# 第12章 小样本学习

[第12章 小样本学习 1](#_Toc83468315)

[12.1 小样本学习问题 1](#_Toc83468316)

[12.2 小样本学习算法 2](#_Toc83468317)

[12.2.1 基于数据增强的方式 3](#_Toc83468318)

[12.2.2 基于模型优化的方式 6](#_Toc83468319)

[12.2.3 基于迁移学习的方式 11](#_Toc83468320)

[12.3 零样本学习算法 14](#_Toc83468321)

[12.3.1 算法说明 15](#_Toc83468322)

[12.3.2 理论基础与基本模型 16](#_Toc83468323)

[12.3.3 归纳式零样本学习 17](#_Toc83468324)

[12.3.4 直推式零样本学习 19](#_Toc83468325)

[12.4 主动学习算法 19](#_Toc83468326)

[12.4.1 算法说明 20](#_Toc83468327)

[12.4.2 理论基础与基本模型 21](#_Toc83468328)

[12.4.3 查询函数设计准则 23](#_Toc83468329)

[12.4.4 样本选择策略 24](#_Toc83468330)

[习题 27](#_Toc83468331)

## 12.1 小样本学习问题

目前深度学习能够通过大量样本，训练出具有良好泛化能力的模型。但实际上，由于隐私性、标注成本等原因，难以获得大量的有标签样本，产生了小样本问题。小样本问题的产生主要有两方面原因：1.数据量足够但数据没有对应的标签，只有少量的带标签数据；2.本身数据量就比较小。

在小样本情况下，训练深度学习模型会出现严重的过拟合现象。例如，我们需要预估抛硬币正面的概率，正常情况下，抛的次数越多，概率越接近0.5，准确度越高，误差越小。但仅在少样本的情况下训练得到的模型，误差值较大。因此，小样本学习的核心问题是：仅使用少量标签数据训练模型，如何使之具有良好的泛化能力。

以深度学习的交通标识识别为例。在获取交通标识数据集的时候，由于某些交通标识不太常见，例如连续急转弯的标识主要出现在偏远山区，数据获取困难，样本量较少，导致深度学习训练模型时产生过拟合，模型预测效果较差。与此相反，人类只需要通过少量数据就能做到快速学习。一个五六岁的小孩子从未了解过交通标识的具体定义，但给他看过一张连续急转弯的交通标识图片，当他在马路上看到对应标识的时候，就会马上认出这还是自己曾经学习过的连续转弯的交通标识。

小样本学习可以概括为N-way、K-shot问题。训练集S包括N种样本，K表示每个类中的样本数，K通常小于20。N-way、K-shot是指从Meta-dataset中随机抽取N类样本，每类样本随机抽取K+1个实例，其中每类样本中的K个实例组成Support set，剩下的实例组成Test set。

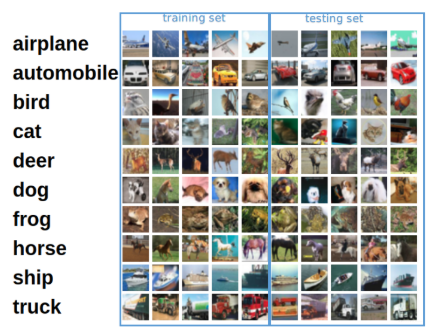


图12.1 传统机器学习数据集划分示意图

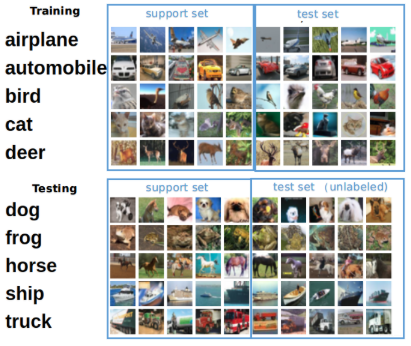


图12.2 小样本学习数据集划分示意图

如图12.1所示，传统的图像分类任务是基于左边的给定训练数据，获得model，然后在右边的数据集上测试model的好坏。而对于小样本问题，其训练数据和测试数据如图12.2所示。我们拥有的是大量的上方这些数据，也就是对于training中的airplane、automobile等，我们有很多类数据，而对于下方Testing中像dog、frog等新的分类问题，就没有那么多类的标注数据。

