

基于轨迹聚类的公共安全异常检测

康 凯, 王家宝, 刘方鑫

KANG Kai, WANG Jiabao, LIU Fangxin

解放军理工大学 指挥信息系统学院, 南京 210007

College of Command Information Systems, PLA Science and Technology University, Nanjing 210007, China

KANG Kai, WANG Jiabao, LIU Fangxin. Anomaly detection of public safety based on trajectory clustering. *Computer Engineering and Applications*, 2016, 52(14): 7-11.

Abstract: The demand of public security anomaly detection is becoming more urgent. The approaches based on trajectories clustering become more popular in surveillance, but the existing methods are not good at high-dimensional and unequal-length trajectories. So this paper presents a new approach to cluster trajectories by combining the dynamic time warping and density-peak algorithm. It measures the distance between trajectories by dynamic time warping, and then clusters trajectories by density-peak cluster algorithm. Dynamic time warping can be directly used to measure the distance of trajectories through non-uniform sampling. Density-peak algorithm is a recently proposed cluster algorithm for non-spherical distribution data by combining the local density and the nearest distance. Experiments are conducted on PETS 2006 surveillance video datasets, and results prove that the proposed approach has an effective ability to discover anomaly patterns.

Key words: trajectory clustering; anomaly detection; density peak algorithm; public safety

摘 要: 公共安全异常检测的需求越来越迫切, 监控中基于轨迹聚类的检测方法越来越流行, 但是现有方法在处理高维不等长轨迹数据时效果并不理想。提出一个新的轨迹聚类方法, 该方法通过组合动态时间弯曲和密度峰算法实现。动态时间弯曲用于度量轨迹间的距离, 密度峰算法根据距离进行聚类。前者可直接度量不等长轨迹聚类, 后者是近年提出的非球体分布数据聚类算法, 以局部密度和最近邻聚类组合实现。实验在 PETS2006 监控视频数据集上进行, 测试结果表明该方法有效地发现了异常的轨迹行为模式。

关键词: 轨迹聚类; 异常检测; 密度峰算法; 公共安全

文献标志码: A **中图分类号:** TP393 **doi:** 10.3778/j.issn.1002-8331.1601-0050

1 引言

近年来, 恐怖主义等威胁社会公共安全的活动更加频繁, 产生了重大的人员伤亡、财产损失, 因此, 对威胁社会公共安全的活动进行提前检测预警成为当前研究的主要热点和重点之一。目前, 此种检测预警可通过对公共区域人员的行为活动进行分析, 以异常行为活动的检测为主要内容, 特别是基于行人在公共场所的行为轨迹进行判断。随着计算机视觉技术的发展, 基于监控视频的行人检测、跟踪等技术已经逐步走向实践, 这些技术进一步推动了基于行人轨迹的异常行为检测, 及预警告警技术的发展。

目前, 基于轨迹的异常检测主要采用非监督的聚类技术。聚类是根据数据点的相似程度将数据点归为两个或多个类别的算法, 其中正常和异常行为由于数据的差异而聚类到不同的类别, 以此实现对异常行为的检测。聚类技术的发展经历了由简单到复杂的过程, 并在发展中不断完善和进步。最早出现的 K -均值聚类算法 (K -means)^[1] 是最经典的聚类算法, 其思想简单且效率较高, 计算复杂度仅为 $O(n)$, 其中, n 是样本数。但是该算法对“噪声”和孤立点数据敏感, 聚类中心会因此而出出现偏离。随后出现的 K 中心点算法 (K -medoids)^[2] 较好地解决 K -均值算法中的“噪声”敏感问题, 不过由于其采

基金项目: 江苏省自然科学基金(No.BK20140065); 江苏省工程技术研究中心(No.BM2014391); 国家自然科学基金(No.61174198); 国家自然科学基金联合基金(No.U1435218)。

作者简介: 康凯(1986—), 男, 硕士, 研究方向为社会应急安全, E-mail: 36442181@qq.com; 刘方鑫(1992—), 男, 硕士, 研究方向为水下传感网、城市计算。

