一种基于自适应高斯过程的基线计算算法

杜占玮 杨永健 肖 敏 白 媛
(吉林大学计算机科学与技术学院 长春 130012)

摘 要 基于自适应高斯过程技术,提出了一种计算网络主动监控中上下基线的新方法,即在满足大型服务器集群对负载性能告警的设置与屏蔽需求下,利用样本噪音的统计特征,结合样本的数据分布,解决了样本数据的回归预测。算法首先分析样本历史数据的噪音,通过结合蚁群算法,提出高斯过程的参数自适应机制,最后实现上下基线的计算。实验结果表明,与其它基线计算算法相比,此算法可以在保证相同准确性的基础上,较大幅度地提高计算效率,保障网络安全,提升网络性能和用户满意度。

关键词 基线计算,高斯过程,机器学习,蚁群算法
中图法分类号 TP391 **文献标识码** A

Baseline Algorithm Based on Adaptive Gaussian Process Machine Learning

DU Zhan-wei YANG Yong-jian XIAO Min BAI Yuan
(College of Computer Science and Technology, Jilin University, Changchun 130012, China)

Abstract The baseline calculation is an important issue in the field of network monitoring. As to deal with the data, most researches just ignore the probability characteristics of the data, which fails to combine data distribution to predict the data and make the related processing. Therefore, this article analysed the historical data's noise first, then made the prediction with the Gaussian process machine learning, and combined with ant colony algorithm, achieved the adaptive mechanism of the parameters, and then calculated the baselines. The experiment shows that compared with other algorithms, our algorithm improves efficiency and accuracy.

Keywords Baseline algorithm, Gaussian process, Machine learning, Ant colony algorithm

1 引言

3G及以上通信网络软交换的普遍采用,使大规模交换机性能、告警、配置和基础数据的实时采集成为可能。由于数据关联复杂,数据处理的实时性要求严格,数据量大,软交换网络配置的可变性、安全性以及故障诊断的实时性要求,采用一般的带有逻辑判断语句的程序显然无法适应这种复杂多变的情况[1]。

早期的网络管理一般采用简单的网络监视手段。网络监视包括收集有关网络状态的信息,将信息综合为关于网络运行状况的表征,并以一种直观和可以理解的方式显示给网络管理人员,但是无法为用户提供任何启发式信息和智能帮助。并且,目前已有的网络管理都属于事后"灭火",没有在技术上真正实现根据用户需求和业务发展情况对通信网络的"主动监控",业务和运维管理系统各自为政,未实现预防故障。

因而,为了实现"主动监控",满足性能告警的设置与屏蔽的需求,提出了一套主动监控算法,其中包括3部分:"软交换网络性能预测算法"、"主动监控实时性能告警算法"和"主动软交换网络优化算法"。算法通过采集交换机海量网络性能、告警、配置和基础数据,对时间序列进行智能分析,实现网络性能主动监控功能、告警和网络优化。"基线算法"、"容忍度计算方法"、"告警产生机制"共同构成"主动监控实时性能告警算法",其中,基线的计算是一个关键环节,容忍度和告警产生机制都是在基线的基础上发展起来的。

目前已有的基线计算算法主要包括一次、二次等多项式拟合算法、概率算法、排序算法、小波理论、神经网络算法[1],主要分为两种[2],即静态基线算法和动态基线算法,不同指标适用一种或多种算法,用于针对不同的监控目的。静态基线的取定方法主要包括手工设定方式和自动设定方式,手工设定方式能够吸收运维人员经验,具备一定的灵活性。与管理需求、设备自身能力有关的指标可以采用手工设定方式,以监控是否低于考核值或者超出处理能力的相应比例;其它适用于该算法的指标由于有一定的波动范围,符合一定的统计规律,因此可以采用自动设定方式,即通过系统的自动学习,按照面的动态基线算法原理产生相应的闭值,自动设定。在指标较少、波动幅度不大的情况下,该方法能够适应工作需要。但是,在纳入主动监控的指标、设备数量较多的情况下,该方式难以适应运维需求,工作效率低下;对于波动幅度较大的指

到稿日期:2011-12-08 返修日期:2012-03-16 本文受吉林省发改委高新技术项目(20106421),吉林省重点科技发展项目(20100309),吉林省教育厅科学技术研究项目(2012184)资助。