小样本学习就是期望机器学习能够更加接近人类思维，主要研究数据集中样本种类多、每类样本数量少的情况下的相关问题。使用远小于深度学习所需要的数据样本量，达到接近甚至超越大数据深度学习的效果。为了达到这一目标，小样本学习过程中利用了样本本身之外的一些信息，例如样本的属性和先验知识，或者其他相似任务下的样本信息，从而获得类似举一反三与触类旁通的能力。

## 12.2 小样本学习算法

小样本学习算法主要分成基于数据增强的方式、基于模型优化的方式和基于迁移学习的方式三种类型，具体分类如图12.3所示。

基于数据增强的方式是借助辅助数据或辅助信息，对原有的小样本数据集进行数据扩充或特征增强。数据扩充是向原有数据集添加新的数据，可以是无标签数据或者合成的带标签数据；特征增强是在原样本的特征空间中添加便于分类的特征，增加特征多样性。因此，基于数据增强的方式可以细分为基于特征增强、无标签/弱标签数据、和基于数据合成的方法三种。

基于模型优化的方式是用缩小假设空间的方式，来提升模型的性能。传统的深度学习问题较为复杂，并不能用一个较小的假设空间来表示。因此，通常会选择较大的假设空间，通过训练样本来优化参数，找到最优解。但在小样本情况下，就需要对模型进行优化，通过先验知识缩小假设空间，降低过拟合风险，提升模型泛化效果。

基于迁移学习的方式是目前比较前沿的方法，是指利用预测目标的相关性，从有大量数据的相关任务中训练好源模型，运用在少量样本的学习任务上，将已经学会的知识迁移到一个新的领域中。主要细分为基于度量学习、基于元学习和基于图神经网络的方法。

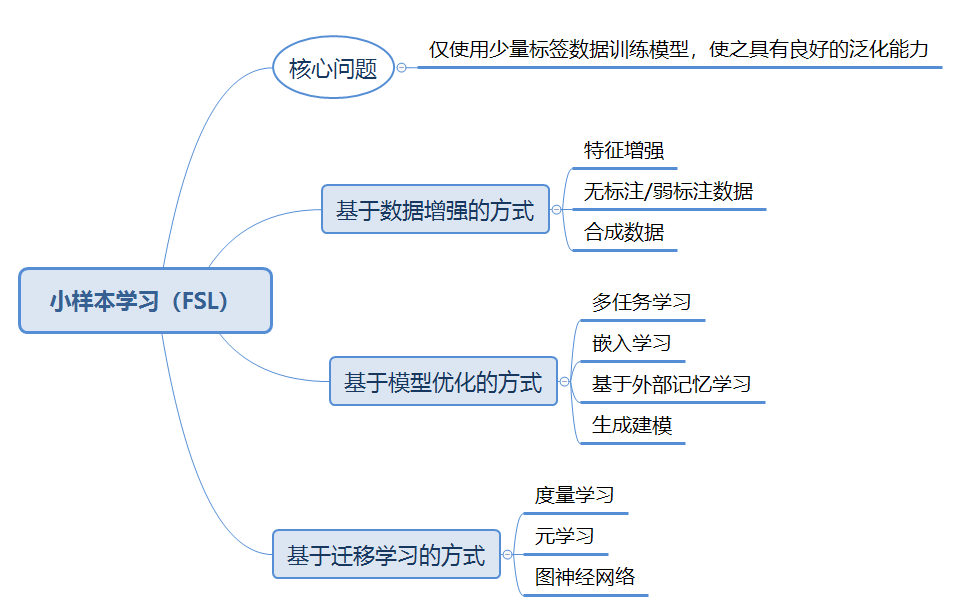


图12.3 小样本学习算法分类

### 12.2.1 基于数据增强的方式

生成对抗网络（GAN）是目前数据增强的主要方式。GAN模型由两部分组成：生成器（generator）和判别器（discriminator）。这里我们认为它们都是由参数确定的神经网络：G和D。判别网络的参数为最大化正确区分真实数据和伪造数据（生成网络伪造的数据）的概率这一目标而优化，而生成网络的目标是最大化判别网络不能识别其伪造的样本的概率。

生成网络生成样本：接受一个输入向量z，该向量取样自一个潜分布，应用由网络定义的函数G至该向量，得到G(z)。判别网络交替接受G(z)和x（一个真实数据样本），输出输入为真的概率。通过适当的超参数调优和足够的训练迭代次数，生成网络和判别网络将一起收敛（通过梯度下降方法进行参数更新）至描述伪造数据的分布和取样真实数据的分布相一致的点。