收稿日期: 2016-01-06 **修回日期:** 2016-03-09 **文章编号:** 1002-8331(2016)14-0007-05

用新的质点计算规则,也使得算法的时间复杂度上升为 $O(n^2)$ 。以上两个聚类算法都是直接根据数据点到各个类中心的距离远近来进行归类,这种归类方式对于高维非球体分布的数据点来说就无法有效了。因此,基于分布的聚类算法开始出现,它通过多个分布混合来解决这种情况,但是其聚类的精度取决于所使用的分布及其混合的方式和数目。为了克服具体分布的影响,基于密度的空间聚类算法(DBSCAN)^[3]取得了较好的效果,且可以处理噪声数据,但是其需要选择一个合适的阈值,而阈值的选择无规律可循,只能靠经验确定。同样,基于密度的均值偏移聚类算法(Mean-Shift)^[4-5]虽没有阈值问题,但它的计算耗费比较大。2014年,《科学》杂志上发表了一篇新的基于密度的聚类算法——密度峰聚类算法(Density-Peak)^[6],使得非球体分布数据的聚类得到了新的突破。本文将该密度峰聚类算法应用到轨迹(时间序列)数据的聚类上,同样取得了很好的效果。

2 密度峰聚类

密度峰聚类算法可视为是过去多个算法思想的融合,与K-means和K-medoids相似,只使用了数据点之间的距离,与DBSCAN和Mean-Shift相似,基于密度极值聚类非球体分布。该算法基于两个基本假设:(1)类中心点比其周围点的局部密度高;(2)类中心点距离局部密度比其高的点具有较大的距离。

2.1 基本假设

基于上述假设,对于数据点 i ,定义局部密度 ρ_i 和最近邻距离 δ_i 分别为。

(1)局部密度 ρ_i :

$$\rho_i = \sum_j \delta(d_{ij} - d_c)$$

$$\delta(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases}$$

其中, d_{ij} 表示数据点 i 与数据点 j 之间的距离; d_c 表示截断距离参数; $\delta(x)$ 表示指示函数。 x 就是函数的定义域,其横坐标的值,当 $x < 0$, $\delta(x) = 1$;否则, $\delta(x) = 0$ 。简言之, ρ_i 就是到数据点 i 的距离小于 d_c 的数据点的数目。

(2)最近邻距离 δ_i :

$$\delta_i = \min_{j: \rho_j > \rho_i} (d_{ij})$$

理解上, δ_i 就是数据点 i 到局部密度比其大的数据点的最小距离。根据该定义,则密度最大的数据点无定义,故重定义为:

$$\delta_i = \max_j (d_{ij})$$

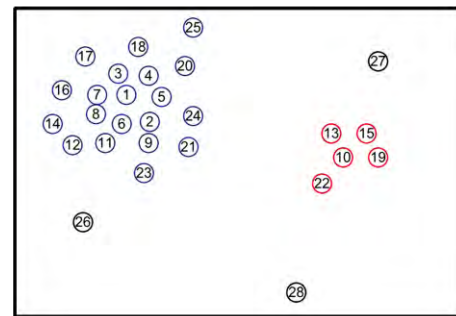
即数据点 i 到其他数据点的最大距离。综合而言,最近邻距离

$$\delta_i = \begin{cases} \max_j (d_{ij}), & \forall j, \rho_j > \rho_i \\ \min_{j: \rho_j > \rho_i} (d_{ij}), & \text{其他} \end{cases}$$