杜占玮(1988—),男,博士生,主要研究方向为无线网络,E-mail:du8491@163.com;杨永健(1960—),男,博士,教授,博士生导师,主要研究方向为无线网络研究,E-mail:yyj@jlu.edu.cn(通信作者)。

标，容易产生设置随意、主观性强的问题，不利于系统维护。

此外，静态基线算法对一个指标在 24 小时监控周期内设置同一水平的闭值，仅仅适用于那些波动不大的指标[3-9]。对于在不同的时间段波峰、波谷差别较大的指标来说，必须针对不同时段设定不同的门限，确定不同时段内指标值的合理分布区域、异常分布区域，因而产生了动态基线算法。已有的动态基线算法主要分为两种，即概率法和排序法。它们都可以根据公司、业务系统、管理等要求，动态调整有效数据比例，实用性强。前者主要以波动程度为指标值计算基线，可以使系统波动程度降低；后者对预处理后的有效数据进行排序，取中间的数字为正常波动区间，可以使系统在序列的角度上，波动最小。

上述算法的共同缺点是忽略了样本噪音的统计特征，未能结合样本的数据分布实现样本的预测和相关处理。因此，本文通过对样本历史数据的噪音分析，引入了高斯过程方法，实现了样本数据的回归预测分析；并且结合蚁群算法[10-13]，提出了高斯过程的参数自适应机制，以计算网络主动监控中的上下基线。最后，通过实验，将本文提出的算法与其它基线计算算法相对比。结果表明，本文提出的算法可以在保证准确性的基础上，较大幅度地提高计算效率，保障网络安全，提升网络性能和用户满意度。

2 相关知识

2.1 基线算法

为了实现“主动监控”，满足性能告警的设置与屏蔽的需求，在项目实施过程中，我们实现了一套主动监控算法，其包括 3 部分：“软交换网络性能预测算法”、“主动监控实时性能告警算法” 和“主动软交换网络优化算法”。算法通过采集交换机海量网络性能、告警、配置和基础数据，对时间序列进行智能分析，实现网络性能主动监控功能、告警和网络优化。其中，“主动监控实时性能告警算法”由“基线算法”、“容忍度计算方法”、“告警产生机制”共同构成[2]，如图 1 所示。

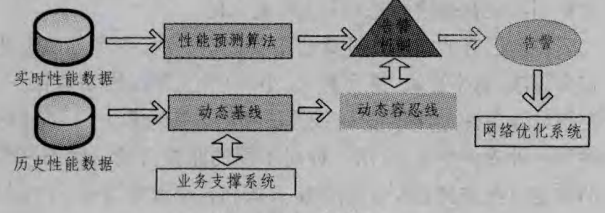

图 1 性能告警算法

计算基线：选取历史正常值为样本空间，经过统计分析得出性能指标的正常波动范围，画出上下基线，如图 2 所示。

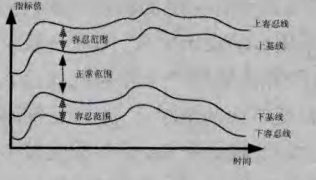

图 2 上下基线示意图

在获得基线之后，通过选取一个合理的容忍度，对基线上(或下)浮动产生容忍线，将其作为触发后续告警产生机制的阈值，当实时监控值超出上(或下)容忍线后，根据性能指标的特性，触发不同的告警产生机制。

2.2 高斯过程学习算法

高斯过程算法是由 Rasmussen 和 Christopher 于 2004 年提出的一类有监督的机器学习算法。该算法自从提出以来一直备受机器学习领域的关注。它以贝叶斯框架为基础，用来研究非线性预测问题。高斯过程由随机变量集合而成，集合中任意的随机变量组合服从联合高斯分布，高斯过程由均值函数和协方差函数来确定。高斯过程模型在统计学中的高斯过程描述为：设连续型随机变量 X 的概率密度为 $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$，$-\infty<x<+\infty$，其中 $\mu,\sigma(\sigma>0)$ 为常数，则称 X 服从参数为 μ,σ 的正态分布或高斯分布，记为 $X\sim N(\mu,\sigma^2)$。与神经网络、支持向量机等方法相比，高斯过程是一种非参数概率模型，它在对未知输入做出预测的同时给出该预测的概率估计，并且高斯过程预测模型中的参数较少，参数优化相对容易，更容易收敛。

