面向云计算的主成分分析多变量局域预测模型*

李洪安，康宝生，张　婧，佟建锋
（西北大学 信息科学与技术学院，西安 710127）

摘　要：针对云平台无法从单变量负荷序列中获取完整预测信息的问题，提出了一种基于主成分分析的多变量局域预测模型并应用于云计算底层资源的预测中。利用主成分分析法综合考虑多种底层资源间的影响关系，确定多变量相空间的嵌入维数，并与局域预测法相结合，由此建立多变量局域预测模型。仿真实验表明，基于主成分分析的多变量局域预测模型的预测精度高于单变量局域预测模型，是面向云计算底层资源预测的一种有效方法。

关键词：云计算；多变量相空间；主成分分析；局域预测模型
中图分类号：TP393.06　　**文献标志码**：A　　**文章编号**：1001-3695(2012)11-4170-06
doi:10.3969/j.issn.1001-3695.2012.11.044

Multivariate-local forecasting model of principal component analysis for cloud computing

LI Hong-an, KANG Bao-sheng, ZHANG Jing, TONG Jian-feng
(*College of Information Science & Technology, Northwest University, Xi'an* 710127, *China*)

Abstract: To solve the problems that cloud computing system can not get enough forecasting information from the univariate load sequence, this paper proposed the multivariate-local forecasting model based on principal component analysis, and applied it in the forecasting of the underlying resource for cloud computing. Using of principal component analysis method to consider the relationship between the underlying resources, it determined the embedding dimension of multivariate phase space and combine with the local forecasting method. The simulation results show that the multivariate-local forecasting model based on principal component analysis can provide more precise of forecasting than the univariate-local forecasting model. Thus, the multivariate-local forecasting model is demonstrated to be efficient for the forecasting of the underlying resource for cloud computing.

Key words: cloud computing; multivariate phase space; principal component analysis; local forecasting model

0　引言

随着云计算的发展，云平台的租户数量不断增长，导致底层资源负荷量也有了较大的变化，如何合理分配及调度底层资源的问题成为业界的挑战[1]。为了科学合理地配置云平台的硬件和软件资源，开展了云平台底层资源的负荷预测研究。一方面，对有限的底层资源进行科学预测、科学规划，提高云平台的资源利用率；另一方面，在云平台规模扩大或云平台升级过程中，可以对云平台各个租户的性能指标进行优化配置，使云平台的投入与产出达到有效化、合理化。

云计算环境中，采集到的底层资源负荷序列类型包括虚拟机 CPU 利用率的负荷序列、内存利用率的负荷序列、硬盘利用率的负荷序列及网络负载利用率的负荷序列等。在预测过程中，一方面，底层资源负荷序列均具有非线性、非平稳的特性；另一方面，负荷序列的特征提取和选择是一项重要且难度较大的工作，由于特征的选择具有很大的主观性，会造成信息的不足或是冗余。针对这些问题，本文利用相空间重构理论对负荷序列进行特征提取。Takens[2] 的嵌入定理为单变量负荷序列重构动力系统奠定了可靠的基础。他认为系统任一个分量的演化是由与之相互作用着的其他分量所决定的。这样，就可以从一批仅仅与时间相关的混沌数据中提取和恢复出系统原来的规律，然后可以在这个空间中用规律来预测轨迹的走向。

云平台涉及多种资源负荷序列的类型，无法从单变量负荷序列中获取完整的系统信息[3]，且不能保证云平台中任意的单变量负荷序列足以重构原系统。因此，云平台的预测模型应利用多变量负荷序列重构出更为准确的原系统，充分考虑各种底层资源间配置的关系问题。目前，针对于多变量负荷序列研究已取得了许多成果。卢山等人[4] 提出由单变量推广的虚假邻近点法计算多变量时间序列的相空间嵌入维数，但其缺少每次扩维选择标准，使重构相空间具有不定性；Cao 等人[5] 对多变量的相空间重构问题进行理论和应用研究，并论证了其优越性；樊重俊[6] 讨论了多观测变量状态空间重构技术，为研究多变量负荷序列的相空间重构提供了理论基础。但是，在相空间重构过程中，这些研究成果的多变量负荷序列均增加了不必要的冗余变量，导致预测模型的结构过于复杂，影响模型学习算法的效率和精度。因此，需要对多变量负荷序列之间的相互影响机制进行研究，选择合理的输入变量，以简化模型结构。如

收稿日期：2012-04-16；**修回日期**：2012-05-24　　**基金项目**：西北大学研究生创新教育项目(10YSY02,08YKC17)

作者简介：李洪安(1978-)，男，博士研究生，主要研究方向为计算机图形图像与信息处理等(an6860@126.com)；康宝生(1961-)，男，教授，博导，博士(后)，主要研究方向为计算机辅助几何设计、计算机图形学、图形图像与多媒体技术等；张婧(1988-)，女，硕士研究生，主要研究方向为服务计算、数据库等；佟建锋(1987-)，男，硕士研究生，主要研究方向为计算机图形图像与信息处理等.

何减少这些冗余信息是确定多变量时间序列嵌入维数的关键。主成分分析方法[7]为研究提供了一种解决途径,既消除了各重构变量间的冗余信息,达到降低多变量相空间维数的目的,又提高了预测模型的预测精度及效率。

在相空间重构基础上,进行负荷序列预测有多种方法。根据拟合相空间中吸引子的方式可分为全局预测法和局域预测法[8]。全局预测法是将轨迹中的全部点作为拟合对象,找出其规律,进而预测轨迹的走向。这种方法理论上是可行的,但当相空间轨迹比较复杂时,预测的结果误差偏大。局域预测法是将相空间轨迹的最后一点作为中心点,把离中心点最近的若干轨迹点作为相关点;然后对这些相关点作出拟合,再估计轨迹下一点的走向;最后从预测出的轨迹点的坐标中分离出所需的预测值。其与全局预测法相比具有柔韧性好、拟合速度快且精度较高的特点,目前得到了广泛的应用和研究[9]。

