指导教师：杨涛　　提交时间：2015/3/20

CVPR2015 Paper Translation

No: 01
姓名：闫士一
学号：2013302614
班号：10011306

多任务的课程学习

Anastasia Pentina, Viktoriia Sharmanska, Christoph H. Lampert
1ST Austria, Am Campus 1,3400 Klosterneuburg, Austria
{apentina,vsharman,chl}@ist.ac.at

摘要

在多个任务之间共享信息能够使算法达到很好的泛化性即使只是进行了小型的数据训练。但是，在一个多任务学习的真实场景中，不是所有的任务都是密切相关的，因此仅仅在最相关的任务中传递信息可能是最有利的。

在这项工作中，我们提出了一个方法，处理多个任务按照一定的顺序通过和后面的任务共享信息而不是一下子处理全部信息。后来，我们针对多任务的课程学习，即找到任务的最好顺序去学习，我们是基于泛化绑定标准的方法来选择任务顺序，这种方法可以优化平均分类期望。 我们的实验结果显示,按顺序学习多种相关任务可以共同比学习更有效，在这种顺序中，任务被解决影响所有的表现，我们的模型能够自动发现有利的秩序的任务。

1.简介

多任务的学习【6】研究解决一些预测任务的问题。虽然传统的机器学习算法可应用于独立解决每个任务，但他们通常需要显著数量的标记数据达到合理的泛化要求。然而，在许多情况下，它是昂贵的和耗时的，注释大量数据，尤其是在计算机视觉应用，如对象分类。另外一种可替代的方法是在几个相关的学习任务中共享信息，实验也表明，这可以用更少的训练点获得更好的泛化。

在这项工作中，我们专注于参数传递的方法，对于多任务学习这取决于相应的模型相关的任务相似性以及每一个方面参数表示。我们专注于线性预测的情况下，并假定模型之间的相似性是通过之间的欧几里得距离测量相应的权值向量[35]。在一个多任务设置中，这个想法是由 Evgeniou 和 PontiI 在[9]中介绍。有作者提出了一种基于支持向量机算法，强制执行对应于不同任务的权重向量为了接近一些正常的原型，他们展示了其在若干数据集的有效性。然而，该算法对称地对待所有的任务，这可能不是最优的在更现实的场景可能会有一些异常任务或任务组，使得来自不同组的任务之间没有相似性。因此，更灵活模型是必要的，能够利用任务关系结构，避免无关任务间的信息传递产生的消极后果。

通过与欧氏距离的正则化理念，不同任务的权重向量也常用在域适配情况下，如学习者有 2 个或多个预测任务，但对执行感兴趣只有在其中的一个。所有其他任务只会作为额外信息源。这种制定产生了有效的算法在各种各样计算机视觉应用方面：物体检测[2]，个性化图像搜索[17]，手假肢[25]和图象分类[33，34]。它的修改转移了用于物体检测功能[]1]和认可[13]间的关联模式。虽然自适应场景是于多任务有显着不同的，因为它集中精力解决的任务只有一个，而不是所有的人，这两个研究领域而言密切相关的方法，因此可受益于彼此。特别是，我们提出的学习算法可以看出作为一种分解一个多任务的问题为一组的领域适应性问题。

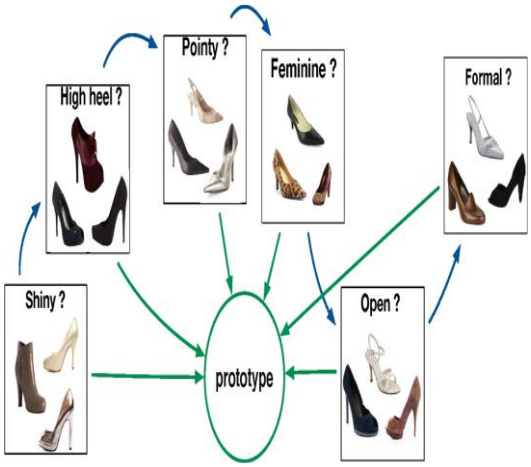

图 1 所提出的多任务学习的示意图方法。如果每一个任务都涉及到一些其他的任务，而不是对所有其他人一样，学习任务有序（蓝色箭头）有利于以一个单一原型共享信息（绿色箭头）为基础的经典多任务学习。

我们的方法从人类的教育过程得到启发。如果我们考虑到在校学生，他们，同样以多任务的学习者，都应该学习很多概念。然而，他们并不是同时学习它们而是按照一定的顺序。通过按照一定有意义的顺序处理任务，学生能够逐步增加他们的知识而且重用以前积累的信息，更有效学习新的概念。受这个例子的启发，我们建议通过信息传递以连续方式解决任务，从以前的任务到下一个，而不是同时解决所有的问题。从任务和内存效率的多变性来说，这种方法使学习更加灵活，因为它不需要同时处理所有训练数据。

对于在校学生中，任务的顺序解决可能影响关键的学习者整体效率。我们采用 PAC 贝叶斯理论[24]问题研究证明一个概括的约束，这取决于用于解决任务的数据表示和算法。这些约束确定了任务被解决顺序的有效性，因此可以被用来找到一个有利任务完成顺序。基于这些约束，我们提出了一个理论上合理的算法，它能够选择一个有利的学习顺序。我们的实验结果表明学习任务在一个序列中可以优于独立学习和标准的同时解决多任务方法，我们的算法能够可靠地发现一个有利的秩序。