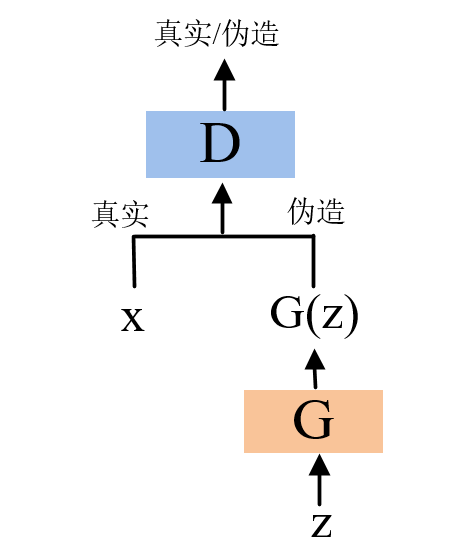


图12.4 GAN模型示意图

样本数据服从概率分布：，对于以为输入的分布，通过学习参数，使得接近，就可以确定生成器。其中参数的确定需用利用极大似然估计。

我们从随机采集一组样本，得到似然函数：，使得取得最大值的就是我们需要的值：

由于，所以：

我们需要找的就是使得取得最小值的参数。

通常情况下是一个神经网络，对于随机一个向量，通过可以生成样本。那么就可以通过一组，生成一个分布，从而与真实分布进行比较。由于神经网络可以拟合任意函数，那么也可以拟合任意分布，所以可以用正态分布，取样去训练一个神经网络，学习得到一个复杂的分布。这个过程可以看作：，其中表示神经网络输入时，输出恰好为。

对于上述过程，如果用最大似然估计就会存在问题，因为神经网络的参数量太大，通过计算最大似然对神经网络的参数进行估计是不现实的，所以就需要使用生成对抗网络GAN，用神经网络更新参数代替上述计算过程：用Gernerator代替，用Discriminator代替去约束Gernerator。

生成器Generator输入为，输出为。判别器Discriminator是一个全连接分类网络，输入为，输出为相应的样本标签。定义其损失函数：

所以期望的Generator：。

整个训练过程主要是交替执行下述两个步骤：（1）固定中所有参数，用梯度下降法修正。（2）固定中所有参数，用梯度下降法修正。以二分类模型为例，判断输入的图片是Real还是Fake。假定判别器输入，对应标签为，则N个样本的交叉熵为：

考虑有概率的样本，概率的样本，其中，那么：

所以：。在的博弈中GAN达到平衡，从而通过GAN模型获得增强数据样本。

### 12.2.2 基于模型优化的方式

对于有限样本的小样本，仅使用简单模型(比如线性分类器)，可选择一个小的假设空间。然而，现实世界中的问题通常很复杂，并且不能由一个较小的假设空间很好地表示。如图12.5(a) 所示，通常会选择较大的假设空间，通过大量训练样本来优化参数，找到最优解。FSL方法通过先验知识设法约束假设空间来进行学习(图12.5(b))，降低过拟合风险，提升模型泛化性能。

