基于特征密集计算与融合算法的教师课堂行为分析

张晓龙[1]，刘剑飞[1*]，郝禄国[2]

[1]河北工业大学电子信息工程学院，天津 300401；

[2]广东工业大学信息工程学院，广东 广州 510006

摘要 针对传统网络结构不能充分利用数据中时空信息的问题，提出了一种时空金字塔池化模型，并将该模型与非局部特征计算操作相结合，设计了一种基于时空信息密集计算与融合的三维密集连接卷积神经网络，从而可以更有效地提取视频的时空特征。将该算法应用于课堂场景下教师行为的分析，实验结果表明，所设计的网络结构在教师行为数据集上达到了较高的识别准确率。

关键词 机器视觉；卷积神经网络；时空金字塔池化；非局部计算；时空特征；行为分析

中图分类号 TP391.4 **文献标识码** A **doi**：10.3788/LOP56.161503

Analysis of Teachers′ Actions Using Feature Dense Computation and Fusion Algorithm

Zhang Xiaolong[1], Liu Jianfei[1*], Hao Luguo[2]

[1] *School of Electronic and Information Engineering, Hebei University of Technology, Tianjin* 300401, *China*;

[2] *School of Information Engineering, Guangdong University of Technology, Guangzhou, Guangdong* 510006, *China*

Abstract Considering the incapacity of the traditional network structure to fully extract spatiotemporal information in data, a spatiotemporal pyramid pooling model is proposed. A three-dimensional, densely connected convolutional neural network based on dense computation and fusion of spatiotemporal features is designed. The combination of this model with non-local computation operation of features improves the effectiveness of spatiotemporal-feature extraction from videos. The algorithm is applied to the analysis of teachers′ actions in classroom scenes. The experimental results show that the designed network structure produces high recognition accuracy on the teachers′ actions dataset.

Key words machine vision; convolution neural network; spatiotemporal pyramid pooling; non-local computation; spatiotemporal features; analysis of actions

OCIS codes 150.1135; 100.2000; 100.4996; 100.6890

1 引 言

当前，视频监控已经逐渐成为了辅助教育教学的一种手段。从视频中准确识别出教师的教学行为方式，有助于教学方式的统计与评价，改善教师教学质量。同时，将得到的数据和结论推送给教师本人或学校，有助于多方位了解教学状况，从而对教师在课堂教学方式上出现的问题制定针对性的解决方案。对课堂场景下教师的行为进行分析，旨在提高教师行为识别的准确率，充分利用视频资源辅助分析教师整体行为，从而自动生成教学方式评估。

在针对目标行为识别方法的研究中，文献[1]中提出一种基于时空方向主成分直方图的人体行为识别方法，解决了由视角、尺度等变化造成的行为类内差别大的问题。文献[2]中基于一种视觉词袋表示方法和融合模型表示人体行为特征，获得了较好的识别性能。然而，上述方法在学习数据特征信息方面受到限制，无法针对识别任务自动生成适应

收稿日期：2019-02-25；**修回日期**：2019-03-15；**录用日期**：2019-03-27

基金项目：天津市自然科学基金(15JCYBJC17000)、河北省高等学校科学技术研究重点项目(ZD2017021)

* **E-mail**：jfliu@hebut.edu.cn

模型[3]。

近年来，深度卷积神经网络在行为识别、目标检测等视觉任务上展现出了较大的优势[4]。基于卷积神经网络的方法具有强大的特征表达能力，但该方法存在对样本监督信息利用不充分的问题[5]，文献[6]和文献[7]中使用了多种特征信息对目标进行组合识别，提高了分类的正确率。