2. 相关工作

虽然我们的工作是基于传输信息和权重向量，其他方法多任务学习也已经提出了。在机器学习文献中的一个流行的观点是，相关任务参数可以被表示为一个小数目共同潜基向量的线性组合。Argyrio 等建议使用稀疏正规化学习这种表述方法[1]。该方法后来扩展到允许在[19]的任务组之间的部分重叠。它也适合于[31]终身设置，其中 Ruvolo 和伊顿公司提出了一种顺序更新模型作为新任务到达，在[27]进行讨论，其中终身学习的推广首先提出。在[30]，模型被扩展到允许学习者选择哪个任务下一个解，也允许做出几个启发性建议。从实验可以得到，基于子空间的方法在很多任务而且任务的基本特征是低维的情况下都取得了良好的效果。当特征维数变大，但是，他们的计算成本快速增长，这使得他们不适用于我们感兴趣的地方 1。一个例外是[15]，其中亚拉曼等人采用基于子空间的方法，共同学习多个属性预测。然而,即使有，维被要求减少.
基于共享的权重向量的方法也被推广，因为它们的原始引入[9]，特别一些宽松的假设如所有的任务都必须相关。在[8]，Evgeniou 等人，通过引入一个图正规化的概念达到了这种效果。另外，陈等人 [7]在权重向量为高度相关任务提出处罚偏差。然而，关于任务间的相似程度的方法需要先验知识。相比之下，在这项工作中，我们提出的算法不承担所有任务都是相关的，而且关于他们的相似之处也不需要先验信息。为了达到最佳性能，如何顺序学习阶段的问

题先前已被研究主要是在单任务学习的背景下，在这里问题是在哪一个那个顺序能够加速训练样本。在[4]Bengio 等人用实验表明在逐步增加的顺序中选择训练实例难度可以导致更快的训练和更高的预测质量。同样，库马尔等[20]介绍了自定步调学习算法，这个算法能够自动选择训练样本进行处理非凸学习问题的顺序。在多任务学习的语境中，以何种方式学习他们在[2]被介绍，在那里拉德等提出了一个算法，基于成对偏好的任务顺序优化。然而，他们只考虑了通过用户交互的顺序执行任务因此，他们的做法是不适用多任务场景的标准。Read 等人[29]提出多标签分类的设定、分解一个多目标问题成的想法、单一目标的人的序列。然而，信息的共享发生通过特征向量的扩增，而不是通过一个正则项。

3. 方法

在多任务的情况下一个学习系统观察多个监督学习任务，例如，识别对象或预测的属性。它的目标是解决所有这些任务由它们之间的信息共享。作为发展最快的现有的方法，但发现在实验中很难。简化的设置产生明显低于其他基准的结果。正式我们假设学习者观察 N 个任务，记$t_1, \ldots, t_n$它们共享相同的输入和输出空间，$\mathcal{X} \subset \mathbb{R}^d \mathcal{Y} = \{-1, +1\}$每一个任务$t_i$都有一个定义：$S_i\{(x_1^i, y_1^i), \ldots, (x_{m_i}^i, y_{m_i}^i)\}$。训练点$m_i$独立分布，根据概率未知比$D_i$属于$\mathcal{X} \times \mathcal{Y}$。我们还假设求解每个任务学习者使用线性预测 f (x) = sign (w, x)，$w \in \mathbb{R}^d$是一个权重向量，同时我们测量了 0/1 的损失分类性能，$l(y_1, y_2) = [[y_1 \neq y_2]]$，学习者的目标是找到$n$个权重矢量，$w_1, \ldots, w_n$这样对任务的平均预期错误$t_1, \ldots, t_n$（鉴于该预测是线性预测）最小化：

$$\mathrm{er}(w_1, \ldots, w_n) = \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbf{E}}_{(x,y)\sim D_i} [\![y \neq \mathrm{sign}\langle w_i, x \rangle]\!]. \quad (1)$$

3.1 以固定的顺序学习

我们建议将一个多任务问题分成n个领域适配问题。具体地，我们假设任务$t_1, \ldots, t_n$以$\pi \in \mathfrak{S}_n$方式处理，$\mathfrak{S}_n$是 n 个元素的所有排列的对称群，并且信息被后续任务之间传输：从$t_{\pi(i-1)}$到$t_{\pi(i)}$，$i = 2, \ldots, n$。在此过程中先前解决任务作为下一个任务附加信息的来源，而且任何的现有的域的适应方法都是可用的。本文中，我们使用一个自适应 SVM[16，36]训练分类为每一任务，由于其在计算机视觉应用中很有效。对于一个给定任务的权重向量$\tilde{w}$和训练数据，所述自适应 SVM 执行以下优化：

$$\min_{w} \ \|w - \tilde{w}\|^2 + \frac{C}{m} \sum_{j=1}^{m} \xi_j \qquad (2)$$

其中$y_j \langle w, x_j \rangle \geq 1 - \xi_j, \xi_j \geq 0$，$1 \leq j \leq m$具体地，学习任务$t_{\pi(i)}$（2）的预测，我们使用权重向量获得先前的任务$w_{\pi(i-1)}$记为$\tilde{w}$，对于第一次任务$t_{\pi(1)}$，我们使用标准的线性 SVM，即$\tilde{w} = 0$。为了简化符号，我们定义$\pi(0)$，即 0，$w_0$是 0 号向量。

请注意，这种方法不依赖于假设，所有的任务$t_1, \ldots, t_n$是相关的。但是效果取决于顺序π，因为它需要随后的任务是相关的。在接下来的部分，我们采用统计学习理论的研究这个问题，并介绍了一种算法自动确定有利的数据顺序。