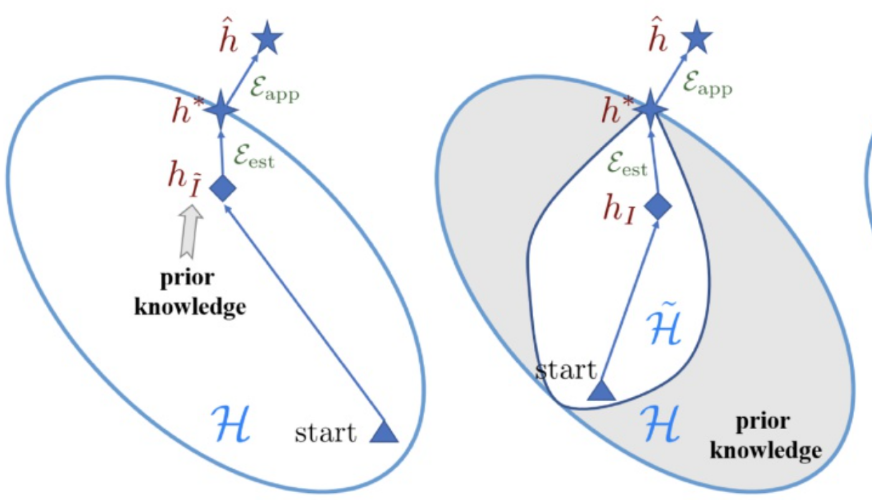


图12.5 a.大量样本情况 b.少量样本情况

**1.多任务学习**

在存在多个相关任务的情况下，多任务学习通过利用通用任务信息和特定任务的信息同时学习这些任务。多任务学习，其实是对目标任务做了一定的约束，或者叫做regularization。我们不希望模型只局限于目标任务的学习，而是能够适应多个任务场景，这样可以大大的增加模型的泛函能力(generalization)。这里的多任务的各个任务之间一定要有强相关性，如果任务之间本身的关联性就不大，多任务学习并不会对模型的提升并不一定会有用。

深度学习中有两种多任务学习模式：隐层参数的硬共享与软共享。

隐层参数硬共享指的是多个任务之间共享网络的同几层隐藏层，只不过在网络的靠近输出部分开始分叉去做不同的任务。如图12.6，通过共享特征和特征的embedding以及隐藏层的网络架构，在最后一层通过全连接+softmax的方式来区分不同任务，最后做一个线性融合来实现多目标排序。不同任务的最终目标不同，共享参数和特征其实会限制目标的差异性，最终模型的效果可能会不太好，如果目标间的相关性比较高，那么这种方式的效果会好一些。

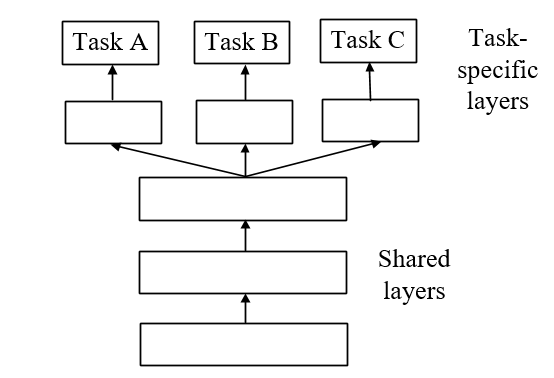


图12.6 参数硬共享示意图

隐层参数软共享是指不同的任务使用不同的网络，但是不同任务的网络参数，采用距离(L1,L2)等作为约束，鼓励参数相似化。如图12.7，每个任务都有自己的参数和模型结构，可以选择哪些共享哪些不共享。最后通过正则化的方式，来拉进模型参数之间的距离，例如使用 L2 进行正则化。在线上服务的时候，花费的时间要比硬参数共享机制化的多，因为模型结构更加复杂了。

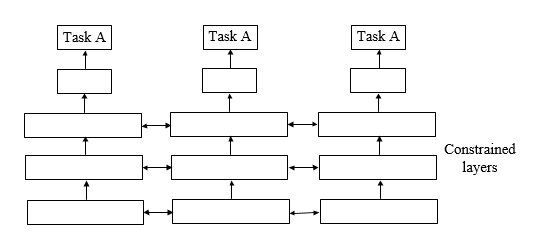


图12.7 参数软共享示意图

**2.嵌入学习**

嵌入（Embedding）的本质是“压缩”，用较低维度的 k 维特征去描述有冗余信息的较高维度的 n 维特征，也可以叫用较低维度的 k 维空间去描述较高维度的 n 维空间。在思想上，与线性代数的主成分分析 PCA，奇异值分解 SVD 异曲同工，事实上，PCA 和 SVD 也可以叫做嵌入方法。

嵌入学习将每个样本嵌入到低维，这样相似的样本会紧密靠近，而异类的样本则更容易区分。然后，在这个较低维的中，可以构造一个较小的假设空间，随后需要较少的训练样本。嵌入功能主要是从先验知识中学到的，并且可以额外使用训练样本任务特定信息。

嵌入学习具有以下关键组成部分：

（1）将测试样本嵌入的函数；

（2）将训练样本嵌入的函数

（3）相似性函数s来测量和中的之间的相似度。

根据该类的嵌入与中的最相似这个原则, 将测试样本分配给类。尽管可以为和使用通用的嵌入函数，但是使用两个单独的嵌入函数可以获得更好的准确性。根据嵌入函数f和g的参数是否随任务而变化，我们将这些FSL方法分类为：特定于任务的嵌入模型；任务不变(即一般)嵌入模型；混合嵌入模型，可同时编码特定于任务和不变任务的信息。