高斯过程回归模型的基本原理：

假设观察目标值 y 被噪音腐蚀，它与真实输出值 t 相差 ε：

$$y=t+\varepsilon \tag{1}$$

式中，ε 为独立的随机变量，符合高斯分布，均值为 0，方差为 σ_n^2，即

$$\varepsilon\sim N(0,\sigma_n^2) \tag{2}$$

观察目标值 y 的先验分布为：

$$y\sim N(0,K+\sigma_n^2 I) \tag{3}$$

式中，$K=K(X,X)$ 为 $n\times n$ 阶对称正定的协方差矩阵，矩阵中的任一项 K_{ij} 度量了 x_i 和 y_j 的相关性。

n 个训练样本的输出 y 和 1 个测试样本的输出 y^* 所形成的联合高斯先验分布为：

$$\begin{bmatrix} y \\ y^* \end{bmatrix}\sim N\left\{0,\begin{bmatrix} [K(X,X)+\sigma_n^2 I & K(X,x^*) \\ K(X,x*)^{\mathrm{T}} & k(X^*,X^*) \end{bmatrix}\right\} \tag{4}$$

式中，$K(X,x^*)$ 是测试点 x^* 与训练集的所有输入点 X 的 $n\times 1$ 阶协方差矩阵，可简写为 $k(x^*)$；$k(x^*,x^*)$ 是测试点 x^* 自身的协方差。

$$\begin{aligned} L&=\ln p(y|X) \\ &=-\frac{1}{2}y^{\mathrm{T}}(K+\sigma_n^2 I)^{-1}y-\frac{1}{2}\ln|K+\sigma_n^2 I|-\frac{n}{2}\ln 2\pi \end{aligned} \tag{5}$$

获得最优超参数后，就可以进行预测，具体过程是：根据贝叶斯原理在训练集的基础上预测出与 x^* 对应的最可能的输出值。采用贝叶斯原理的目的是利用观察到的真实数据不断更新概率预测分布，即给定新的输入 y^* 的最大可能的预测后验分布 $p(y^*|x^*,X,y)$，预测后验分布是高斯型的：

$$y^*|x^*,X,y\sim N[\hat{y}(x^*),\sigma(x^*)] \tag{6}$$

y^* 的均值和方差为：

$$\hat{y}(x^*)=k^{\mathrm{T}}(x^*)(K+\sigma_n^2 I)^{-1}y \tag{7}$$

$$\sigma(x^*)=k(x^*,x^*)-k^{\mathrm{T}}(x^*)(K+\sigma_n^2 I)^{-1}k(x^*) \tag{8}$$

3 基于蚁群的高斯过程算法

首先，需要分析样本噪音的概率特征，判断其是否符合高斯分布。为了检验样本噪音是否符合高斯分布，首先假设每天采集数据的噪音都是符合高斯分布的，根据高斯分布的性

质，可知：

如果 $X\sim N(\mu_X,\sigma_X^2)$ 与 $Y\sim N(\mu_Y,\sigma_Y^2)$ 是统计独立的常态随机变量，那么：

它们的和 U 也满足常态分布，即

$U=X+Y\sim N(\mu_X+\mu_Y,\sigma_X^2+\sigma_Y^2)$

它们的差 V 也满足常态分布，即

$V=X-Y\sim N(\mu_X-\mu_Y,\sigma_X^2+\sigma_Y^2)$

U 与 V 两者是相互独立的。

根据式(1)，假设 $y_i=t+\varepsilon_i$，其中 y_i 表示第 i 天采集到的数据，t 表示当天的真实数据，ε_i 表示第 i 天采集到数据的噪音，需要检测噪音 ε_i 是否符合 $\varepsilon\sim N(0,\sigma_n^2)$。由于 t 未知，因此计算

$$y_i-y_{i-1}=\varepsilon_i-\varepsilon_{i-1}\sim N(0,\sigma_i^2+\sigma_{i-1}^2) \tag{9}$$

判断 $\varepsilon_i-\varepsilon_{i-1}\sim N(0,\sigma_i^2+\sigma_{i-1}^2)$ 是否成立，从而在一定程度上反应噪音的概率分布特性。为使用高斯过程机器学习方法，对其进行预测提供理论依据。