大多数行为识别网络模型采用的是视频处理的方法，从视频中提取时间和空间信息，然后组合预测视频中的行为类别[8]。3D DenseNets[9]、C3D[10]和3D ResNets[11]都是基于3D卷积神经网络设计的模型结构。其中，3D DenseNets为密集连接型卷积结构，因其能够重复利用各层特征，行为识别准确率得到了很大提升。然而，3D DenseNets虽然采用了特征复用的方式提高特征利用率，但该网络在进行综合预测时未考虑时空特征点之间的关联性[12]和融合多尺度特征[13]，不能充分利用视频特征信息。此外，传统算法要求输入的三维视频数据的时长和空间分辨率均固定，因此，不能使用不同时间和空间尺寸的训练样本进行训练。为此，本文提出一种基于时空信息密集计算与融合的深度卷积神经网络。该网络在加深三维密集连接型结构的同时，引入了时空金字塔池化模块和时空特征非局部计算模块，将时空信息密集计算和时空信息密集融合技术相结合，改善特征空间的行为表征能力，从而提高视频数据行为识别的准确率。

2 基于时空信息密集计算与融合的深度卷积网络

设计的基于时空信息密集计算与融合的深度卷积神经网络(TSDCFN)是在改进型三维密集连接卷积神经网络的基础上引入非局部特征密集计算算法提取时空信息，然后使用一种三维金字塔池化模块得到时空信息的密集融合。

TSDCFN整体结构如图1(a)所示，其中密集连接模块(dense connection layer)的结构如图1(b)所示。通过裁切、水平翻转方法产生的不同时长、不同分辨率的数据首先经过3×3×3卷积层，然后进入多个密集连接块(dense block)、由1×1×1卷积与1×2×2最大池化构成的转换层(transition layer)和只包含1×1×1卷积的转换层(transition w/o pooling layer)，再进行人体行为特征的提取，其中每个dense block均由BN-ReLU-1×1×1Conv以及BN-ReLU- NonLocal-3×3×3Conv混合计算层组成。接着通过时空金字塔池化层(spatio-temporal pyramid pooling layer)产生多种尺度的特征图。最后融合多尺度特征图，在分类层(classification layer)预测输出行为类别。

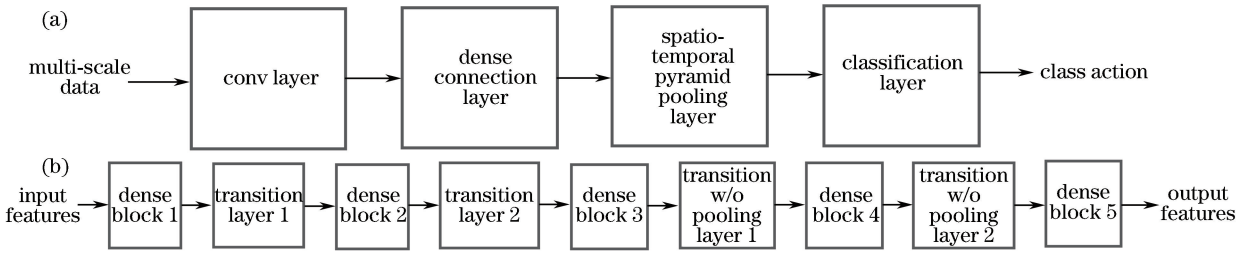

图1 TSDCFN结构框图。(a) TSDCFN整体结构；(b)密集连接模块结构(其中，w/o表示不包含)

Fig. 1 Schematic of TSDCFN structure. (a) Overall structure of TSDCFN; (b) structure of densely connected module (in which, w/o represents without)

在TSDCFN中，非局部特征计算模块使特征图的每个特征点都是由邻近的其他特征点计算得出，同时时空金字塔池化模块将特征图池化成多种尺度，从而得到更有效的行为类别特征表示。

2.1 时空金字塔池化模块

针对输入视频数据尺寸不一致的问题，提出了一种时空金字塔池化模型，如图2所示。在全连接层之前通过参数集bins设置池化的层数以及每层特征图的大小，得到相应维数的多尺度特征图。加入时空金字塔池化块之后，网络对于不同尺寸、不同时长的视频动作片段具有更强的泛化能力。同时，多尺度池化提高了特征对行为类别的表示能力。因此，时空金字塔池化模型可以实现多尺寸视频的输入以及多尺度时空特征的密集融合。时空金字塔模型的参数集bins为$\{k,l,m,n\}$，表示池化层共有4层，第1层池化得到的特征图的维度是$k\times k\times k$，第2层池化得到的特征图的维度是$l\times l\times l$，第3层和第4层以此类推。池化过程中沿着空间维度方向和时间维度方向的核尺寸S和步长T分别为