本文基于混沌理论,首先,对采集到的云平台底层资源负荷序列进行多变量相空间重构,并利用 Lyapunov 指数[10]对云平台底层资源负荷序列进行混沌特征的分析和判定。其次,利用主成分分析法对重构后的多变量相空间进行主成分提取,从而实现降低多变量相空间维数的目的,使得最终的相空间具有较少的冗余度。最后,构建多变量局域[11]预测模型,将多变量相空间轨迹的最后一点作为中心点,通过广义自由度[12]方法确定中心点的邻域子空间,采用最小二乘法对邻域子空间中的点进行线性拟合,构建出邻域子空间中最邻近点的轨道,根据轨道的趋势确定最终的预测结果。图 1 表示本文多变量局域预测模型的基本原理。其中:序列 $x_i(n)$ 表示云平台涉及的多种资源负荷序列,$i=1,2,\cdots,K$ 为云平台的资源负荷种类,$n=1,2,\cdots,N$ 为云平台的各种资源负荷序列的序号;$x_i(n)$,$x_i(n-\tau)$,$\cdots$,$x_i(n-(m_i-1)\tau)$ 表示经过多变量相空间重构后的相空间向量;$x'_1,x'_2,\cdots,x'_K$ 表示经过主成分分析法简化后的相空间向量;$Z(n+h)$ 表示利用局域预测法得到 $(n+h)$ 时刻的预测结果,h 为预测步长。

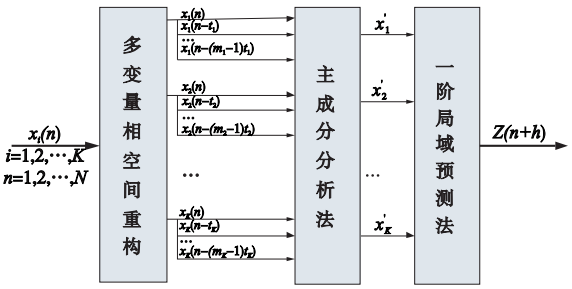


图1　多变量局域预测模型的基本原理

1　多变量负荷序列相空间重构

在一定程度上单变量时间序列是多变量时间序列的特殊情况,因此可将单变量的重构方法应用到多变量中。单变量相空间重构过程中有两个非常重要的参数:延迟时间 τ 和嵌入维数 m[13]。对单变量时间序列 $x(n)$ $(n=1,2,\cdots,N)$,在确定嵌入维数 m 和时间延迟 τ 后,重构相空间如式(1):

$$X(n)=[x(n),\cdots,x(n-(m-1)\tau)]^{\mathrm{T}}\in \mathrm{R}^m$$
$$(n=N,\cdots,(m-1)\tau+1) \tag{1}$$

设云平台的多变量负荷序列为 $X_1,X_2,\cdots,X_K$。其中,$X_i=[x_i(1),x_i(2),\cdots,x_i(\mathrm{N})]^{\mathrm{T}}$ 表示第 i 个变量的时间序列,K 表示云平台底层资源的种类数。如果每个变量都选择恰当的时间延迟 τ_i 和嵌入维数 m_i $(i=1,2,\cdots,K)$,则多变量时间序列重构的相空间如式(2):

$$\begin{aligned}Y(n)=&[x_1(n),x_1(n-\tau_1),\cdots,x_1(n-(m_1-1)\tau_1),\\&x_2(n),x_2(n-\tau_1),\cdots,x_2(n-(m_1-1)\tau_1),\\&\cdots\\&x_k(n),x_k(n-\tau_1),\cdots,x_k(n-(m_1-1)\tau_1)]^{\mathrm{T}}\in \mathrm{R}^{m_1+\cdots+m_k}\\&n=N,N-1,\cdots,\max_{1\leqslant i\leqslant k}(m_i-1)\tau_i+1\end{aligned} \tag{2}$$

1.1　多变量负荷序列延迟时间的确定

目前确定重构延迟时间 τ 的常用方法有自关联函数法[14]、平均位移法[15]、互信息法[16]三种。由于自关联函数法是对数据序列线性相关性的描述,本质上不适合非线性分析,只是由于计算简便得到了应用。平均位移法是靠经验值,具有随意性,且理论性不强。互信息法是从信息论的角度出发,这种方法既可以分析线性也可以分析非线性系统,在理论上对非线性分析非常合适。

本文利用互信息法确定云平台多变量相空间的延迟时间。设两个离散型随机变量 $X\in\{x_1,x_2,\cdots,x_n\}$ 和 $Y\in\{y_1,y_2,\cdots,y_n\}$,其先验概率分别为 $\{p(x_i)\}_{i=1,2,\cdots,n}$ 和 $\{p(y_j)\}_{j=1,2,\cdots,n}$。信息熵可定义为

$$H(X)=-\sum_{i=1}^{n}p(x_i)\log p(x_i) \tag{3}$$

它描述了随机变量 X 的不定性。类似可以定义联合熵:

$$H(X,Y)=-\sum_{i=1}^{n}\sum_{j=1}^{k}p(x_i,y_j)\log p(x_i,y_j) \tag{4}$$

其中,$p(x_i,y_j)$ 为联合概率。

根据条件概率 $p(x_i|y_j)$ 定义条件熵:

$$H(X|Y)=-\sum_{i=1}^{n}\sum_{j=1}^{k}p(x_i|y_j)\log p(x_i|y_j) \tag{5}$$

它描述了已知 Y 后 X"残留"的不定性。显然,条件熵 $H(X|Y)$ 的值越大,已知 Y 后 X"残留"的不定性也越大,即 Y 和 X 的关联性越弱;反之,$H(X|Y)$ 的值越小,Y 和 X 的关联性就越强。由式(3)~(5)可以推出:

$$H(X|Y)=H(X,Y)-H(Y) \tag{6}$$

变量 Y 和 X 之间的互信息定义为

$$\begin{aligned}I(X;Y)&=H(X)-H(X|Y)=\\&H(X)+H(Y)-H(X,Y)\end{aligned} \tag{7}$$

根据式(7)可推出时间序列和其延迟时间序列的互信息:

$$\begin{aligned}I(\tau_i)&=I(X_i;X_i^{\tau i})=\\&H(X_i)+H(X_i^{\tau i})-H(X_i,X_i^{\tau i})\end{aligned} \tag{8}$$

根据互信息法,计算出互信息函数 $I(\tau_i)$ 在不同延迟 τ_i 下的值,取 $I(\tau_i)$ 的第一个局部极小值点对应的时间 τ_i 为第 i 个变量时间序列 X_i 的延迟时间。