3.2 学习数据关联的顺序

这里我们检查顺序π的作用是为了获得预期错误（1）最小。然而，我们并没有限制我们使用自适应支持向量机的理论分析。具体地，我们只假设学习算法用于解决每个单独的任务，$t_{\pi(i)}$和所有任务一样。$\mathcal{A}(w_{\pi(i-1)},\ S_{\pi(i)})$算法根据$w_{\pi(i-1)}$返回$w_{\pi(i)}$，$w_{\pi(i-1)}$来自于先前解决的任务和训练数据$S_{\pi(i)}$。下面的定理提供了平均预期的上界误差（1）所获得的预测因子（证明可以被发现在补充材料中）。

定理 1 对于任何确定的学习算法$\mathcal{A}$和$\delta > 0$，下列不等式的概率保持至少$1-\delta$（过采样集$S_1, \ldots, S_n$）一致的对于任何$\pi \in \mathfrak{G}_n$：

$$\frac{1}{2n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{(x,y)\sim D_i} [\![y \neq \operatorname{sign}\langle w_i, x\rangle]\!] \leq$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m_{\pi(i)}}\sum_{j=1}^{m_{\pi(i)}} \bar{\Phi}\left(\frac{y_j^{\pi(i)}\langle w_{\pi(i)}, x_j^{\pi(i)}\rangle}{\|x_j^{\pi(i)}\|}\right) + \frac{\|w_{\pi(i)} - w_{\pi(i-1)}\|^2}{2\sqrt{\bar{m}}}\right] + \frac{1}{8\sqrt{\bar{m}}} - \frac{\log\delta}{n\sqrt{\bar{m}}} + \frac{\log n}{\sqrt{\bar{m}}}, \quad (3)$$

当$\overline{m}$是样本大小的调和平均值，$m_1, \ldots, m_n$，$\overline{\Phi}(z) = \frac{1}{2}(1 - \operatorname{erf}(\frac{z}{\sqrt{2}}))$，$\operatorname{erf}(z)$是高斯误差函数[12， 23]，$\pi(0) = 0, w_0 = 0$ 和 $w_{r(i)} = \mathcal{A}(w_{\pi(i-1)},\ S_{\pi(i)})$。

不等式（3）左边产生任务的一半错误，这正是学习者想要减少的。然而，因为底层数据分布$D_1, \ldots, D_n$是未知的，所以它是不可计算。与此相反，不等式右边（3）给出了上限，但也仅包含可计算数量。这是一个n项平均数（是常数不依赖于$t_{\pi(i)}$），其中每一项都对应一个任务。如果我们认为与任务$t_{\pi(i)}$相对应的话，它的第一部分是训练误差的类似物。每一个项目$\overline{\Phi}(y_j^{\pi(i)}\langle w_{\pi(i)}, x_j^{\pi(i)}\rangle \|x_j^{\pi(i)}\|^{-1})$都有一个在 0、1 之间的值，而且是训练点$x_j^{\pi(i)}$和超平面$w_{\pi(i)}$的线性递减函数。具体来说，当它是正确分类，而且离超平面的距离较大，它接近 0，反之，接近 1。因此，抓住训练集预测的信心。这项任务的第二部分是一个复杂术语。它测量$t_{\pi(i-1)}$和$t_{\pi(i)}$的根号距离。其结果是（3）的右手侧的值可能对后续任务产生影响。它会导致算法通过最大限度地减少右手边（3）的基础上的数据公式而获得一个任务公式。因为（3）在公式中保持一致，它也为学习的秩序提供了保障。

最大限度地减少（3）的右手边是一个代价较大组合问题，因为它需要搜索所有可能的排列$\pi \in \mathfrak{S}_n$。为了执行此搜索的程序，我们提出了一种增量。我们先后确定$\pi(i)$通过最小化对应的上限（3）与尚未解决的任务。具体而言，在第 i 步，$\pi(1), \ldots, \pi(i-1)$已经定义，我们寻找一个任务t_k最小化目标函数，如下不包括π在内：

$$\frac{1}{m_k}\sum_{j=1}^{m_k} \bar{\Phi}\left(\frac{y_j^k\langle w_k, x_j^k\rangle}{\|x_j^k\|}\right) + \frac{\|w_k - w_{\pi(i-1)}\|^2}{2\sqrt{\bar{m}}}, \quad (4)$$

对于$w_k = \mathcal{A}(w_{\pi(i-1)},\ S_k)$，我们让公式的$\pi(i)$最大限度地减少（4）的指标。在每一步，我们选择是容易的任务（具有低的经验误差）和类似的前一个（相应的权值向量是接近）。因此，这一优化过程适合用最简单的任务开始，也适合大多数类似的程序。在

使用自适应 SVM（2）求解每一个任务的情况下，所得到的过程总结在算法 1，我们将其称为 SeqMT。但是，它可以被并行化，因为每个诸如虚拟机独立地与其它训练。

3.3 多序列学习

本文所提出的算法 SeqMT，依赖于所有任务可以在一个序列，其中每个任务相关的前一个订购的想法。在实践中，这并非总是如此，因为我们可以有不相关的任何其它任务离群任务，或者我们可以有任务的几组，在这种情况下，有利的是形成所述组内子序列，但它不利于它们连接成一个单一的序列。

因此，我们提出了 SeqMT 模型的一个扩展，允许任务形成子序列，在信息传递的唯一任务之间在序列。我们的多个版本，多 SeqMT，也选择任务迭代，但是在任何阶段，允许学习者选择是否继续一个现有的子序列或开始一个新的。为了决定哪些任务解决下，继续它的子序列，学习者可以进行两级优化。首先，每个现有的公式（包括空序列对应于没有转让的情况下）的学习者发现已有任务的公式，是最有前途的做法。在 SeqMT 算法以相同的方式完成下一个任务是如何选择。然后，学习者比较标准值（4）为每对公式和选择序列公式与最小值和继续它的任务公式。请参阅扩展技术报告[28]的确切公式。