考虑到基线的特性，在使用高斯过程机器学习方法计算基线的时候，会产生一条预测曲线 mu。依据该预测曲线，然后产生置信区间$[mu-LOW*\sigma_\varepsilon,mu+UP*\sigma_\varepsilon]$，从而对应下基线和上基线，其中，$\sigma_\varepsilon$ 代表有高斯过程估算出的样本噪音方差，LOW 代表下基线范围。UP 代表上基线范围，在经典统计理论中，LOW 和 UP 为一个值，代表置信度，这里为了基线计算的方便，将其分为两个值。

LOW 和 UP 的合理取值，对下基线和上基线的性能有很大的影响，这里使用启发式的蚁群算法实现参数自适应配置。

对于该优化问题的求解，蚁群算法对应的思路是：(1)基于问题的背景，估计最优解的大致范围，估计出各个变量 x_i 的上限 u 和下限 l：$l\leqslant x_i\leqslant u$。在可行域内绘制网格，每一个网格点都代表一个不同的状态，蚂蚁在网格点间来回移动，并且留下相应的信息量。循环一段时间后，目标值较小的网格点上对应的信息量相对较大。以信息量为依据，搜索信息量较大的网格点，并缩小变量范围，蚂蚁在此点附近进行移动，并且重复上述的过程，直到满足一定的条件，算法最后结束[14]。

上述的蚁群算法总结如下[10-13]：

1)估计各个变量的上限和下限；

2)对每个变量进行 N 等份划分；

3)若 $\max(h_1,h_2,\cdots,h_n)<\varepsilon$，算法结束，输出最优解；

4)$nc\leftarrow 0$(nc 为最大迭代次数)，初始化蚁群信息素矩阵；

5)每只蚂蚁按概率选择下一个节点；

6)更新信息素方程，且 $nc\leftarrow nc+1$；

7)直到迭代到最大次数，从信息素矩阵中找到每一列的最大元素对应的行，组合成$(m_1,m_2,\cdots,m_n)$，并且减小变量相应的取值范围。之后，跳转到第 2)步。

其中，按照如下算式减小变量相应的取值范围：

$$l_{x_i}=l_{x_i}+(m_i-\Delta)h_i \tag{10}$$

$$u_{x_i}=u_{x_i}+(m_i+\Delta)h_i \tag{11}$$

式中，l_{x_i} 是 x_i 的下限，u_{x_i} 是 x_i 的上限，m_i 是 x_i 的对应列的最大元素所对应的行，Δ 是 x_i 上限或下限中行值变化的增量值，h_i 是算法第 2)步中变量 x_i 被 N 等分后，其中一份的值。

这里，蚁群算法中的适应值函数设定为测试样本落入由上下基线包围的区间的数量。

4 实验结果与分析

为了验证本文算法的性能，做了如下一些实验。分别使用 RBF 神经网络和本文的高斯过程机器学习算法对样本库求解。

4.1 样本噪音的概率特征分析

采集一个月(28 天)的历史数据，每天的数据集大小为 1440，使用前一天的数据减去后一天的数据得到新的数据，并对其进行正态分布分析，可以得到图 3 和表 1。可以明显地看出，数据基本符合高斯分布，从而在一定程度上验证了样本的高斯分布特性。

表 1 样本噪音正态分布参数估计

序列	均值	方差	序列	均值	方差
1	−0.00926	89.8436	15	0.015926	74.004
2	−0.01037	109.5037	16	−0.01556	67.63914
3	−0.05259	111.2537	17	−0.01926	63.60409
4	−0.03889	114.051	18	0.005556	61.2816
5	−0.05667	113.5536	19	−0.0363	56.07392
6	−0.04037	101.2624	20	0.031481	61.55756
7	−0.03444	97.63631	21	−0.00741	65.87238
8	−0.04778	95.02861	22	0.041481	65.30578
9	−0.06778	91.03038	23	0.012963	64.86641
10	−0.06	89.94092	24	−0.01926	62.28233
11	−0.01185	84.50101	25	−0.02037	62.06943
12	−0.03963	86.82521	26	−0.03074	59.82396
13	−0.0137	80.09664	27	−0.0137	56.75401
14	−0.01593	73.64197			

