡抖动现象，$\sigma(t)$取值要让训练邻域的覆盖范围尽量小，如图2所示，典型覆盖范围为训练邻域中心点n_C的“最近邻域”。

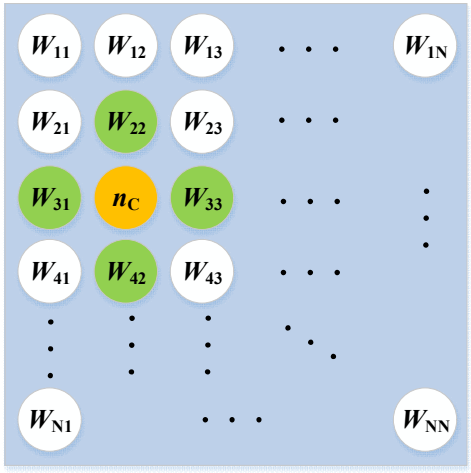

图2 训练邻域中心点n_C的“最近邻域”

根据公式（3）可知，针对训练样本$M(t)$进行SOM网络训练时，其中特定神经元n_{ij}受训练影响程度可由该$H_{n_C}^{(i,j)}$神经元的训练邻域函数的值来调控。当神经元n_{ij}的训练邻域函数值很小或趋近于零时，训练样本$M(t)$对该神经元的影响也趋近于零（即无影响），反之亦然。

而根据训练邻域函数公式（2）可知，当训练邻域中心点n_C和训练迭代次数t确定的情况下，$\|n_C(i,j)-(i,j)\|$和$\alpha(t)$的取值是确定的，神经元n_{ij}的训练邻域函数值由$\sigma(t)$的取值决定。根据上述原理可知，针对特定训练样本$M(t)$的第t轮迭代训练以及针对$M(t)$设定的训练影响范围大小，可以精确计算所需的$\sigma(t)$取值。

下面以规模$N\times N$的SOM网络训练初始阶段为例说明针对$\sigma(t)$的计算方法。训练初始阶段要求针对特定训练样本$M(t)$的第t轮迭代训练影响范围为：以训练邻域中心点n_C为圆点，半径为N/2的圆形范围。具体计算过程如下面公式（6）所示，其中ε为一个趋近于零的实数。

$$\alpha(t)\cdot\exp\left(-\frac{\left(\frac{N}{2}\right)^2}{2\sigma^2(t)}\right)=\varepsilon \Rightarrow \ln\alpha(t)-\frac{\left(\frac{N}{2}\right)^2}{2\sigma^2(t)}=\ln\varepsilon$$

$$\Rightarrow \frac{\left(\frac{N}{2}\right)^2}{2\sigma^2(t)}=\ln\frac{\varepsilon}{\alpha(t)} \Rightarrow \sigma^2(t)=\frac{N^2}{8\cdot\ln\frac{\varepsilon}{\alpha(t)}}$$

$$\Rightarrow \sigma(t)=-\frac{N}{2\sqrt{\ln\left(\frac{\varepsilon}{\alpha(t)}\right)^2}},\ \sigma(t)=\frac{N}{2\sqrt{\ln\left(\frac{\varepsilon}{\alpha(t)}\right)^2}} \quad \because\sigma(t)>0,\ \therefore\Rightarrow \sigma(t)=\frac{N}{2\sqrt{\ln\left(\frac{\varepsilon}{\alpha(t)}\right)^2}} \tag{6}$$

五、基于SOM网络的虚拟机异常运行状态检测识别

通过训练样本集合多次迭代训练获得收敛的SOM网络后，可基于该网络对目标虚拟机的待检测运行状态$M(t_d)$进行推理计算，并根据该计算结果判断$M(t_d)$是否为异常状态。本小节仍然以规模$N\times N$的SOM网络为例展示上述推理计算过程，具体如下面公式（7）所示：

$$Anomaly(M(t_d))=\begin{cases} true,\ \min\left\{\|M(t_d)-W_{ij}\| \mid i,j=1,2,3,\ldots,N\right\}\geq\delta \\ false,\ \min\left\{\|M(t_d)-W_{ij}\| \mid i,j=1,2,3,\ldots,N\right\}<\delta \end{cases} \tag{7}$$

公式（7）中，$W_{ij},1\leq i,j\leq N$为收敛后SOM网络中各神经元的权重值，δ表示一个足够小的常数。该公式含义为：收敛后SOM网络中各神经元的权重值代表了目标虚拟机常见的运行状态（即正常状态），如果该虚拟机的待检测运行状态$M(t_d)$与SOM网络各神经元权重值之间的最小差异小于一个足够小的阈值，表明$M(t_d)$与SOM网络中某个神经元的权重值是相似的，即$M(t_d)$被判定为虚拟机的某个正常运行状态。反之，如果该虚拟机的待检测运行状态$M(t_d)$与SOM网络各神经元权重值之间的最小差异大于阈值，表明$M(t_d)$与SOM网络中各神经元的权重值存在明显差异，即$M(t_d)$被判定为虚拟机的异常运行状态。

结语

本文以云数据中心虚拟机的运行状态作为检测目标，综合考虑云环境下虚拟机的部署运行特点及其异常运行状态检测需求，针对性提出了基于增量SOM的无监督学习方法的虚拟机运行状态建模方法，以及基于运行状态模型的虚拟机异常运行状态检测设别方法。本文提出的针对云数据中心虚拟机运行状态的异常检测方法的有效性已经通过相关仿真实验和实际应用得到验证，具有较好的实际应用前景。

参考文献

[1]屈慧姣.云计算环境下虚拟机资源优化配置与调度研究[D].太原科技大学，2021.

[2]杨光.云架构下的虚拟机异常行为检测方法研究及系统实现[D].四川师范大学，2022.

[3]胡悦.云中虚拟机异常检测与动态迁移研究[D].中国矿业大学，2022.

基金项目：This work is supported by Sichuan Science and Technology Program（Grant No.2020JDRC0040）

作者简介：周真，西南民族大学计算机科学与工程学院讲师，工学博士。