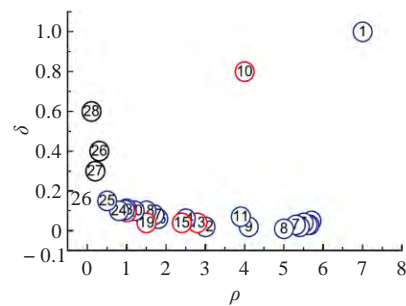
最后,聚类中心要求局部密度 ρ_i 大,且最近邻距离 δ_i 也大。

2.2 聚类过程

图1(a)是一幅数据点的二维空间分布,共有28个数据点,按局部密度由高到低编号。计算每个数据点的局部密度 ρ_i 和最近邻距离 δ_i ,可得到一张关于 $\rho \times \delta$ 的决策图,如图1(b)。在决策图上,具有较大 ρ_i 和较大 δ_i 的数据点即为类中心,如数据点1和10;同时,局部密度 ρ_i 较小但是最近邻距离 δ_i 较大的点可视为异常点或孤立点,如数据点26,27和28。



(a) 数据点分布图



(b) $\rho \times \delta$ 决策图

图1 数据点分布及其 $\rho \times \delta$ 决策图

确定类中心点之后,需对剩余数据点进行类别归属。基本思想就是按剩余数据点的局部密度降序确定它们的类别,使得待归类数据点的类别等同于局部密度比其高且距离其最近的数据点的类别,即与其具有最近邻距离的数据点的类别。归类过程如下:

(1)计算数据点 i 的“邻居”,即与其具有最近邻距离的点:

$$neighbour(i) = \arg \min_{j: \rho_j > \rho_i} (d_{ij})$$

(2)将剩余数据点按局部密度从大到小排序。

(3)对每个数据点,将其类别设置为其“邻居”的类别。

由于 $\rho_i < \rho_{neighbour(i)}$ 保证了当前点的“邻居”的类别必定已经完成了归类,故一次遍历即可完成类别归属过程。图1中,数据点1和10首先判定为类中心,其余点按密度从高到低依次进行类别归属。首先是点2,其“邻居”是点1,故将点1的所属类别赋予点2;其次是点3,其“邻居”还是点1,故将点1的所属类别赋予点3;接着是

点4,其“邻居”仍是点1,再次将点1的所属类别赋予点4;然后是点5,其“邻居”是点4,故将点4的所属类别赋予点5号,……,直至完成所有点的类别归属。

该算法对非球体分布数据点和参数 d_c 具有好的鲁棒性。实际中, d_c 的选择一般要求使得平均每个点的邻居数为所有点的1%~2%。

3 基于DTW的密度峰聚类轨迹

3.1 轨迹特性分析

轨迹 G 作为刻画对象随时间运动的动态序列,可以用向量序列表示:

$$G_T = \{g_1, g_2, \dots, g_t, \dots, g_N\}, g_t = [x^t, y^t]^T$$

其中, g 表示对象中心(重心)位置, t 表示时序标号。由于轨迹是一个和时间有关的向量序列,它的向量长度随时间 t 不同而发生变化。要处理这种不等长向量间的比较问题,常常采用一些预处理的方法将序列特征变为同等长度,如归一化方法、降维法。归一化方法是让所有的轨迹都具有相同的长度,主要方法有补零法^[7]和轨迹延长法^[8]。降维法是通过将不同长度的轨迹减少维数,使其变成同等长度后再进行处理。新的维数空间用来假设轨迹模型,同时通过选择参数来描述这个模型,参见文献[9-10]。

虽然通过轨迹预处理方法,将轨迹变为等长的向量进行聚类,计算比较简单。但是轨迹的很多信息在预处理中遭到破坏,很难保留轨迹中的原有信息。而且轨迹自身由于其自身噪声与波动性的特点,常规距离不能很好地适应相似度度量,相似的轨迹之间会呈现多种变形,如振幅平移和伸缩、线性漂移、不连续、时间轴伸缩等,传统的欧氏距离、马氏距离就变得不再合适。

3.2 DTW距离度量

从轨迹特性分析可知,传统距离函数无法有效度量不等长的轨迹。本文采用动态时间弯曲(Dynamic Time Warping, DTW)距离进行度量。DTW距离是一种基于时域建模的时间序列匹配方法,最早用于语音识别中计算不同长度的信号距离^[11],在匹配简单的轨迹上应用比较广泛,例如手写体识别。它可以同时得到两个距离,一个是时间上的扭曲程度,另一个是空间上的轨迹相似度。