算法 1：顺序多任务的学习

Input $S_1,\ldots,S_n$ {training sets}
$\pi(0) \leftarrow 0, w_0 \leftarrow \mathbf{0}$
$T \leftarrow \{1,2,\ldots,n\}$ {indices of yet unused tasks}
for $i = 1$ to n **do**
  for all $k \in T$ **do**
    $w_k \leftarrow$ solution of (2) using $S_k, w_{\pi(i-1)}$
  end for
  $\pi(i) \leftarrow$ minimizer of (4) w.r.t. k
  $w_{\pi(i)} \leftarrow w_k$ where $k = \pi(i)$
  $T \leftarrow T \setminus \{\pi(i)\}$
end for
Return $w_1,\ldots,w_n$ and $\pi(1),\ldots,\pi(n)$

4. 实验

在本节中，我们验证我们的两个主要要求：1）学习顺序的方式多个任务可以比共同学习起来更有效；2），我们可以自动查找平均分类精度方面的有利秩序。我们用两个可公开获得的数据集：动物与属性（AWA）3[22]和鞋子 4[5]增加带有属性 5[18]。在第一个实验中，我们研究的情况下，每个任务都有一定的难度，学习的对象类，这是由人类注释中定义的范围从最简单到最难的。我们展示了一个课程学习模式的优势，在学习多个任务，独立地学习每个任务。我们还研究更详细的自动确定订单，它们与订单相比较学习人类学习的精神，从最简单到最艰巨的任务去的时候。在第二个实验中，我们研究得知在不同的靴鞋的特点视觉属性的场景。在这种背景下，一些任务有明显关系，如高的鞋跟，并有光泽，但有些任务是没有，比如高跟鞋和运动。因此，我们我们也调整了我们算法，使它允许多个序列，表明它更好捕捉任务结构，因此是有利的学习战

略。

4.1 学习简单和困难的任务的顺序

我们专注于八类数据集从 AwA：黑猩猩、大熊猫、金钱豹、波斯猫、河马、浣熊、老鼠、密封，而人类的注解标明一个对象是否是容易图像识别[32]。对于每一个类的注释指定的是指图像从最容易到最难的排名分数。要创建简单的硬任务，我们将每个类中的数据拆分成五个相等的部分，根据它们的排名和使用的部件来创建五个类的任务。每个部分都有平均 120 的样品除了类鼠，因为 AwA 包含很少图片，所以有每部分只有约 60 的样品。每个任务是对其余七类的部件中的一个的二进制分类。对于每一个任务中，我们平衡 21 和 21 个训练图像和 77 比 77 的测试图像（35 比 35 在大鼠类的情况下），每个充当反面典型的类有等量样本。不同任务之间的数据不重叠。作为我们的特征表示，我们使用数据集 SURF 描述符[3]得到 2000 尺寸袋的词直方图。我们　使用规范 2 的特点，增加一个单位元素作为一个偏置的术语。

评价指标

为了评估我们使用的分类错误率的性能。我们重复各实验用不同的随机数据分裂 20 次，并测量整个任务的平均差错率。我们报告这个数值取的均为平均值和标准差。

基线

我们比较了连续的学习模式（SeqMT）与来自多任务算法[9]，[10]该把所有任务对称（MT）。具体来说，MT 转正为所有任务的权重向量得到了一个类似的原型。wo 通过优化问题与任务权重向量共同学习。

$$\min_{w_0,w_i,\xi_j^i} \|w_0\|^2 + \frac{1}{n}\sum_{i=1}^{n}\|w_i - w_0\|^2 + \frac{C}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\xi_j^i$$

$$\text{subject to } y_j^i\langle w_i, x_j^i\rangle \geq 1 - \xi_j^i, \quad \xi_j^i \geq 0 \quad \text{for all } i, j. \quad (5)$$

为了研究如何相关的知识转移实际上是，我们比较 SeqMT 用线性 SVM 基线独立解决了各个任务（工业 SVM）。作为参考，我们也训练对从所有的任务，并输出一个线性预测所有任务（合并 SVM）合并的数据的线性 SVM。

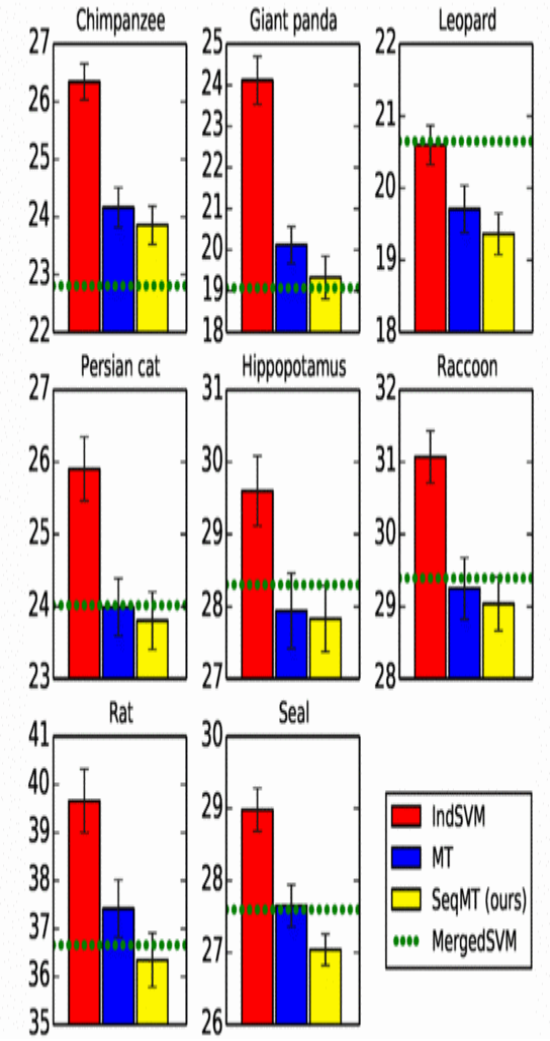

图 2，在 AwA 数据集中学习简单和困难任务的顺序：多任务（MT）和单任务（在 SCCM）基线用 SeqMT 方法的比较。条的高度对应于在 20 个重复（越低越好）的平均误码率性能超过 5 的任务。作为参考’我们还提供合并 SVN 基线，培养对从所有任务合并的数