$$S=\mathrm{cell}(s/i),\tag{1}$$

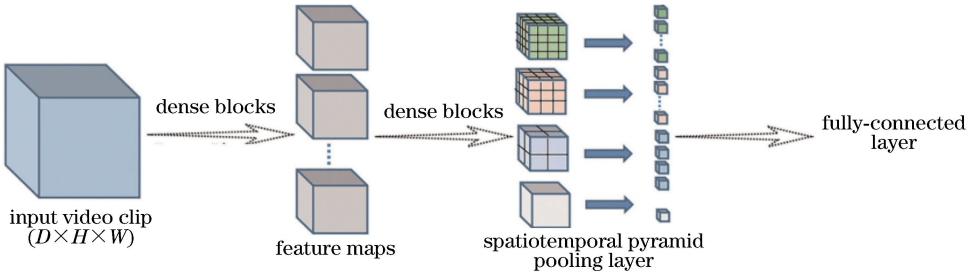

图 2 时空金字塔池化模型

Fig. 2 Spatiotemporal pyramid pooling model

$$T=\mathrm{floor}(s/i), \tag{2}$$

式中：cell 为向上取整函数；floor 为向下取整函数；i 为 bins 中设置的参数($i=k,l,m,n$)；s 为金字塔池化模型的输入尺寸。池化运算时采取补零的方式。

最后通过 Flatten 函数和 Concatenation 函数将特征图展开并连接成一维特征向量，送入全连接层。

2.2 改进型三维密集连接网络

改进型三维密集连接卷积神经网络由深度监督的 DenseNets 网络经过膨胀化卷积和池化得到。为了扩展更多的 dense blocks 以加深网络，同时防止特征图的分辨率被过分降低，引入 DSOD 算法[14]中的 transition w/o pooling layer，使网络不只是通过增加每个 block 内部的层数来加深网络。改进型三维密集连接卷积神经网络包含 1 个 3D 卷积层、5 个 dense blocks、2 个 transition layers、3 个 transition w/o pooling layers 和 1 个 classification layer。表 1 给出了网络深度 $d=58$、生长率 $k=48$ 时的改进型三维密集连接卷积神经网络的模型结构，其中，D、H 和 W 分别表示数据的序列长度、帧高度和帧宽度。

2.3 非局部特征计算模块

三维视频数据中存在时间维度的信息，特征点不仅与同一帧空间内部的其他特征点有关，还与相邻帧的特征点有关。非局部(non-local)特征计算模块[12]可将某点处的特征与相邻点处的特征联系起来，充分挖掘空间帧内部以及时间帧之间的关系。

图 3 为一种时空非局部特征计算模块框图。若输入特征图维度为 $T\times H\times W\times 1024$，分别经过 3 次 $1\times1\times1$ 卷积运算，维度均变为 $T\times H\times W\times 512$，经过两个乘法器维度不变，再经过一次 $1\times1\times1$ 卷积运算，特征图维度增加到 $T\times H\times W\times 1024$，经过加法器输出的特征图的维度与输入特征图的维度相等。在所提出的网络中，将该模块添加到每个 dense block 最后一层 $1\times1\times1$ 卷积的前面，相应的