与传统距离不同,DTW距离不要求两条时间序列点与点之间进行一一对应的匹配,允许序列点自我复制后再进行对齐匹配^[12]。当时间序列发生时间轴弯曲时,可以在弯曲部分进行自我复制,使两条时间序列之间的相似波形进行对齐匹配。图2(a)中,两个序列都是均匀采样的,然后比较对应点的距离,显然在均匀采样时两个序列的距离比较大。图2(b)中,上边的序列为均匀采样,下边的序列采样不均匀,这个不均匀的时间采样函数就是时间扭曲函数。

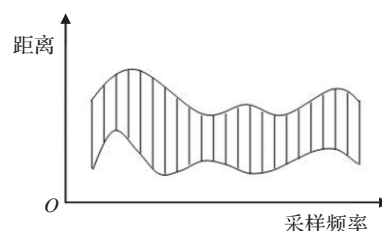


图2(a) 均匀采样度量

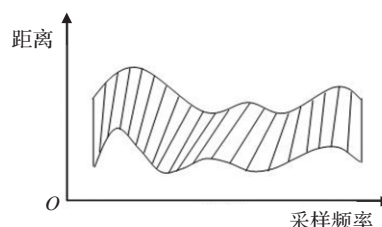


图2(b) DTW度量

DTW距离定义:

设时间序 $X = \langle x_1, x_2, \dots, x_n \rangle$, $Y = \langle y_1, y_2, \dots, y_m \rangle$, 其中长度分别为 n 和 m 。它们之间的DTW距离定义为:

$$D_{\text{dtw}}(\Phi, \Phi) = 0$$

$$D_{\text{dtw}}(X, \Phi) = D_{\text{dtw}}(Y, \Phi) = \infty$$

$$D_{\text{dtw}}(X, Y) = d(x_1, y_1) + \min \begin{cases} D_{\text{dtw}}(X, Y[2:m]) \\ D_{\text{dtw}}(X[2:n], Y) \\ D_{\text{dtw}}(X[2:n], Y[2:m]) \end{cases}$$

其中 Φ 表示空的轨迹序列, $d(x_i, y_i)$ 表示序列点 x_i 和 y_i 之间的距离,可以由用户选择不同的距离度量,比如马氏距离。

通常,两条轨迹序列之间的DTW距离计算采用构造累计距离矩阵的动态规划方法,累计距离矩阵实际上是一个递推关系,可以通过回溯法求解^[13-14],DTW的时间复杂度为 $O(nm)$ 。

基于DTW距离的密度峰聚类算法:

根据DTW距离定义计算两两数据点之间的距离矩阵 $\{d_{ij}\}_{i,j=1}^n$ 。

根据距离矩阵,计算所有数据点的局部密度 $\{\rho_i\}_{i=1}^n$ 。

根据局部密度和距离矩阵,计算所有数据点的最近邻距离 $\{\delta_i\}_{i=1}^n$,并记录具有最近邻距离的数据点(即“邻居”)的索引。

根据局部密度和最近邻距离,绘制 $\rho \times \delta$ 的决策图,再根据给定的类数依次确定左上角孤立数据点为类中心,并依次标号。

根据标号的类中心,执行2.2节描述的归类过程,即完成了整个聚类过程。

4 实验和应用

实验选取PETS2006数据集进行轨迹聚类分析,实现常规运动模式发现和异常模式检测。PETS2006数据集是火车站场景的监控视频,如图3所示,主要包括4个

摄像头拍摄的不同角度下的同一场景视频。该数据主要用于检测多摄像头发现“丢失行李”事件。实验选取3号摄像机拍摄的7段视频(同一场景)构建轨迹数据,该部分数据共持续11 min 9 s,包括17 394帧画面。结合视频内容分析可知,视频中的轨迹行为主要分为4类(见图4):类1表示站台处上车,类2表示车站门口进入,类3表示车站门口离开,类4表示进出停留区域。其中,