据。与所有的结果的完整的表，请参考技术报告[28]。

要了解影响的任务顺序对分类精度，我们比较性能SeqMT与学习任务，按随机顺序基线（随机），并且为了从最简单到最难（语义），我们发现相关的另一个基线的多样性启发式[30]。它定义了下一个任务要解决可以通过放大（4），而不是最小化它。我们称它为作为多元化。所有基线进行了重新实施，使用相同的特征在所有的实验。

模型选择

我们进行交叉验证，选择正则化的模型选择方法，每一种方法有折中参数。在所有我们的实验中，我们选择超过8个参数值$\{10^{-2}, 10^{-1} \ldots, 10^{5}\}$使用5×5折交叉验证。

结果

我们提出这个实验的结果在图 2和图3显示。我们从图2中可以看到，该SeqMT方法在8个例子中均优于MT和 IndSVM 算法。这表明，知识转移之间的任务是显然是有利的，在这种情况下，它支持我们的假设，顺序的学习比共同学习他们更有效的尽管不是所有的任务都同样相关。正如预期的那样，在单任务的基线IndSVM 但在所有情况下，提高了基线 MergedSVM ，更多的数据训练具有更好的泛化能力。在某些情况下，执行的 MergedSVM 看齐或比 SeqMT和 MT 方法更好，例如，在黑猩猩和大熊猫的情况。我们预计，当任务非常相似，一个超平面可以解释大部分情况。在这种情况下，MergedSVM可以从超平面的数据量获得优势。当任务不同的话，MergedSVM 无法解释所有的情况共用一个平面，失去了SeqMT 和 MT 车型，每个学习任务都有一平面。这可以看出，在河马和密封的情况下，特别是在豹的情况下，MergedSVM 甚至独立训练也不获得提高。

接下来，我们检查哪项任务正在被解决的重要性，研究结果在图 3。在这项研究中，所有的方法使用自适应支持向量机作为一个学习算法，用于解决下一个任务和不同的任务的顺序如何被定义的问题。在所有 8 个例子中 SeqMT 算法优于随机基线，多样性算法比其他基线差很多，大概是因为选择的下一个任务的最大启发不是很有效。作为参考，我们还检查语义基线，从最简单的任务到最难解决的，（如果我们对任务 77 的易难顺序有先验信息），8 个类中有 6 个，我们的 SeqMT 模型学到的秩序（黄色菱形）优于或等同于语义（绿色广场），除了上课的黑猩猩和大熊猫，在这两个类中我们没有管理学习的最佳顺序。

有趣的是，对于某些类，语义学习比随机学习的顺序更糟糕（如密封和河马），我们认为这是事实，对人类比较容易的对于机器来说可能不那么简单。事实上，在密封和河马的情况下，人类和机器的理解而相反：人类最艰苦的工作对于机器学习来说最简单。因此，学习这些类随机顺序导致更好的结果比在一个固定的不利的顺序学习。我们通过单个支持向量机的错误率的分类为：最简单，容易，普通，困难和最难结果在图 4。

最后，对每类我们计算所有可能的订单的效果，学习 5 个任务，这导致 120 低音线，我们想象的所有订单的表现小提琴阴谋[14]，其中的阴影区域的每个水平切片反映多少不同的顺序实现这一误差率（在纵轴上表示的性能）。总的来说，SeqMT 是固定顺序之间的竞争激烈，显然在大鼠和密封两例（菱形低于黄色区域）占据