（1）特定任务嵌入模型

通过仅使用来自该任务的信息为每个任务量身定制嵌入函数。例如：给定任务Tc的少量数据，中样本之间的所有成对排名都列举为样本对，训练样本的数量因此增加，仅使用特定任务的信息也可以学习嵌入函数。

（2）任务不变嵌入模型

任务不变的嵌入方法从包含足够样本的大规模数据集中学习一个通用的嵌入函数，无需重新训练将其直接用于新的小样本。第一个FSL嵌入模型使用内核嵌入样本。

（3）混合嵌入模型

尽管任务不变嵌入方法可以应用在新任务上且计算成本低,但不利用当前任务的特定知识。当任务特性是只有少量样本的情况下，简单地应用任务不变的嵌入函数可能是不合适的。为了缓解这一问题，混合嵌入模型采用了一般的任务不变嵌入模型，该模型是根据数据训练中的特定任务信息从先验知识中学习而来的。这是通过学习一个函数来实现的，该函数将从中提取的信息作为输入并返回一个嵌入，该嵌入作为f的参数。

**3.基于外部记忆学习**

基于外部记忆学习：使用外部记忆学习从中提取知识，并将其存储在外部存储器中。然后，每个新样本用从内存中提取内容的加权平均值表示。这限制了由内存中的内容表示，因此实质上减小了假设空间H的大小。这个引入辅助记忆单元一般称为外部记忆，以区别与循环神经网络的内部记忆。

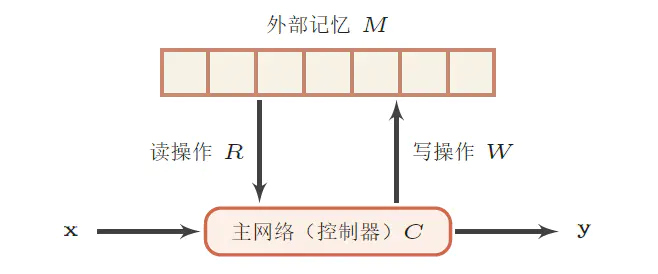


图12.8 外部记忆结构示意图

如图12.8时外部记忆结构的示意图，其中主网络C也称为控制器（Controller），负责信息处理，并与外界的交互（接受外界的输入信息并产生输出到外界）。主网络还同时通过读写模块和外部记忆进行交互。

外部记忆单元M用来存储信息，一般可以分为很多记忆片段（Memory Segment），这些记忆片段按照一定的结构来进行组织。记忆片段一般用向量来表示，外部记忆单元可以用一组向量来表示。这些向量的组织方式可以是集合、树、栈或队列等。大部分信息存储于外部记忆中，不需要全时参与主网络的运算。

读取模块R根据主网络生成的查询向量，从外部记忆单元中读取相应的信息。写入模块W：根据主网络生成的查询向量和要写入的信息a来更新外部记忆。

这种结构化的外部记忆是带有地址的，即每个记忆片段都可以按地址读取和写入。要实现类似于人脑神经网络的联想记忆能力，就需要按内容寻址的方式进行定位，然后进行读取或写入操作。按内容寻址通常使用注意力机制来进行。通过注意力机制可以实现一种“软性”的寻址方式，即计算一个在所有记忆片段上的分布，而不是一个单一的绝对地址。

比较典型的记忆网络是端到端记忆网络，可以多次从外部记忆中读取信息。在端到端记忆网络中，外部记忆单元是只读的。给定需要存储的信息M，将其转换成两组记忆片段：用于寻址的片段A和用于进行输出的片段C。主网络根据输入来生成q，并从外部记忆中读取相关信息r，产生输出：。为了实现更复杂的计算，可以进行k轮交互，第k轮交互产生的查询向量则为：，相应的输出为：。

**4.生成建模**

生成方法：由数据学习联合概率分布, 然后由求出概率分布作为预测的模型，这就是需要确定的生成模型。该方法表示了给定输入X与产生输出Y的生成关系。生成建模方法借助先验知识从观测到的估计概率分布，此类中的方法可以处理许多任务，例如生成，识别，重构和图像翻转。根据变量代表的信息，现有方法可分为三种：可分解的元件、相似类的先验概率分布和推理网络参数。