1.2　多变量负荷序列嵌入维数的确定

对于多变量负荷序列,一个很自然的想法就是对每个变量的负荷序列分别求一个嵌入维数,然后将所有变量的重构相空间进行简单迭加,形成多变量负荷序列的相空间如式(2)。但是,由于多变量相空间在迭代过程中包含大量的冗余信息,如何减少这些冗余信息,是确定多变量负荷序列嵌入维数的关键。本文利用主成分分析方法对云平台的多变量相空间向量

进行主成分提取,消除各重构变量间的冗余信息,降低多变量相空间维数。

1.2.1 多变量相空间嵌入维数的确定

目前关于确定嵌入维数 m 的方法有 G-P 法[17]、伪最邻近点法(FNN)[18]和 Cao 方法[19]等。G-P 法通常采用系统特征值饱和法,往往因缺乏先验知识而导致算法冗余,加上关联维数对噪声反应最为敏感,因此造成计算复杂,关联维数波动大等缺点。Cao 方法是伪最邻近点法的改进,能够有效区分随机序列和确定性序列,使用较小的数据量就可求得嵌入维数。

本文利用 Cao 方法确定云平台多变量相空间的嵌入维数。在 m 维相空间中,$X(i)=\{x(i),x(i+\tau),\cdots,x(i+(m-1)\tau)\}$ 为相点矢量,每个 $X(i)$ 都有一个某距离内的最邻近点 $X_N(i)$,距离为 $R_m(i)$ 则有

$$R_m^2(i)=\|X(i)-X_N(i)\|^2=\sum_{j=1}^{m}\{x[i+(j-1)\tau]-x_N[i+(j-1)\tau]\}^2 \tag{9}$$

当相空间维数从 m 维增加到 $m+1$ 维时,这两个相点的距离就会发生变化,两者距离成为 $R_{m+1}(i)$,且有

$$R_{m+1}^2(i)=\sum_{j=1}^{m+1}\{x[i+(j-1)\tau]-x_N[i+(j-1)\tau]\}^2=R_m^2(i)+|x(i+m\tau)-x_N(i+m\tau)|^2 \tag{10}$$

由 m 维增加 1 维而引起的两邻近点间距离的变化是

$$[R_{m+1}^2(i)-R_m^2(i)]^{1/2}=|x(i+m\tau)-x_N(i+m\tau)| \tag{11}$$

可见,如果两邻近点距离不随维数 m 增大而变化,即 $R_{m+1}(i)=R_m(i)$,则邻近点是真实的;如果 $R_{m+1}(i)$ 比 $R_m(i)$ 大很多,可以认为这是由于高维混沌吸引子中两个不相邻的点投影到低维轨道上时变成相邻的两点造成的,因此这样的邻近点是虚假的。令

$$a(i,m)=\frac{\|X_{m+1}(i)-X_{Nm+1}(i)\|}{\|X_m(i)-X_{Nm}(i)\|} \tag{12}$$

定义:

$$E(m)=\frac{1}{N-m\tau}\sum_{i=1}^{N-m\tau}a(i,m) \tag{13}$$

$$E_1(m)=\frac{E(m+1)}{E(m)} \tag{14}$$

如果时间序列是确定的,则嵌入维是存在的,即 $E_1(m)$ 将在 m 大于一定值 m_0 后不再变化,若为随机信号则逐渐增加。但在实际应用不容易判断有限长序列 $E_1(m)$ 是在缓慢变化还是已经稳定,因此补充一个判断准则 $E_2(m)$,定义如下:

$$E^*(m)=\frac{1}{N-m\tau}\sum_{i=1}^{N-m\tau}|x(i+m\tau)-x_N(i+m\tau)| \tag{15}$$

$$E_2(m)=\frac{E^*(m+1)}{E^*(m)} \tag{16}$$

对于随机序列,数据间没有相关性,$E_2(m)$ 将始终为 1;对于确定性序列,数据间的相关关系依赖于嵌入维 m 值变化,总存在一些 m 值使得 $E_2(m)$ 不等于 1。

1.2.2 多变量相空间嵌入维数的主成分分析法

假设云平台包含 K 个底层资源负荷序列类型,这 K 个底层资源相空间的第一个向量[20]分别记为 $Y_1,Y_2,\cdots,Y_K$,将不同底层资源相空间的第一个向量 $Y_q=[x_{q1},x_{q2},\cdots,x_{qm}]^{\mathrm{T}}$ 作为一个变量(其中 $q=1,\cdots,K;m=\min_{1\leqslant h\leqslant K}(m_h)$),由此构造 $m\times K$ 矩阵,矩阵中相应位置的变量表示为 $y_{ij}(i=1,2,\cdots,m;j=1,2,\cdots K)$,利用此矩阵求相关系数矩阵[21]:

$$\boldsymbol{R}=(r_{ij})\quad i=1,2,\cdots,m;j=1,2,\cdots,K \tag{17}$$

其中:

$$\overline{y_j}=\frac{1}{n}\sum_{i=1}^{n}y_{ij} \tag{18}$$

$$r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \tag{19}$$

$$S_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(y_{ki}-\overline{y_i})(y_{kj}-\overline{y_j}) \tag{20}$$

对相关系数矩阵 $\boldsymbol{R}$ 计算其特征值 $\lambda_i(i=1,2,\cdots,p)$,并使其按大小顺序排列 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant0$;然后求出对应于特征值 λ_i 的特征向量 $v_i=(\gamma_{i1},\gamma_{i2},\cdots,\gamma_{im})'$。根据主成分思想,$\lambda_i/\sum_{k=1}^{p}\lambda_k$ 是第 i 种主成分的贡献率,贡献率越大,说明波动幅度越大,那么在云平台底层资源负荷序列预测中所占的影响比例越多。$\sum_{k=1}^{i}\lambda_k/\sum_{k=1}^{p}\lambda_k$ 是第 i 种主成分的累积贡献率。若前几个主成分的累积贡献率超过 85%,则不考虑之后的主成分。利用公式确定主成分为

$$Z=\sum_{j=1}^{p}a_{ij}X_j \tag{21}$$

$$a_{ij}=\frac{\gamma_{ij}}{\sqrt{\lambda_i}} \tag{22}$$

根据主成分分析的比例,确定云平台各种底层资源的嵌入维数为 $m_j(j=1,\cdots,K)$,则多变量相空间的嵌入维数为 $m=m_1+m_2+\cdots+m_K$。