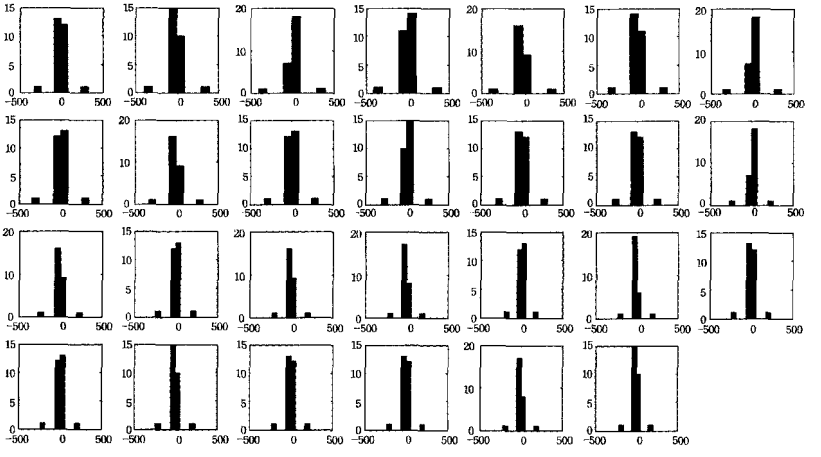

图 3 样本噪音正态分布分析

4.2 样本高斯过程分析

选取其中一天作为高斯过程机器学习样本集，大小为1440，训练时每6个数据为一组，其前5个样本作为输入，后1个作为输出，从而生成输入集和输出集，大小为1335。选取前1000组作为训练样本集，后335组作为测试样本集。分别训练高斯过程机器学习方法和RBF神经网络，最后可以得到图4－图7。可以比较直观地发现，高斯过程机器学习方法比RBF神经网络的误差小很多，前者对测试样本的预测结果与真实值有较高的匹配。而RBF神经网络预测方法由于训练样本偏少而使预测结果与真实值偏差较大。

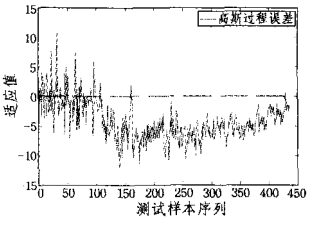

图4 高斯过程机器学习方法预测误差图

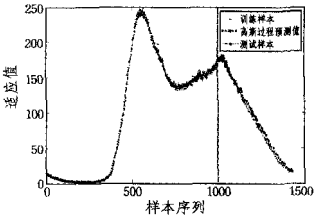

图5 高斯过程机器学习方法预测结果

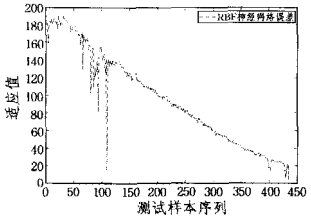

图6 RBF神经网络预测误差图

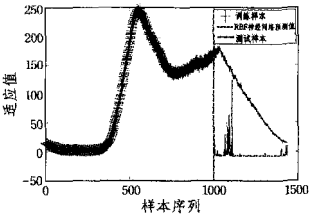

图7 RBF神经网络预测结果

4.3 基于蚁群的高斯过程算法参数自适应

通过多次实验，高斯过程机器学习算法的主要参数设定如下：*LOW* 和 *UP* 的区间设置为[1,3]，蚂蚁数 Nants 设置为100，区间数设置 N 为50，最大迭代数设置为10，信息素的更新速度 $\rho=0.99$，信息素初值 $\tau_0=0.01$。从图8－图11可以发现，在适应值达到一定高度后，*UP* 和 *LOW* 的组合最小的一对为(2.76,2.92)，其对应的上、下基线如图8所示，相对于95%的置信度区间，即 *UP* 和 *LOW* 的值为(3,3)，可以发现，其有了一定的改进。