优势，然而在黑猩猩处于劣势。基于训练集的自适应学习顺序比一个固定的学习顺序有利。

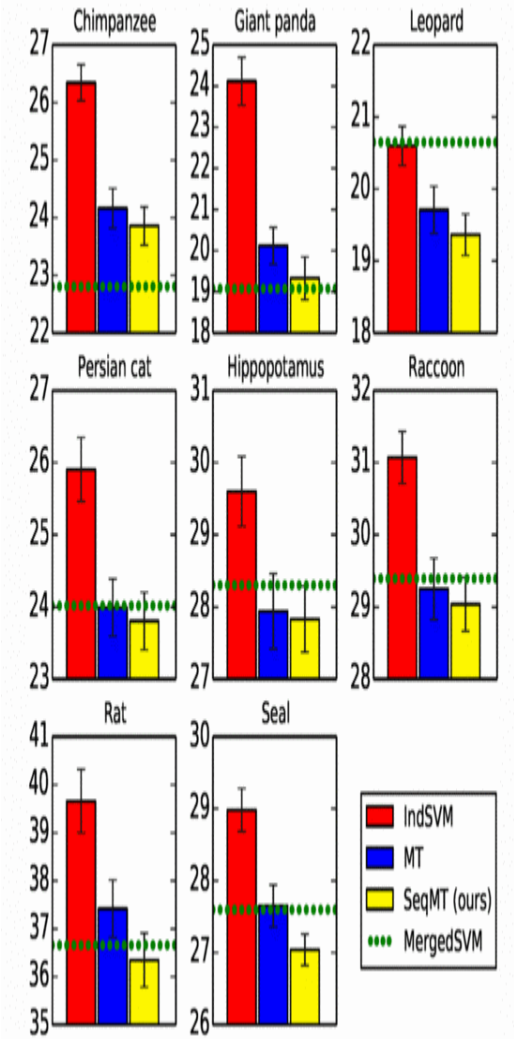

图 3. 不同的任务顺序策略研究与 AwA 数据集中的实验。四个主要的基准段，语义，随机的，分集有一个独特的标记物和颜色，以及它们的垂直位置捕获平均错误率性能（在纵轴上示出）。所有可能的订单的性能被可视作为背景小提琴情节，其中该阴影区域中的一个水平片段反映了许多不同的顺序如何实现这一错误率性能。请注意，此区域的对称性仅用于美观的目的。想要获得所有的结果，请参阅补充材料。最好看的颜色。

为了更好地了解定理 1 的实际利益，我们评估的性能之间的依赖关系的一个特定的任务顺序和它的排名，根据（3）。对于这一点，对于每一个数据拆分，我们排序所有可能的任务订单，根据相应的值（3），并计算它们的测试误差（平均超过 20 重复）以及相关系数之间的误差率和绑定的值。结果如图 5 所示。在所有情况下，除了黑猩猩有一个正相关的任务顺序的基础上的理论分析和测试性能的基础上，为定理 1 的有效性提供了证据。

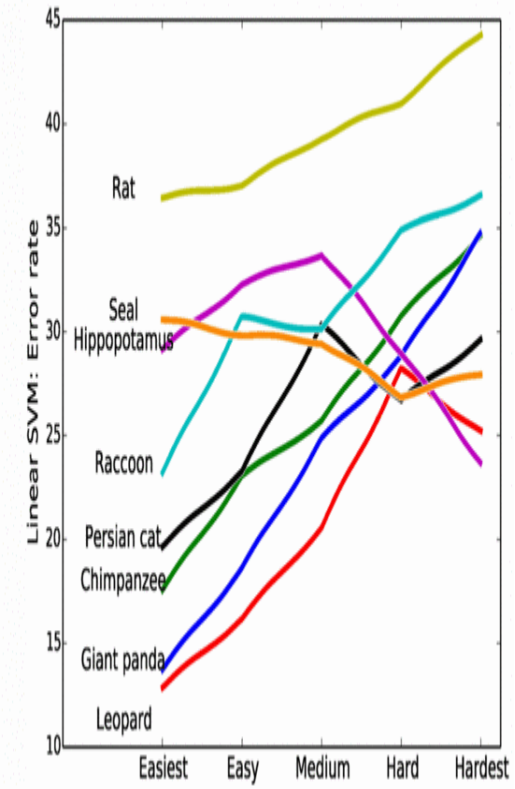

图 4。机器学习的性能可视化（线性 SVM）为该数据集上的实验结果提供了人类的注解。理想情况下，当人与机器的理解重合时，它将是一个自底向上的对角线。因此，在密封和河马的情况下我们不会期望在语义顺序学习会使我们最好的策略，在其他类也一样。最好看的颜色。

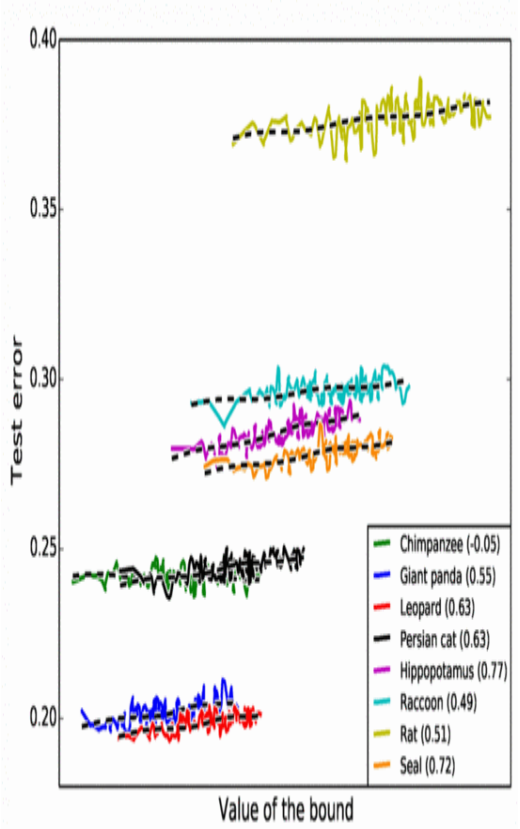

图 5.可视测试误差和在 AWA 实验结合的值之间的依赖。在每一个数据分裂的所有可能的任务命令进行排序，根据目标函数的相应值（3）和他们的测试错误的平均超过 20 个重复。在括号中的数字之间的相关系数测试错误和绑定的值。黑色虚线说明最佳的线性拟合。最好看的颜色。

	Chimpanzee	Giant panda	Leopard	Persian cat	Hippopotamus	Raccoon	Rat	Seal
Error+Compl	23.86 ± 0.33	19.33 ± 0.52	19.36 ± 0.29	23.81 ± 0.40	27.83 ± 0.46	29.04 ± 0.37	36.34 ± 0.57	27.04 ± 0.22
Error	24.47 ± 0.42	20.02 ± 0.58	19.97 ± 0.27	24.51 ± 0.16	29.07 ± 0.55	29.75 ± 0.31	38.00 ± 0.54	28.27 ± 0.28
Compl	23.94 ± 0.32	19.44 ± 0.70	19.36 ± 0.29	23.81 ± 0.40	27.83 ± 0.46	29.04 ± 0.37	36.34 ± 0.57	27.04 ± 0.22