1.3 多变量相空间重构

本文综合互信息法和 Cao 方法的优点,求得参数延迟时间 τ 和嵌入维数 m。在相空间重构的过程中,设 N 是时间序列长度,M 是相空间中点的个数,则 $M=N-(m-1)*\tau$,相空间轨迹的表达式为

$$X(t+\tau)=f(X(t)) \tag{23}$$

$X(t+\tau)$ 可视为 $f(X(t))$ 的映射,则有

$$X(t)=[x(t),x(t+\tau),\cdots,x(t+(m-1)\tau)] \tag{24}$$

上述映射可表示为时间序列

$$X(1)=[x(1),x(1+\tau),\cdots,x(1+(m-1)\tau)]$$
$$X(2)=[x(2),x(2+\tau),\cdots,x(2+(m-1)\tau)]$$
$$\cdots$$
$$X(M)=[x(M),x(M+\tau),\cdots,x(N)] \tag{25}$$

2 多变量局域预测模型

本文构建了多变量局域预测模型,其基本思想如下:首先,基于重构后的云平台多变量相空间,利用广义自由度方法确定相空间中心点的邻域子空间。其次,采用最小二乘法对邻域子空间中的点进行线性拟合,根据拟合趋势确定最终的预测结果。

2.1 确定中心点的邻域子空间

在相空间中,由于选取的邻域点个数不同,对应的输入输出也不相同,因此需要选取适当的邻域点个数,以取得更加精确的预测结果。本文采用基于广义自由度的方法确定邻域点个数 K,基本步骤如下:

a)依次选取邻域点个数 $K=2m+1,2m+2,\cdots,2m+10$。

b)计算每个邻域点数对应的自由度 D 和最小方差

$\sigma^2(K)$,如式(27)(28)所示:

$$H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} \tag{26}$$

$$D(K) = tr(H) \tag{27}$$

$$\sigma^2(K) = \frac{[y - yX^{\mathrm{T}}(XX^{\mathrm{T}})^{-1}X][y - yX^{\mathrm{T}}(XX^{\mathrm{T}})^{-1}X]^{\mathrm{T}}}{K - D(K)} \tag{28}$$

c)选择使 $\sigma^2(K)$ 最小的 K 作为最优邻域点个数。

2.2　多变量局域预测模型

在预测过程中,选取相空间轨迹的最后一点作为中心点。设中心点 X_M 的参考向量集为 $\{X_{Mi}\}$,$i = 1,2,\cdots,q$,其演化 k 步后的相点集 $\{X_{Mi+k}\}$,$k = 1,2,\cdots$;d_i、d_m 分别为邻域子空间中各点到中心点的空间距离和最小距离。定义点 X_{Mi} 权值为

$$P_i = \frac{\exp(-c(d_i - d_m))}{\sum_{i=1}^{q}\exp(-c(d_i - d_m))} \tag{29}$$

其中:c 为参数,一般情况下取值为 1。

构建一阶局域线性预测模型为

$$X_{Mi+k} = a_k e + b_k X_{Mi} \quad i = 1,2,\cdots,q \tag{30}$$

根据加权最小二乘法有

$$\min\sum_{i=1}^{q}P_i\left[\sum_{j=1}^{m}(x^j_{Mi+k} - a_k - b_k x^j_{Mi})^2\right] \tag{31}$$

其中:x^j_{Mi} 是参考向量 X_{Mi} 的第 j 个元素。将式(31)看成是关于未知数 a_k、b_k 的二元函数,两边求偏导并化简得:

$$\begin{cases} a_k\sum_{i=1}^{q}P_i\sum_{j=1}^{m}x^j_{Mi} + b_k\sum_{i=1}^{q}P_i\sum_{j=1}^{m}(x^j_{Mi})^2 = \sum_{i=1}^{q}P_i\sum_{j=1}^{m}x^j_{Mi+k}x^j_{Mi} \\ a_k m + b_k\sum_{i=1}^{q}P_i\sum_{j=1}^{m}x^j_{Mi} = \sum_{i=1}^{q}P_i\sum_{j=1}^{m}x^j_{Mi+k} \end{cases}$$

写成矩阵形式为

$$\begin{pmatrix} \alpha & \beta \\ m & \alpha \end{pmatrix}\begin{pmatrix} a_k \\ b_k \end{pmatrix} = \begin{pmatrix} e_k \\ f_k \end{pmatrix} \tag{32}$$

其中:$\alpha = \sum_{i=1}^{q}P_i\sum_{j=1}^{m}x^j_{Mi}$,$\beta = \sum_{i=1}^{q}P_i\sum_{j=1}^{m}(x^j_{Mi})^2$,$e_k = \sum_{i=1}^{q}P_i\sum_{j=1}^{m}x^j_{Mi+k}x^j_{Mi}$,$f_k = \sum_{i=1}^{q}P_i\sum_{j=1}^{m}x^j_{Mi+k}$,则

$$\begin{pmatrix} a_k \\ b_k \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ m & \alpha \end{pmatrix}^{-1}\begin{pmatrix} e_k \\ f_k \end{pmatrix} \tag{33}$$

根据求得的 a_k、b_k,代入 k 步预测公式 $X_{M+1} = a_k e + b_k X_M$,即可得到演化 k 步后的相点预测值 X_{M+k}:

$$X_{M+k} = (x_{M+k}, x_{M+k+\tau}, \cdots, x_{M+k+(m-1)\tau}) \tag{34}$$

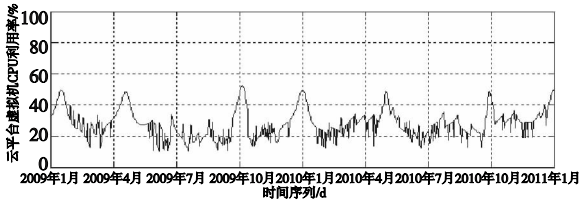
其中:X_{M+k} 中的第 m 个元素 $x_{M+k+(m-1)\tau}$ 即为原序列的 k 步预测值。