（1）可分解的元件

尽管在FSL问题中缺少具有监督信息的样本，但它们可能与其他任务的样本共享一些较小的可分解组件。这种方法利用从先验数据中学习到的可分解的各个元件，如人脸识别中的五官等，从中学习各元件与目标类的关系，将其与目标类联结起来。

（2）相似类的先验概率分布

通常，相似的任务具有相似的先验概率，并且可以在FSL中使用。例如，考虑“橙色猫”，“豹”和“孟加拉虎”的三级分类，这三个物种相似，但孟加拉虎濒临灭绝，而橙色猫和豹则丰富，因此，人们可以学习一种 来自“橙色猫”以及“豹”的先验概率，并以此作为few-shot类别“孟加拉虎”的先决条件。该方法利用相似数据类的先验概率分布，将小样本数据分类至若干个对应的相似数据类中，利用类的先验概率分布建模。

（3）推理网络参数

该方法利用现有的深度网络的参数的强特征表征能力来表征新的任务数据，更加高效且降低了对人类知识的要求。通常使用一些辅助大规模数据集对推理网络进行训练，典型的包括变分自编码器（VAE），自回归模型，生成对抗网络（GAN）, VAE和GAN的组合等。

如图12.9，生成模型种类多样，根据概率密度函数的处理方式不同，生成模型可以分为图示几种类型。

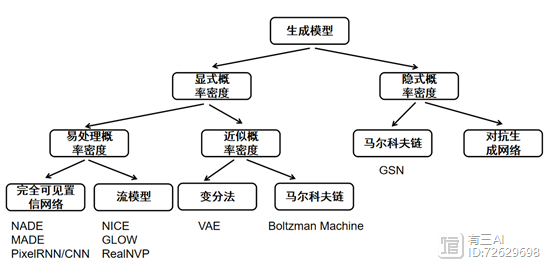


图12.9 生成模型的类别

### 12.2.3 基于迁移学习的方式

迁移学习是指利用旧知识来学习新知识，主要目标是将已经学会的知识很快地迁移到一个新的领域中。例如，一个程序员在掌握了 C 语言的前提下,能够更快地理解和学习 Python 语。迁移学习能够解决小样本问题。迁移学习解决了源数据集和目标数据集分布不一致时所产生的过拟合问题，只需要源领域和目标领域存在一定关联，使得模型学习到的知识和特征能够帮助在目标领域训练分类模型，从而实现知识在不同领域之间的迁移，通常可以细分为基于度量学习、基于元学习和基于图神经网络的方法。

**1.度量学习**

度量学习是一种空间映射的方法，其能够学习到一种特征空间，在此空间中，所有的数据都被转换成一个特征向量，并且相似样本的特征向量之间距离小，不相似样本的特征向量之间距离大，从而对数据进行区分。在度量空间中的学习异常高效，在小样本分类时效果很好。

在训练模型的过程，我们随意的选取两个样本，使用模型提取特征，并计算他们特征之间的距离。如果这两个样本属于同一个类别，那我们希望他们之间的距离应该尽量的小，甚至为0；如果这两个样本属于不同的类别，那我们希望他们之间的距离应该尽量的大，甚至是无穷大。

大边界最近邻算法LMNN是最常使用的一种度量学习算法，其可以通过对训练集学习来得到一种原始数据的新度量，这种方法可以在一定程度上对原始数据分布进行重构，得到一个更加合理的数据分类空间。

LMNN学习了为k近邻分类设计的伪测量，是以监督方式学习该全局度量的算法，以提高k最近邻规则的分类准确性。该算法基于半定规划，是凸优化的子类。LMNN背后的主要直觉是学习伪测量，在该伪测量下，训练集中的所有数据实例被至少k个共享相同类标签的实例包围，该算法学习该类型的伪测量：，其中矩阵M 需要是正半正定的。

以文本分类为例，首先要对文本进行特征提取将待测试文本和训练文本表示成向量空间模型。定义表示训练文本集合，为类别集合，其中表示第i篇文章，表示文本向量的第i维，采用IG算法作为特征提权算法，然后采用LMNN方法对训练数据集进行重构，最后使用 K 近邻分类器来实现文本分类，评价标准使用F1值和查准率、查全率。