表 1 改进型三维密集连接卷积神经网络参数

Table 1 Parameters of modified 3D densely connected convolutional neural network

Layer	Output size	3D DenseNet
3D Convolution	$D\times H\times W$	$3\times3\times3$ conv
Dense block 1	$D\times H\times W$	$\begin{bmatrix}1\times1\times1\text{ conv}\\3\times3\times3\text{ conv}\end{bmatrix}\times10$
Transition layer 1	$D\times H\times W$	$1\times1\times1$ conv
	$D\times\frac{H}{2}\times\frac{W}{2}$	$1\times2\times2$ max pooling
Dense block 2	$D\times\frac{H}{2}\times\frac{W}{2}$	$\begin{bmatrix}1\times1\times1\text{ conv}\\3\times3\times3\text{ conv}\end{bmatrix}\times10$
Transition layer 2	$D\times\frac{H}{2}\times\frac{W}{2}$	$1\times1\times1$ conv
	$\frac{D}{2}\times\frac{H}{4}\times\frac{W}{4}$	$2\times2\times2$ max pooling
Dense block 3	$\frac{D}{2}\times\frac{H}{4}\times\frac{W}{4}$	$\begin{bmatrix}1\times1\times1\text{ conv}\\3\times3\times3\text{ conv}\end{bmatrix}\times10$
Transition w/o pooling layer 1	$\frac{D}{2}\times\frac{H}{4}\times\frac{W}{4}$	$1\times1\times1$ conv
Dense block 4	$\frac{D}{2}\times\frac{H}{4}\times\frac{W}{4}$	$\begin{bmatrix}1\times1\times1\text{ conv}\\3\times3\times3\text{ conv}\end{bmatrix}\times10$
Transition w/o pooling layer 2	$\frac{D}{2}\times\frac{H}{4}\times\frac{W}{4}$	$1\times1\times1$ conv
Dense block 5	$\frac{D}{2}\times\frac{H}{4}\times\frac{W}{4}$	$\begin{bmatrix}1\times1\times1\text{ conv}\\3\times3\times3\text{ conv}\end{bmatrix}\times10$
Transition w/o pooling layer 3	$\frac{D}{2}\times\frac{H}{4}\times\frac{W}{4}$	$1\times1\times1$ conv
Classification layer		fully connected softmax and prediction

计算公式为

$$\boldsymbol{y}=\mathrm{softmax}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}_{\theta}^{\mathrm{T}}\boldsymbol{W}_{\varphi}\boldsymbol{x})\boldsymbol{W}_{g}\boldsymbol{x}, \tag{3}$$

式中：softmax(·)为归一化指数函数；θ、φ、g 均为 $1\times1\times1$ 卷积操作；$\boldsymbol{W}_{\theta}$、$\boldsymbol{W}_{\varphi}$、$\boldsymbol{W}_{g}$ 为对应的权重学习矩阵；$\boldsymbol{x}$ 为输入特征图；$\boldsymbol{y}$ 为输出特征图。

3 实验与分析

利用课堂场景下的教师行为数据库对提出的

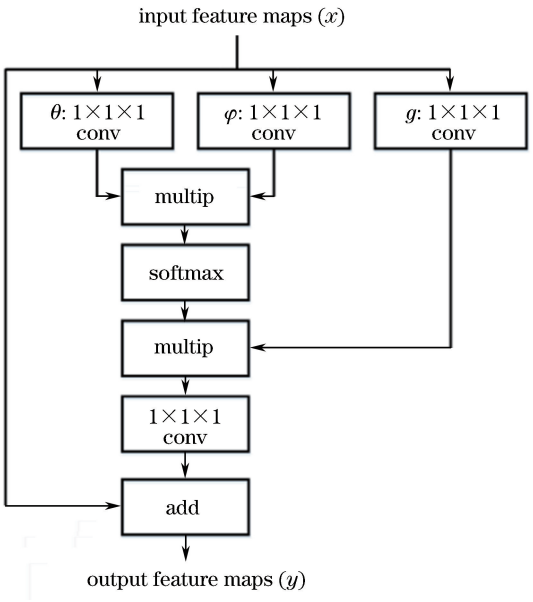

图 3 时空非局部特征计算操作框图

Fig. 3 Scheme of spatiotemporal and non-local feature computation

TSDCFN 进行训练和测试。下面首先介绍实验使用的教师行为数据库，给出网络训练相关的设置，然后通过实验验证不同的网络参数配置、非局部特征计算模块以及三维金字塔池化模块对教师行为识别准确率的影响。