3　模型仿真实验

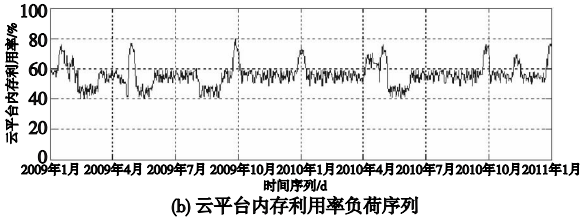
为了检验和评价多变量局域预测模型应用于云平台资源负荷序列的预测效果,本文利用单变量局域预测模型的预测效果与其进行对比。仿真数据来源于某公司 2009 年 1 月 ~2011 年 1 月的云计算监测系统,观测周期为天,共 730 天的观测序列中包含了春节、五一节假日、十一节假日三个使用高峰期,比较具有代表性,可以看出这些序列具有明显的非线性、非平稳特性,适合于利用相空间重构理论对其进行特征提取。

本文以云平台的虚拟机 CPU 利用率序列为例建立预测模型。首先,选取第 1 ~720 天的 CPU 利用率序列,对第 721 ~730 天的 CPU 利用率序列建立单变量局域预测模型,评估预测模型精度;其次,利用 CPU 利用率的三个相关资源(内存利用率、硬盘利用率、网络负载利用率)序列,根据主成分分析法构建针对于 CPU 利用率的多变量局域预测模型。

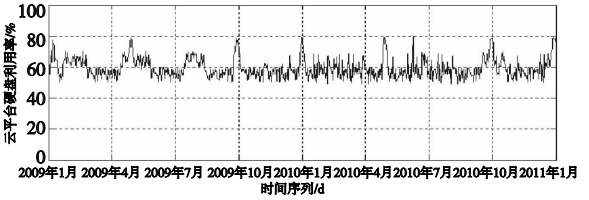
采集的资源负荷序列趋势分别如图 2 所示。其中(a)表示云平台虚拟机 CPU 利用率的负荷序列 x_1;(b)表示云平台内存利用率的负荷序列 x_2;(c)表示云平台硬盘利用率的负荷序列 x_3;(d)表示云平台网络负载利用率的负荷序列 x_4。



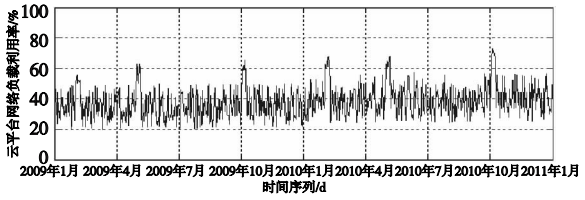
(a) 云平台虚拟机CPU利用率负荷序列



(b) 云平台内存利用率负荷序列



(c) 云平台硬盘利用率负荷序列



(d) 云平台网络负载利用率负荷序列

图2　云平台资源负荷序列

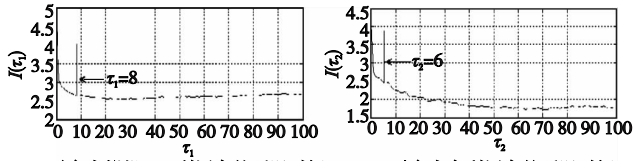
3.1　多变量负荷序列延迟时间的确定

在单变量与多变量相空间重构过程中,均利用互信息法计算时间延迟。首先,确定虚拟机 CPU 利用率序列的时间延迟 τ_1;其次,确定虚拟机 CPU 利用率的三个相关资源的时间延迟 τ_2、τ_3、τ_4。各种资源负荷序列的互信息函数 $I(\tau_i)$($i = 1,2,3,4$)随延迟时间 τ_i 的关系变化如图 3 所示。(a)中可以看出,在 $\tau_1 = 8$ 时,虚拟机 CPU 利用率的 $I(\tau_1)$ 出现第一个极值,故而可得延迟时间 $\tau_1 = 8$;(b)中可以看出,在 $\tau_2 = 6$ 时,内存利用率的 $I(\tau_2)$ 出现第一个极值,故而可得延迟时间 $\tau_2 = 6$;(c)中可以看出,在 $\tau_3 = 2$ 时,硬盘利用率的 $I(\tau_3)$ 出现第一个极值,故而可得延迟时间 $\tau_3 = 2$;(d)中可以看出,在 $\tau_4 = 1$ 时,网络负载利用率的 $I(\tau_4)$ 出现第一个极值,故而可得延迟时间 $\tau_4 = 1$。

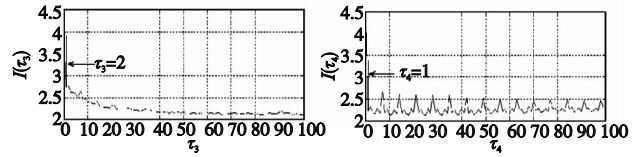
3.2　多变量负荷序列嵌入维数的确定

3.2.1　相空间的嵌入维数确定

在单变量与多变量相空间重构过程中,均利用 Cao 方法计算嵌入维数。首先,确定云平台虚拟机 CPU 利用率序列的嵌入维数 m_1;其次,确定虚拟机 CPU 利用率的三个相关资源的嵌入维数 m_2、m_3、m_4。各种资源负荷序列的 $E_1(m)$ 与 $E_2(m)$ 函数随嵌入维数 m_i($i = 1,2,3,4$)的关系变化如图 4 所示。

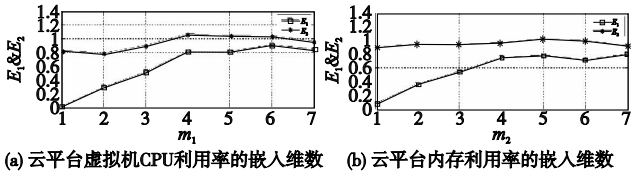


(a) 云平台虚拟机CPU利用率的延迟时间 (b) 云平台内存利用率的延迟时间

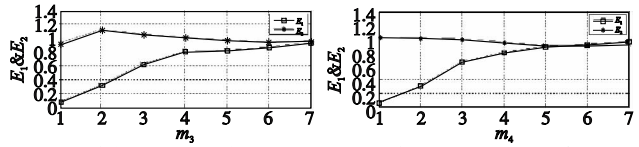


(c) 云平台硬盘利用率的延迟时间 (d) 云平台网络负载利用率的延迟时间

图3 云平台利用率延迟时间



(a) 云平台虚拟机CPU利用率的嵌入维数 (b) 云平台内存利用率的嵌入维数



(c) 云平台硬盘利用率的嵌入维数 (d) 云平台网络负载利用率的嵌入维数

图4 云平台利用率的嵌入维数

从图4(a)中可以看出,虚拟机CPU利用率的$E_1(m)$在m_1大于4时不再变化,故而可得最佳嵌入维数$m_1=4$;$E_2(m)$的值总体离1较远,说明序列的确定性较强。