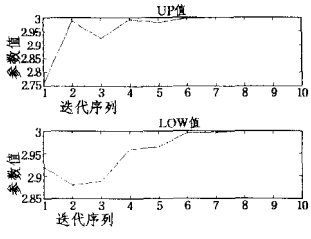

图8 UP和LOW变化曲线

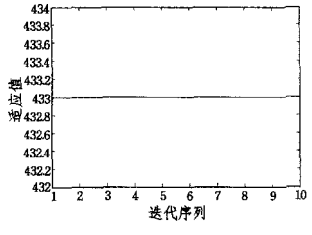

图9 适应值变化曲线

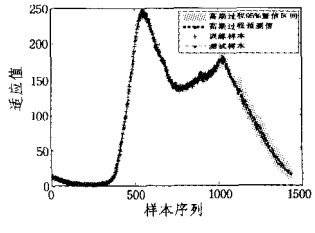

图10 高斯过程机器学习方法的95%置信区间

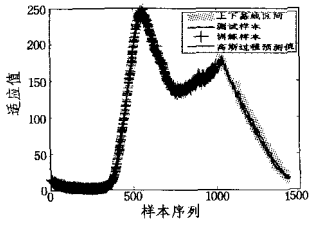

图11 参数自适应后的高斯过程机器学习方法上下基线区间

结束语 本文首先通过对样本历史数据的噪音分析，引入高斯过程机器学习方法对样本数据进行分析，并结合蚁群算法，实现了高斯过程机器学习方法的参数自适应机制，实现了网络主动监控中的网络性能预测算法。实验证明，与以往的基线计算方法相比，该方法可以实现更高的精确度，并且较好地利用了噪声的统计特征。以后的工作将进一步研究在噪声环境下数据集依据时间序列的分段的算法，同其他的分段方法进行全面的比较，并且将其推广到多变量的时间序列。

参 考 文 献

[1] 韩玲. 实时网络流量监测系统的研究与实现[M]. 北京：北京邮电大学，2011：20-21

[2] Jie Y, Ling H, Yun X. The research on Real-Time Alarm Algorithm in network traffic monitoring systrm[C]// 2010 3rd IEEE International Conference on Broadband Network and Multimedia Technology, IC-BNMT2010. IEEE Computer Society, 2010：243-246

[3] Crewther B T, et al. Baseline Strength Can Influence the Ability of Salivary Free Testosterone to Predict Squat and Sprinting Performance[J]. Journal of Strength and Conditioning Research, 2012, 26(1)：261-268

[4] Lass J H. Baseline Factors Related to Endothelial Cell Loss Following Penetrating Keratoplasty [J]. Archives of Ophthalmology, 2011, 129(12)：1640-1640

[5] Loeb S, et al. Baseline Prostate-Specific Antigen Testing at a Young Age[J]. European Urology, 2012, 61(1)：1-7

[6] Martinot M L P, et al. Baseline Brain Metabolism in Resistant Depression and Response to Transcranial Magnetic Stimulation [J]. Neuropsychopharmacology, 2011, 36(13)：2710-2719

[7] Sandborn W J, et al. Baseline C-reactive protein (CRP) and plasma anti-TNF concentration in patients with active Crohn's disease treated with certolizumab pegol[J]. Inflammatory Bowel Diseases, 2011, 17：S20-S21

[8] Seftel A, Ni X, Mckay L. Baseline Factors Associated with Incomplete Response to Tadalafil on-Demand：Analysis of Pooled Data from 17 Randomized Clinical Studies[J]. Journal of Sexual Medicine, 2011, 8：396-397

[9] Verhoeven J J, et al. Baseline insulin/glucose ratio as a marker for the clinical course of hyperglycemic critically ill children treated with insulin[J]. Nutrition, 2012, 28(1)：25-29

[10] Xiao J, Li L P. A hybrid ant colony optimization for continuous domains[J]. Expert Systems with Applications, 2011, 38(9)：11072-11077

[11] Qi C M. Ant Colony Optimization with Local Search for Continuous Functions[J]. Advanced Research on Industry, Information Systems and Material Engineering, Pts 1-7, 2011, 204-210：1135-1138

[12] Chen G H, Lu Y. Application of a Novel Ant Algorithm Termed Continuous Gridded in Aidded Drug Design[J]. Chinese Journal of Chemistry, 2011, 29(10)：2019-2026

[13] Chen L, Sun H Y, Wang S. Solving Continuous Optimization Using Ant Colony Algorithm[C]// 2009 Second International Conference on Future Information Technology and Management Engineering. Fitme 2009, 2009：92-95

[14] 黄翰，郝，吴春国，等. 蚁群算法的收敛速度分析[J]. 计算机学报，2007，30(8)：1344-1353