3.1 教师行为数据库的建立

课堂场景的教师行为数据库共包含 1800 个动作短视频，设计了 9 种不同的行为类别，分别为黑板板书（blackboard writing）、纸上写字（paper writing）、长时注视学生（long time staring）、坐下（sitting）、使用手机（phone using）、讲授投影仪（projector teaching）、讲话（speaking）、操作计算机（computer operating）和翻阅资料（material browsing），各类样本的数量分别为 169，225，250，130，182，230，255，139，220。为了增强数据库的丰富性，提高网络模型的泛化性，制作样本时考虑了光线、背景、分辨率和尺度等变化因素。

3.2 训练设置

训练时，超参数 Nesterov momentum 和 weight decay 分别设置为 0.9 和 10^{-4}，根据文献[15]中的方法进行参数初始化，使用 Xavier 对全连接层的权重进行初始化[16]。当序列长度为 16 时，使用批大小 8；当序列长度为 32 时，使用批大小 4，这样可以保证每次学习时图形处理器（GPU）处理的帧数据量均为 128。迭代训练次数为 200。学习率最初设置为 0.1，在迭代次数 Epoch 为 90 和 150 处分别减小为原来的 1/10 和 1/100。dense block 中 1×1×1 卷积层的输出通道数均设置为 128，时空金字塔池化模块的参数集 bins 设置为[4，3，2，1]。

3.3 实验结果与分析

不同参数配置的实验结果见表 2。其中，d 为网络深度；k 为 dense blocks 中每层产生的通道数；θ 为转换层输出通道数与输入通道数的比值；L 为网络每次处理的数据长度。表中最后一行采用了数据增强的方法。全连接层过渡矩阵要求网络的输入数据尺寸一致，所以数据增强实验是在引入时空金字塔池化模块的基础上进行的。

图 4 为实验训练过程中识别准确率的对比曲线，给出了本文设计的 TSDCFN 网络和传统 3D DenseNet 网络的比较。由图 4 可知，网络在迭代到 60 个 Epoch 时 TSDCFN 的识别准确率已接近 90%，而 3D DenseNet 的识别准确率在 80%以下；同时，TSDCFN 在 180 个 Epoch 时识别准确率稳定在 99%左右，收敛速度较快。

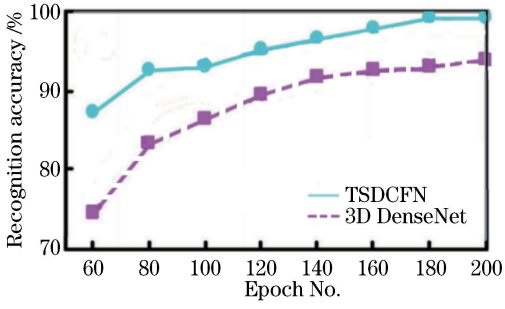

图 4 识别率与迭代次数的关系

Fig. 4 Relationship between recognition accuracy and Epoch No.

由表 2 可知，传统 3D DenseNet 算法的识别准确率仅为 85.62%，本文通过扩展密集块并优化参数，改进的网络 Modified 3D DenseNet 识别准确率可达 90%以上；引入非局部特征密集计算模块时的准确率为 93.02%；加入时空金字塔池化模块时的准确率为 93.76%；若同时引入两个模块，即构成 TSDCFN 结构，行为识别准确率达 96.87%。由于在网络设计中引入了时空金字塔池化模块，网络能够对多时长、多空间尺寸的输入样本进行训练，这种采用数据增强策略之后的网络识别准确率达 98.13%。由此可知，本文设计的基于特征密集计算与融合算法的三维密集连接卷积模型在教师行为数据集上具有较好的识别性能。

表 2 不同参数配置的网络在教师行为数据集上的测试结果(其中,Data Aug 表示数据增强)

Table 2 Test results of networks with different configuration parameters on dataset of teachers' actions (in which, Data Aug represents Data Augmentation)