表 1：在错误性和复杂性之间权衡，通过 seqmt 策略选择下一任务。数字是在五个任务中重复 20 次以上的平均错误率。

我们还研究了在目标函数（4）在选择下一个任务的重要性。为此，我们比较我们的算法选择下一个任务：基于训练误差（误差）和基于复杂度（并发症）。表 1 的结果表明选择下一个任务，复杂度，即任务之间的相似性，是更重要的组成部分，但它误差项组合有时更够达到更好的结果。最后，我们提出顺序学习的算法，以达到最佳的性能结果，并有利于所有其他策略，包括人类学习的顺序。

4.2 学习相关属性的子序列

我们专注于 10 个属性描述的鞋模型[18]：尖在前面，敞开的，颜色明亮，覆盖装饰品，有光泽，高的鞋跟，覆盖到腿上的，正式的，运动的，女性的。而且也从 10 鞋的数据集中获得了 10 个分类：运动、靴子、高跟鞋、木屐、公寓，泵、雨鞋、球鞋，高跟鞋，婚礼鞋。属性描述在类中从 1 到 10，10 表示该类有最多，1 表示该类有最少。我们形成 10 个二分类任务，顶端的两类样本为阳性（10 和 9 等级类）和最底端的两类样本为阴性（1 和 2 等级的类）。在属性类的描述更加澄清，见补充材料。对于每一个任务，我们平衡 50 比 50 的训练图像和 300 与 300 的测试图像，从每个类中随机采样相等数量。不同任务之间的数据不重叠。作为特征表示，我们使用 960 维的要点描述符的级联与公式化的 30 维的颜色描述符，增强与单位元作为偏置项。

基准线

除了所有基线在上一节所述，我们添加 MultiSeqMT 方法允许学习属性的多个序列（与信息传递的顺序）。此外，我们还包括一个基线 RandomMultiSeq 学习与选择随机启动一个新的子序列的随机属性。

Methods	Average error
IndSVM	10.34 ± 0.13
MergedSVM	29.67 ± 0.10
MT	10.37 ± 0.13
SeqMT (ours)	10.96 ± 0.12
MultiSeqMT (ours)	$\mathbf{9.95 \pm 0.12}$
Diversity	12.66 ± 0.17
Random	12.14 ± 0.20
RandomMultiSeq	10.89 ± 0.14

表 2 学习相关属性在鞋子的数据集。我们比较了 multiseqmt 和 seqmt 方法与多任务（MT）和单任务（indsvm）基线，并报告 mergedsvm 结果作为参考基准。我们研究的子序列的任务正在解决和比较我们的方法具有多样性的重要性，随机 randommultiseq 基线。我们研究的子序列的重要性，数字对应的平均错误率超过 10 个任务的 20 个重复（较低的更好的）。结果最好是用粗体突出。

结果：

我们目前的实验结果表 2。我们看到，该 multiseqmt 法优于所有其他的基线是一个有利的策略，在这种情况下。它比 seqmt 模型证实了学习的多序列是有利的，当不是所有的任务都有关。单任务学习基线 indsvm 实力较强，等同于多任务学习 MT 基准线，可能是因为多任务学习的强迫转移之间的任务无关的负面影响。正如预期的那样，mergedsvm 是无法用一个超平面解释所有的任务，在这种情况下表现很差。

类似于先前的实验中，我们研究序列和子序列其中任务正在解决的重要性。首先，我们比较多 SeqMT，并与学习一定的顺序任务（最后三行在表 2）基线 SeqMT 方法的性能，然后我们将分享我们的调查结果关于子序列属性的学习顺序。

我们可从表 2 可知，MultiSeqMT 是能将任务划分为子序列的最有效方法。学习随机 MultiSeq 做多个随机的子序列比学习的所有任务的单一序列，SeqMT，随机和多样性的基线做的更好。然而，由于 SeqMT 与随机 MultiSeq 相提并论，并明显比随机基线更好执行，我们得出结论，即使有一个序列，我们能够学习到一个良好的秩序，很少受到不相关的任务之间传输的影响。多样性基线比其他基线在这种情况下表现的更加糟糕。最后，我们分析了 MultiSeqMT 的序列。平均而言有 4.6 的序列，典型地，最长的是 5-6 的元素，有几对和几个单身。特别是有六个属性，有光泽，高的鞋跟，尖尖的正面，女性，公开，正式可以相互受益，常形成的相关任务的序列。在组内，高的鞋跟，有光泽的频繁启动子序列使两者之间发生传输互换。往往遵循前两次都是正面和女性，尖尖的下一个属性，它们也是为了密切相关可互换。属性公开并不总是在亚序列，但一旦它被包括在内，它将会转移到正式，这往往结束该序列。

剩下的四个属性，颜色明亮，布满装饰，腿部运动长，要么形成更小的子序列，有时两个任务，否则显示为单独的任务。偶尔也有从长的腿上的属性，以覆盖的饰品，我们相信这一事实的鞋靴类股票的这些属性的高等级。在一半的情况下，属性颜色明亮的运动和不为其他相关工作并形成自己的子序列。

5. 结论

在这项工作中，我们提出了一个连续的方式来解决多个任务，并研究

了问题，如果和如何学习者解决了一组任务的顺序影响其整体性能。首先，我们提供了一个理论结果：一个概括的约束，可以用来访问的学习秩序的质量。其次，我们提出了一个原则性的算法，选择一个有利的顺序的基础上的理论结果。最后，我们测试了我们的算法对2个数据集，并显示：1）学习多任务顺序可以更有效地比学习他们共同；2）的顺序中，任务的整体分类性能的影响；3）我们的方法是能够自动发现一个有益的顺序。