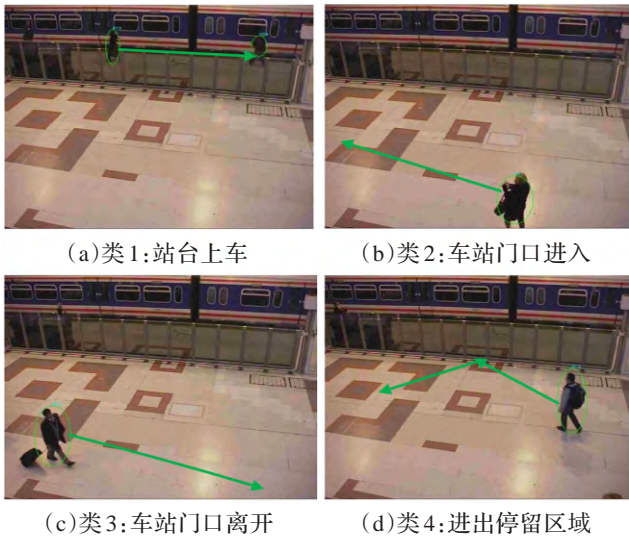
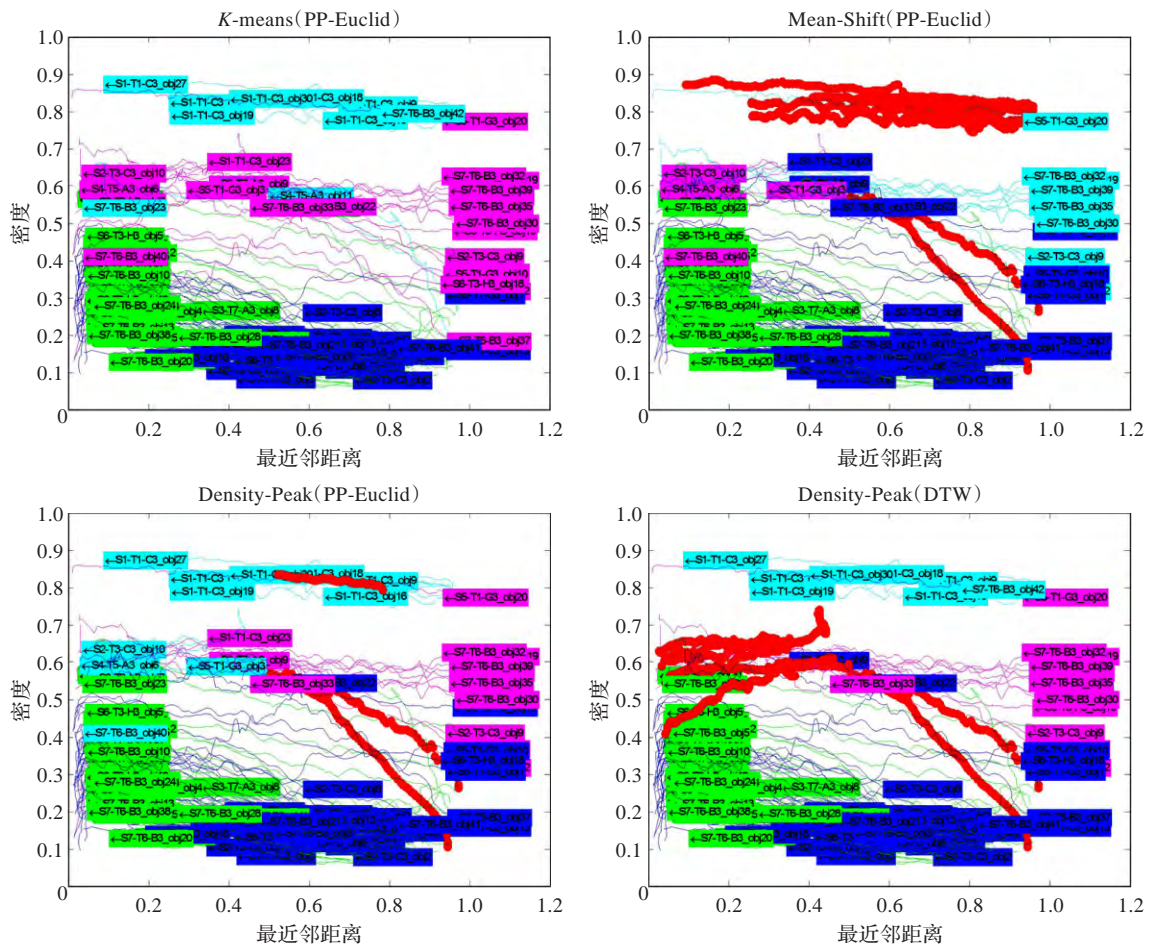