从图4(b)中可以看出,内存利用率的$E_1(m)$在m_2大于4时不再变化,故而可得最佳嵌入维数$m_2=4$;$E_2(m)$的值总体离1较远,说明序列的确定性较强。

从图4(c)中可以看出,硬盘利用率的$E_1(m)$在m_3大于4时不再变化,故而可得最佳嵌入维数$m_3=4$;$E_2(m)$的值总体离1较远,说明序列的确定性较强。

从图4(d)中可以看出,网络负载利用率的$E_1(m)$在m_4大于5时不再变化,故而可得最佳嵌入维数$m_4=5$;$E_2(m)$的值总体离1较近,说明序列的波动性较强。

由此,根据云平台虚拟机CPU利用率、内存利用率、硬盘利用率、网络负载利用率的时间延迟τ_i与嵌入维数m_i,重构出各种资源负荷序列的单变量相空间。根据文献[10]的方法,计算出各种资源负荷序列相空间的最大Lyapunov指数分别为$\lambda_1=0.1543>0$,$\lambda_2=0.1457>0$,$\lambda_3=0.1254>0$,$\lambda_4=0.1622>0$,表明云平台资源负荷序列中存在混沌特性,根据Takens的嵌入定理,提取和恢复出系统原来的规律。

3.2.2 多变量相空间嵌入维数的主成分分析法

在多变量相空间重构过程中,针对云平台虚拟机CPU利用率的三个相关资源因素(内存利用率、硬盘利用率、网络负载利用率)相空间的向量进行主成分分析。分别将三个相关资源的相空间的第一个向量作为主成分分析法的输入序列Y_1、Y_2、Y_3,得到包含原始信息的主成分。各主成分的特征根和贡献率如表1所示。

由表1可知,第一主成分的贡献率达64.08%,第二主成分的贡献率达33.95%,第一、二主成分的累积贡献率达98.03%,符合累计贡献率达到85%的原则。作为提取的主成分,进一步计算得到第一、二主成分的特征向量,即主成分表达式的变量系数,见表2。

表1 主成分特征根与贡献率

主成分	特征值	贡献率/%	累积贡献率/%
Z_1	1.922 3	64.08	64.08
Z_2	1.018 6	33.95	98.03
Z_3	0.059 1	1.97	100

表2 主成分特征量表

资源类别	Z_1	Z_2
内存利用率	-0.706 2	0.110 1
硬盘利用率	-0.707 8	-0.088 0
网络负载利用率	0.015 6	0.990 0

计算后主成分可用下列线性组合表示为

$$Z_1=-0.5094x_2-0.5105x_3+0.0113x_4$$

$$Z_2=0.1101x_2-0.0880x_3+0.9900x_4$$

综合各主成分的加权平均值,权数为各主成分的贡献率:

$$Y=0.6408\times Z_1+0.3395\times Z_2=$$

$$-0.2890Y_1-0.3570Y_2+0.3433Y_3$$

根据三个相关资源因素在主成分分析中的比例,确定其在虚拟机CPU利用率多变量相空间中的嵌入维数。已知三个相关资源的单变量相空间嵌入维数分别为4、4、5,那么内存利用率的嵌入维数为$4\times0.289=1$,硬盘利用率的嵌入维数为$4\times0.3570=1$,网络负载利用率的嵌入维数为$5\times0.3433=2$。已知虚拟机CPU利用率的嵌入维数为4,则虚拟机CPU利用率的多变量相空间嵌入维数为$m=m_1+m_2+m_3+m_4=4+1+2+1=8$。

3.3 多变量局域预测模型

a)采用基于广义自由度的方法确定相空间中心点的邻域子空间。(a)基于重构后的虚拟机CPU利用率的单变量相空间,由于嵌入维数$m=4$,可知邻域点数为10个;(b)基于重构后的虚拟机CPU利用率的多变量相空间,由于嵌入维数$m=8$,可知邻域点数为21个。

b)分别对虚拟机CPU利用率的单变量与多变量相空间中心点的邻域子空间中的点,采用最小二乘法进行线性拟合,根据拟合趋势确定最终的预测结果。单变量与多变量局域预测模型的预测结果,如图5所示。

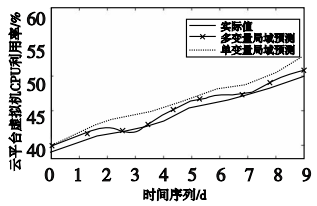


图5 单变量与多变量局域预测模型的预测结果

由图5可以看出,多变量局域预测模型的预测结果曲线比较接近于实际值,而单变量局域预测模型的曲线与实际值相差较大。

对预测结果的误差分析选用平均绝对百分比误差E_{mape}和相对误差E。E_{mape}定义如下:

$$E_{\text{mape}}=\frac{1}{N}\sum_{i=1}^{N}\left|\frac{f_i-y_i}{y_i}\right|\times100\% \tag{35}$$

其中,f_i和y_i分别表示预测值和实际值,N是预测值的个数。相对误差E定义如下:

$$E = \left| \frac{f_i - y_i}{y_i} \right| \times 100\% \quad (36)$$

表 3 列出了单变量与多变量局域预测模型的平均绝对百分比误差 E_{mape} 及相对误差 E 的最大值 $E_{\max}$。

表 3　单变量与多变量局域预测模型的预测误差对比

预测结果	E_{mape}/%	$E_{\max}$/%
单变量局域预测	3.07	6.83
多变量局域预测	2.71	5.61

根据表 3 可以看出,多变量局域预测模型的平均绝对误差 E_{mape} 只有 2.71%,相对误差 E 的最大值是 5.61%,与单变量局域预测模型相比优势明显,预测精度较高。