我们的模型的一个限制是，目前只允许从以前的任务转移到解决当前的一个，因此它输出的一系列相关的任务或多个任务序列。在未来的工作中，我们计划延长我们的模型，通过放松这种情况下，允许在一棵树，或更一般的图形结构的任务。

感谢

我们感谢诺维 Novi Quadrianto 有益的讨论。这项工作是由欧洲研究理事会的欧盟第七框架计划下的部分（FP7 / 2007-2013）/ ERC 拨款协议 308036。

References

[1] A. Argyriou, T. Evgeniou, and M. Ponti!. Convex multi-task feature learning. Machine Learning, 200S. 2

[2] Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In International Conference on Computer Vision (ICCV), 2011. 1

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Goo!. Speeded-up robust features (SURF). Computer Vision and Image Understanding (CV1U), 200S. 5

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In International Conference on Machine Learing (ICML), 2009. 2

[5] T. L. Berg, A. C. Berg, and 1. Shih. Automatic attribute discovery and characterization from noisy web data. In European Conference on Computer Vision (ECCV), 2010. 4

[6] R. Caruana. Multitask learning. Machine Learning, 1997. 1

[7] X. Chen, S. Kim, Q. Lin, J. G. Carbonell, and E. P. Xing. Graph-structured multi-task regression and an efficient optimization method for general fused lasso. arXiv: 1005.3579 [stat.ML], 2010. 2

[S] T. Evgeniou, C. A. Micchelli, and M. Ponti!. Learning multiple tasks with kernel methods. Journal of Machine Learning Research (JMLR), 6, 2005. 2

[9] T. Evgeniou and M. Ponti!. Regularized multi-task learning. In International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2004. 1,2,5

[10] J. R. Finkel and C. D. Manning. Hierarchical Bayesian domain adaptation. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2009. 5

[11] T. Gao, M. Stark, and D.

Koller. What makes a good detector?
- structured priors for learning from few examples. In European Conference on Computer Vision (ECCV), 2012. 1
[12] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In International Conference on Machine Learing (ICML), 2009. 3
[13] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In European Conference on Computer Vision (ECCV), 2012. 1
[14] J. L. Hintze and R. D. Nelson. Violin plots: A box plotdensity trace synergism. The American Statistician, 1995.
6
[15] D. layaraman, F. Sha, and K. Grauman. Decorrelating Semantic Visual Attributes by Resisting the Urge to Share. In Computer Vision and Pattern Recognition (CVPR), 2014. 2
[16] W. Kienzle and K. Chellapilla. Personalized handwriting recognition via biased regularization. In International Conference on Machine Learing (ICML), 2006. 3
[17] A. Kovashka and K. Grauman. Attribute adaptation for personalized image search. In International Conference on Computer Vision (ICCV), 2013. 1
[IS] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image Search with Relative Attribute Feedback. In Computer Vision and Pattern Recognition (CVPR), 2012. 4,7
[19] A. Kumar and H. Daume III. Learning task grouping and overlap in multi-task learning. In International Conference on Machine Learing (ICML), 2012. 2
[20] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In Conference on Neural Information P rocessing Systems (NIP S), 2010. 2
[21] A. Lad, Y. Yang, R. Ghani, and B. Kisie!. Toward optimal ordering of prediction tasks. In SIAM International Conference on Data Mining (SDM), 2009. 2
[22] C. H. Lampert, H. Nickisch, and S. Harmeling. Attributebased classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), 2013. 4
[23] J. Langford and J. Shawe-Taylor. PAC-Bayes and margins. In Conference on Neural Information Processing Systems (NIPS),2002. 3
[24] D. A. McAllester. Some PAC-Bayesian theorems. Machine Learning, 1999. 2
[25] F. Orabona, C. Castellini, B. Caputo, A. E. Fiorilla, and G. Sandini. Model adaptation with least-squares SVM for adaptive hand prosthetics. In International Conference on

Robotics and Automation (ICRA), 2009. 1
[26] S. 1. Pan and Q. Yang. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 2010. 1
[27] A. Pentina and C. H. Lampert. A PAC-Bayesian bound for lifelong learning. In International Conference on Machine
Learing (ICML), 2014. 2
[2S] A. Pentina, V. Sharmanska, and C. H. Lampert. Curriculum learning of multiple tasks. arXiv: 1412.1353 [stat.ML], 2014.
4,5
[29] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. In P roceedings of the
European Conference on Machine Learning and Knowledge
Discovery in Databases (ECML PKDD), 2009. 2
[30] P. Ruvolo and E. Eaton. Active Task Selection for Lifelong
Machine Learning. In Conference on Artificial Intelligence
(AAA!), 2013. 2,5
[31] P. Ruvolo and E. Eaton. ELLA: An efficient lifelong learning algorithm. In International Conference on Machine Learing
(ICML), 2013. 2
[32] V. Sharmanska, N. Quadrianto, and C. Lampert. Learning to transfer privileged information. arXiv:1410.03S9 [cs.CV],
2014. 5
[33] T. Tommasi and B. Caputo. The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In British Machine Vision Conference (BMVC),
2009. 1
[34] T. Tommasi, F. Orabona, and B. Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In Computer Vision and Pattern Recognition (CVPR), 2010. 1
[35] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In International Conference on Multimedia (ICM), 2007. 1
[36] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In International Conference on Multimedia (ICM), 20