图3 PETS2006中事件类别

“丢失行李”事件产生的行为属于类4,其轨迹条数较少。实验采用OpenCV视觉库(www.opencv.org)中的高斯背景模型检测运动目标,采用均值滤波算法跟踪运动目标,由跟踪结果自动生成数据作为聚类的轨迹数据。经运动目标检测和跟踪后,共检测出170条轨迹,其中有效跟踪的139条(本文设定有效跟踪的对象持续帧数大于50帧)。轨迹点坐标根据场景像素的高度和宽度分别量化到区间[0,1]。

为了检验基于DTW距离的Density-Peak聚类轨迹的效果。实验分别采用轨迹线性插值补齐后的欧式距离(PP-Euclid距离)和轨迹直接DTW距离(DTW距离)来度量不等长轨迹的相似度,在PP-Euclid距离下分别采用K-means、Mean-Shift和Density-Peak聚类轨迹,在DTW距离下采用Density-Peak聚类轨迹(K-means和Mean-Shift无法计算不等长轨迹的类中心点)。其中,K-means和Mean-Shift调用Matlab自带函数,设置K-means的聚类个数为4,设置Mean-Shift的聚类半径也为4,Density-Peak聚类个数同样为4。实验结果见图4,其中,红色粗线条表示在4个主要类别之外的轨迹(异常轨迹),蓝色、绿色和粉红色为发现的轨迹运动模式。由图4可知,在PP-Euclid距离条件下,Density-Peak的聚类结果比K-means和Mean-Shift的聚类结果都好;在



Density-Peak聚类算法条件下,采用DTW距离比PP-Euclid距离的聚类结果更好。因此,基于DTW距离的Density-Peak聚类算法是一种相对K-means和Mean-Shift更好的轨迹运动模式发现和异常检测的聚类算法,且因其聚类中心为样本数据点(而非样本均值点),故而更适合聚类不等长轨迹数据。

5 结束语

本文基于密度峰聚类算法研究了基于运动目标轨迹的模式发现和异常检测,提出了采用DTW距离解决高维不等长轨迹数据点的距离度量的问题,并结合轨迹数据点的局部密度和最近邻距离,通过密度峰聚类完成了异常轨迹行为检测。该方法在监控视频中可以实现行人常规行为和异常行为的测试,实验效果显著,提供了公共安全异常行为的预警和告警基础。

参考文献:

- [1] MacQueen J. Multivariate observations[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, 1: 281-297.
- [2] Kaufman L, Rousseeuw P J. Finding groups in data: an introduction to cluster analysis[M]. New York: Wiley-Interscience, 2009.
- [3] Ester M, Krieger H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996: 226-231.
- [4] Fukunaga K, Hostetler L. The estimation of the gradient of a density function in pattern recognition[J]. IEEE Trans on Inf Theory, 1975, 21: 32-40.
- [5] Cheng Y. Mean Shift, model seeking, and clustering[J]. IEEE Trans on Pattern Anal Mach Intell, 1995, 17(8): 790-799.
- [6] Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.
- [7] Hu W, Xiao X, Fu Z, et al. A system for learning statistical motion patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(9): 1450-1464.
- [8] Hu W, Tan T, Wang L, et al. A survey of visual surveillance of object motion and behaviors[J]. IEEE Transactions on Systems, Man and Cybernetics, 2004, 34(3): 334-352.
- [9] Jiao L, Wu Y, Wu G, et al. Anatomy of a multicamera video surveillance system[J]. ACM Multimedia Systems, 2004, 210(2): 144-163.
- [10] Zhong H, Shi J, Visontai M. Detecting unusual activities in video[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, 2: 819-826.
- [11] Rabiner L, Juang B. Fundamentals of speech recognition[M]. Englewood Cliffs, NJ: Prentice-Hall, 1993.
- [12] 何涛. 视频中的稀疏多目标跟踪和轨迹异常检测研究[D]. 成都: 西南交通大学, 2014.
- [13] 孙玉砚, 孙利民, 朱红松, 等. 基于车牌识别系统车辆轨迹的行为异常检测[J]. 计算机研究与发展, 2015, 52(8): 1921-1929.
- [14] 田国会, 尹建芹, 闫云章, 等. 基于混合高斯模型和主成分分析的轨迹分析行为识别方法[J]. 电子学报, 2016(1).
- [15] Bernstein L, Bosch P, Canziani O, et al. Climate change 2007: synthesis report: an assessment of the intergovernmental panel on climate change[Z]. Geneva, Switzerland: IPCC, 2007.
- [16] Li L, Lin P, Yu Y, et al. The flexible global ocean-atmosphere-land system model, grid-point version 2: FGOALS-g2[J]. Advances in Atmospheric Sciences, 2013, 30(3): 543-560.
- [17] Dai Y, Zeng X, Dickinson R E, et al. The common land model[J]. Bull Amer Meteor Soc, 2003, 84(8): 1013-1023.
- [18] Intel® VTune™ Amplifier 2016[EB/OL]. [2015-11-10]. <https://software.intel.com/en-us/intel-vtune-amplifier-xe>.
- [19] 张理论, 叶红, 吴建平, 等. 基于最大负载偏移率的并行负载平衡性能分析[J]. 计算机研究与发展, 2010, 47(6): 1125-1131.

(上接6页)

- [5] Dongarra J J, Piotr L, Antoine P. The LINPACK Benchmark: past, present and future[J]. Concurrency & Computation Practice & Experience, 2003, 15(9): 803-820.
- [6] Dongarra J, Luszczek P. LINPACK Benchmark[M]//Encyclopedia of parallel computing. US: Springer, 2011: 1033-1036.
- [7] TOP500.org. Top500 supercomputer sites[EB/OL]. [2015-11-10]. <http://www.top500.org/>.
- [8] Meuer H W. The TOP500 project: looking back over 15 years of supercomputing experience[J]. Informatik Spektrum, 2008, 31(3): 203-222.
- [9] Gates W L, Boyle J S, Covey C, et al. An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I)[J]. Bull Am Meteorol Soc, 1999, 80: 29-56.
- [10] Scaife A A, Kucharski F, Folland C K, et al. The CLIVAR C20C project: selected twentieth century climate events[J].

Climate Dynamics, 2009, 33(5): 603-614.

- [11] Bernstein L, Bosch P, Canziani O, et al. Climate change 2007: synthesis report: an assessment of the intergovernmental panel on climate change[Z]. Geneva, Switzerland: IPCC, 2007.
- [12] Li L, Lin P, Yu Y, et al. The flexible global ocean-atmosphere-land system model, grid-point version 2: FGOALS-g2[J]. Advances in Atmospheric Sciences, 2013, 30(3): 543-560.
- [13] Dai Y, Zeng X, Dickinson R E, et al. The common land model[J]. Bull Amer Meteor Soc, 2003, 84(8): 1013-1023.
- [14] Intel® VTune™ Amplifier 2016[EB/OL]. [2015-11-10]. <https://software.intel.com/en-us/intel-vtune-amplifier-xe>.
- [15] 张理论, 叶红, 吴建平, 等. 基于最大负载偏移率的并行负载平衡性能分析[J]. 计算机研究与发展, 2010, 47(6): 1125-1131.