4　结束语

本文提出了面向云计算的主成分分析多变量局域预测模型。一方面解决了传统单变量预测模型无法获取完整系统信息的问题,另一方面消除了多变量相空间的冗余变量,提高了预测模型的精度及效率。仿真实验结果表明:面向云计算的主成分分析多变量局域预测模型的预测精度要明显优于单变量预测模型的精度。因此,进一步探索基于主成分分析的多变量局域预测模型,对提高云计算底层资源的预测精度很有意义:

a)云平台的底层资源预测一方面为合理分配及调度底层资源提供了依据,优化运营结构,降低运营成本;另一方面保证了云平台的稳定性,为云平台能否持续提供服务作出判断。

b)云平台的底层资源预测为用户的资源租赁决策提供量化的方法,以便满足需求各异的不同租户在共享资源的基础上构建专属于自己的业务流程。

参考文献:

[1] SUN Wei, ZHANG Kuo, ZHANG Xin, *et al*. Software as a service: an integration perspective[C]//Proc of the 5th International Conference on Service-Oriented Computing. Berlin: Springer-Verlag, 2007: 558-569.

[2] TAKENS F. Detecting strange attractors in turbulence[J]. Lecture notes in Math, 1981, 898: 361-381.

[3] 张春涛,马千里,彭宏,等. 基于条件熵扩维的多变量混沌时间序列相空间重构[J]. 物理学报,2011,60(2):1-5.

[4] 卢山,王海燕. 多变量时间序列最大李雅普诺夫指数的计算[J]. 物理学报,2006,55(2):572-575.

[5] CAO L Y. Practical method for determining the minimum embedding dimension of a scalar time series[J]. Physica D: Nonlinear Phenomena, 1997, 110(1-2): 43-50.

[6] 樊重俊. 多观测变量状态空间重构技术及应用[J]. 上海理工大学学报,2009,31(3):273-277.

[7] 刘宝英,杨仁刚. 基于主成分分析的最小二乘支持向量机短期负荷预测模型[J]. 电力自动化设备,2008,28(11):13-16.

[8] 吕金虎,陆君安,陈士华. 混沌时间序列分析及其应用[M]. 武汉:武汉大学出版社,2002.

[9] 丁涛,周惠成. 混沌时间序列局域预测法[J]. 系统工程与电子技术,2004,26(3):338-340.

[10] 杨绍清,章新华,赵长安. 一种最大李雅普诺夫指数估计的稳健算法[J]. 物理学报,2000,49(4):636-640.

[11] CAI Ming-lun, CAI Feng, SHI Ai-guo, *et al*. Chaotic time series prediction based on local-region multi-steps forecasting model[M]// Advances in Neural Networks. Berlin: Springer-Verlag. 2004: 418-423.

[12] JAYEWARDENA A W, LI W K, XU P. Neighborhood selection for local modeling and prediction of hydrological time series[J]. Journal of Hydrology, 2002, 258: 40-57.

[13] 李智勇,吴为麟. 基于相空间重构的电能质量扰动特性提取方法[J]. 浙江大学学报:工学版,2008,42(4):672-675.

[14] ALBANO A M, MUENCH J, SCHWARTZ C. Singular-value decomposition and the Grassberger-Procaccia algorithm[J]. Physical Review A, 1988, 38(6): 3017-3026.

[15] ROSENSTEIN M T, COLLINS J J, De LUCA C J. Reconstruction expansion as a geometry-based framework for choosing proper delay times[J]. Physica D, 1994, 73(1/2): 82-98.

[16] FRASER A M, SWINNEY H L. Independent coordinates for strange attractors from mutual information[J]. Physical Review A, 1986, 33(2): 1134-1140.

[17] 孙金花,胡健,李向阳. 基于分形理论的离群点检测[J]. 计算机工程,2011,37(3):33-25.

[18] SANGOYOMI T B. Nonlinear dynamics of the Great Salt Lake: dimension estimation [J]. Water Resources Research, 1996, 32(1): 975-985.

[19] 张淑清,贾健,高敏,等. 混沌时间序列重构相空间参数选取研究[J]. 物理学报,2010,59(3):1576-1582.

[20] 韩敏,魏茹. 基于相空间同步的多变量序列相关性分析及预测[J]. 系统工程与电子技术,2010,32(11):2426-2430.

[21] 牛京考. 基于主成分回归分析法预测中国铁矿石需求[J]. 北京科技大学学报,2011,33(10):1177-1181.

(上接第 4164 页)

参考文献:

[1] 计智伟,胡珉,尹建新. 特征选择算法综述[J]. 电子设计工程,2011,19(9):46-51.

[2] 黄海燕,顾幸生,刘漫丹. 求解约束优化问题的文化算法研究[J]. 自动化学报,2007,33(10):1115-1120.

[3] ROBERT R G. An introduction to cultural algorithms[C]//Proc of the 3rd Annual Conference on Evolution Programming. Singapore: World Scientific Publishing, 1994: 131-136.

[4] 焦李成,杜海峰,刘芳,等. 免疫优化计算、学习与识别[M]. 北京:科学出版社,2006:92-99.

[5] CHANG Zhi-yang, HAN Li, JIANG Da-wei. Improved clone selection algorithm and its application[J]. Computer Engineering, 2011, 37(1): 173-174.

[6] BLAKE C L, KEOGH E, MERZ C J. UCI repository of machine learning databases[M]. Irvine, CA: University of California, 1998.

[7] 薄翠梅,张浞,张广明,等. 基于特征样本核主元分析的 TE 过程快速故障辨识方法[J]. 化工学报,2008,59(7):1783-1789.

[8] DASH M, LIU H. Feature selection for classification[J]. Intelligent Data Analysis: An International Journal, 1997, 1(3): 131-156.

[9] BURGES C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 127-167.