Method	d	k	θ	L	Data Aug	Accuracy /%
3D DenseNet	30	24	0.5	32	—	85.62
Modified 3D DenseNet	58	48	1.0	16	—	91.44
Modified 3D DenseNet with non-local feature computation block	58	48	1.0	32	—	93.02
Modified 3D DenseNet with spatio-temporally pyramid pooling	58	48	1.0	—	—	93.76
TSDCFN with non-local feature computation	58	48	1.0	—	—	96.87
block and spatio temporal pyramid pooling	58	48	1.0	—	√	98.13

本文设计的网络对教师行为的识别示例如图 5 所示。由图 5 可见,该网络对尺度变化、区分度低的教师行为均可获得准确的识别效果,有效验证了其准确性及稳健性。

图 5 动作视频识别效果示例

Fig. 5 Examples of recognition effects of action videos

4 结 论

提出了一种基于时空信息密集计算与融合的深度卷积神经网络,该网络通过引入非局部特征计算模块和三维金字塔池化模块,使视频中的时空信息得到了充分利用,从而大大提高了对行为特征的表示能力。将该网络应用于课堂场景下教师行为数据集,实验结果表明,时空特征密集计算和多尺度特征密集融合的互补特性可以有效提升教师行为识别的准确率。

致谢 感谢广州海昇计算机科技有限公司对本项目的支持

参 考 文 献

[1] Xu H Y, Kong J, Jiang M, *et al*. Action recognition based on histogram of spatio-temporal oriented principal components[J]. Laser & Optoelectronics Progress, 2018, 55(6): 061009.
徐海洋, 孔军, 蒋敏, 等. 基于时空方向主成分直方图的人体行为识别[J]. 激光与光电子学进展, 2018, 55(6): 061009.

[2] Peng X J, Wang L M, Wang X X, *et al*. Bag of visual words and fusion methods for action recognition: comprehensive study and good practice [J]. Computer Vision and Image Understanding, 2016, 150: 109-125.

[3] Hinton G E. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[4] Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[5] Chen B, Zha Y F, Li Y Q, *et al*. Person re-identification based on convolutional neural network discriminative feature learning [J]. Acta Optica Sinica, 2018, 38(7): 0720001.
陈兵, 查宇飞, 李运强, 等. 基于卷积神经网络判别特征学习的行人重识别[J]. 光学学报, 2018, 38(7): 0720001.

[6] Liu F, Shen T S, Ma X X. Convolutional neural network based multi-band ship target recognition with feature fusion[J]. Acta Optica Sinica, 2017, 37(10): 1015002.
刘峰, 沈同圣, 马新星. 特征融合的卷积神经网络多波段舰船目标识别[J]. 光学学报, 2017, 37(10): 1015002.

[7] Li C J, Liu Y P. Abnormal driving behavior detection based on covariance manifold and LogitBoost [J]. Laser & Optoelectronics Progress, 2018, 55(11): 111503.
李此君, 刘云鹏. 基于协方差流形和LogitBoost的异常驾驶行为识别方法[J]. 激光与光电子学进展, 2018, 55(11): 111503.

[8] Ng J Y H, Hausknecht M, Vijayanarasimhan S, *et al*. Beyond short snippets: deep networks for video classification [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 4694-4702.

[9] Gu D F. 3D densely connected convolutional network for the recognition of human shopping actions[D]. Ottawa, Ontario: University of Ottawa, 2017.

[10] Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks[C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, Santiago, Chile. New York: IEEE, 2016: 4489-4497.

[11] Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition [C] // 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), October 22-29, 2017, Venice, Italy. New York: IEEE, 2018: 3154-3160.

[12] Wang X L, Girshick R, Gupta A, *et al*. Non-local neural networks[C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 7794-7803.

[13] He K M, Zhang X Y, Ren S Q, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[14] Shen Z Q, Liu Z, Li J G, *et al*. DSOD: learning deeply supervised object detectors from scratch[C] // 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 1937-1945.

[15] He K M, Zhang X Y, Ren S Q, *et al*. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification [C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1026-1034.

[16] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [C] // Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, May 13-15, 2010, Chia Laguna Resort, Sardinia, Italy. Cambridge: PMLR, 2010, 9: 249-256.