分类率,本文所设计的适应度函数为

$$f(x_i) = A(i) + \left(\frac{1}{m_c}\sum_{i,j=1}^{m_c}\omega_{ij}\right)$$

其中:$A(i)$为用该个体进行特征选择所得到的特征变量在故障诊断中的正确率,m_c 是抗体中被选中的变量总和,ω_{ij}为归一化的海明距离。

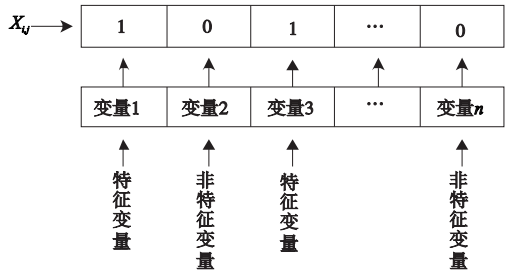


图3 粒子编码

基于免疫文化算法的故障特征选择方法基本步骤如下:

a)选择支持向量机作为分类器,对采集到的原始数据进行预处理,根据采集数据构造训练集和测试集;

b)初始化抗体种群,设定免疫文化算法参数,对抗体进行二进制编码,计算初始种群的适应度;

c)执行上一章中免疫文化算法的步骤;

d)根据抗体编码确定相应的特征变量,并生成相对应的故障训练数据集和测试数据集,对 SVM 进行训练和测试,得到不同抗体的故障正确分类率;

e)产生一个记忆抗体,储存于文化算法的信念空间,代表到目前为止所获得的最优特征子集。

以上五个步骤全部执行一次为一代。

3 仿真实验及结果分析

为了验证基于免疫文化算法的故障特征选择方法的有效性,本文选择 UCI 数据集[6]中的 wine、vote 二类数据进行测试,详细信息见表 1。

表 1 测试数据统计特性

数据集	样本数量	维数	类别数
wine	178	13	3
vote	435	16	2

对于 wine 数据,从三类中分别选取 29、30、30 个数据作为训练集,其余 89 个数据为测试集;从 vote 数据两类中各选取 190 个共计 380 个数据作为训练集,其余 155 个数据为测试集。

算法参数选取如下:抗体种群规模为 30,迭代次数为 50,变异率为 0.03,抗体编码长度为数据集维数 L,种群更新率为 15%,支持向量机核函数选用 RBF 径向基核函数 $K(x,x_i) = \exp(-\gamma \| x - x_i \|^2)$,$\gamma = 3$,惩罚系数 $C = 4$,不敏感损失函数 Epsilon = 0.001。计算结果取 10 次计算的平均值。通过该算法对 wine、vote 数据进行特征变量选择情况如图 4 和 5 所示。

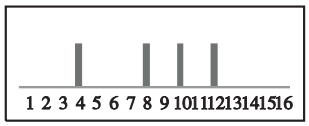


图4 Wine数据特征变量选择结果 图5 Vote数据特征变量选择结果

通过实验可以看出,该特征选择方法可以准确地找到最能体现 wine、vote 数据特征的特征变量。从图 6 和 7 可以看出,对于 vote 数据分类准确率达到 100%,对 wine 数据达到 98%以上,只有第三类中的一个样本点被错分到了第二类。为了验证该算法在工业过程故障特征选择的准确性,将其用于 TE 过程的故障分类中。

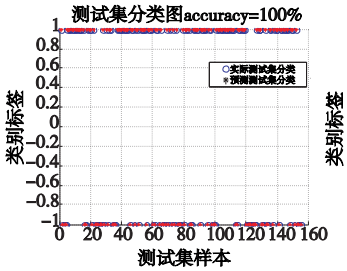


图6 Vote数据分类图

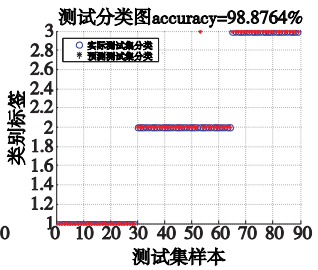


图7 Wine数据分类图

TE 过程[7]是由美国 Tennessee Eastman 化学公司过程控制小组提出的一个基于实际工业过程的仿真案例,包括 41 个测量变量和 12 个操作变量(控制变量)以及 21 个预先设定好的故障。本文分别以 TE 过程中的故障 8、14 和 20 为对象进行故障特征选择。算法参数选择同上,测试样本是故障从 8 小时后引入,即从第 160 组数据开始采集,输入变量共取 1 ~ 22 维;训练样本是故障从 1 小时后引入,即从第 20 组数据开始采集,输入变量共取 1 ~ 22 维,并将上述所得的数据先进行预处理、归一化后再进行故障特征选择。

如图 8 所示,如果对采集到的高维故障数据不进行特征选择而直接进行分类,其分类准确率为 80%,预测会分到故障 20 的部分数据在实际分类中被分到了故障 8 和 14 中;经过基于免疫文化算法的特征选择后,得到降维数据的故障分类准确率达到了 90%,其中故障 14 的预测分类基本与实际分类重合,显示了良好的分类效果,如图 9 所示。由验证了本文提出的基于免疫文化算法的特征选择方法的有效性,大幅提高了分类准确率。

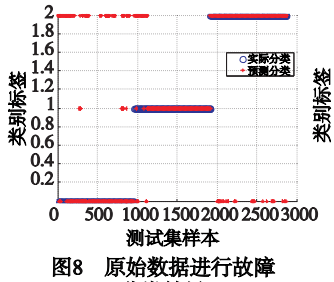


图8 原始数据进行故障分类结果

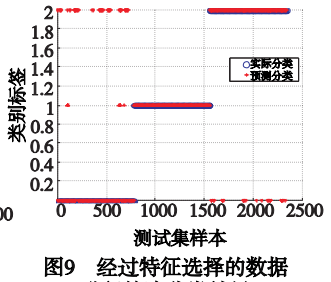


图9 经过特征选择的数据进行故障分类结果

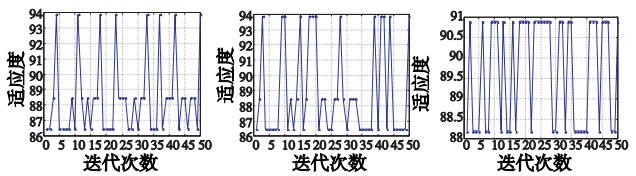


图10 三类故障的适应度曲线

4 结束语

本文提出了一种基于免疫文化算法的封装式特征选择方法,并采用 SVM 分类器对其性能进行评估,将其运用到 TE 过程故障诊断中,通过实验验证了该方法的可行性和有效性。实验表明该方法在降低数据维度和提高分类准确率上有着良好的效果。由于 SVM 主要用于小样本的分类,因此在以后的工作中将尝试将该方法用于其他分类器中,进一步提高特征选择的质量和效率。

(下转第 4175 页)