具体流程：

（1）首先，对中文文本进行分词、去停用词等预处理。

（2）对文本进行特征选择，可以选用IG 这种常用的的特征提取算法来对文本进行特征提取。

（3）构造向量空间模型(Vector SpaceModel，VSM)，可以采用经典TF\*IDF 法。

（4）对训练样本以欧氏距离用留一法计算出训练集中每个数据点的先验知识 K近邻，并做好标签，设定此 K 值为 Kp 。

（5）利用 LMNN 算法对训练集进行学习，求出映射矩阵 L。

（6）对训练样本和测试样本分别作映射。

（7）跟据基于 LMNN 的文本分类算法对测试集进行分类。

**2.元学习**

元学习（Meta-Learing），又称“学会学习”（Learning to learn）, 即利用以往的知识经验来指导新任务的学习，使网络具备学会学习的能力。元学习中要准备许多任务来进行学习，而每个任务又有各自的训练集和测试集，是解决小样本问题（Few-shot Learning）常用的方法之一。

元学习的本质是增加学习器在多任务的泛化能力，元学习对于任务和数据都需要采样，因此学习到的F(x)可以在未出现的任务中迅速（依赖很少的样本）建立起mapping。因此“元”体现在网络对于每个任务的学习，通过不断的适应每个具体任务，使网络具备了一种抽象的学习能力。

Meta-learning中为了区别概念，将训练过程定义为“Meta-training”、测试过程定义为“Meta-testing”, 如下图所示：

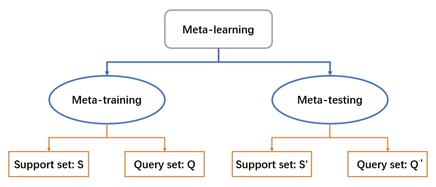


图12.10 元学习示意图

区别于一般神经网络端到端的训练方式，元学习的训练过程和测试过程各需要两类数据集（Support/Query set）。小样本分类任务属于“N-way， k-shot”问题，其中，N代表选择的Testing data中样本的种类，k代表选择的K类Testing data中每类样本的数量，一般来说N小于Testing data的总类别数。

我们从Testing data中随机选出 N 个类。然后，再从这 N 个类中按照类别依次随机选出 k+x个样本（x 代表可以选任意个），其中的 k 个样本将被用作 Support set S'，另外的x个样本将被用作 Query set Q'。S和Q的构建同理，不同的是从Training data中选择的样本类别和每类样本数量均不做约束。

Meta-learning 通常采用一种被称为 Episodic Training 的方法来进行训练，它偏重于任务和数据的双重采样，即任务和数据一样是需要采样的。具体来说对于一个10分类任务，元学习通过可能只会建立起一个5分类器，每个训练的episode都可以看成是一个子任务，而学习到的F(x)可以帮助在未见过的任务里迅速建立mapping。

**3.图神经网络**

图（Graph）数据包含着十分丰富的关系型信息。从文本、图像这些非结构化数据中进行推理学习，例如句子的依赖树、图像的场景图等，都需要图推理模型。图是一种结构化数据，它由一系列的对象（nodes）和关系类型（edges）组成。作为一种非欧几里得形数据，图分析被应用到节点分类、链路预测和聚类等方向。图神经网络（Graph neural networks）是一种链接主义模型，它靠图中节点之间的信息传递来捕捉图中的依赖关系。

图神经网络GNN的思想是利用嵌入将图中的节点通过向量来表示，然后再利用得到的嵌入结果，经过一般的深度神经网络模型进一步研究。本质上，图中的每个节点都与一个标签相关联，我们的目的是预测没有 ground-truth 的节点的标签。图神经网络有很多种变体，比较常用的有图卷积神经网络(graph convolutional network)、门控图神经网络(gated graph neural network)和图注意力网络(graph attention network)等

在图神经网络里，每一个样本被看作图中的一个节点，该模型不仅学习每个节点的嵌入向量，还学习每条边的嵌入向量。卷积神经网络将所有样本嵌入到向量空间中，将样本向量与标签向量连接后输入图神经网络，构建每个节点之间的连边；然后通过图卷积更新节点向量，再通过节点向量不断更新边的向量，这就构成了一个深度的图神经网络。

如图12.11所示，5个不同的节点输入到 GNN 中，根据公式A构建边，然后通过图卷积更新节点向量，再根据A更新边，再通过一层图卷积得到最后的点向量，最后计算概率。

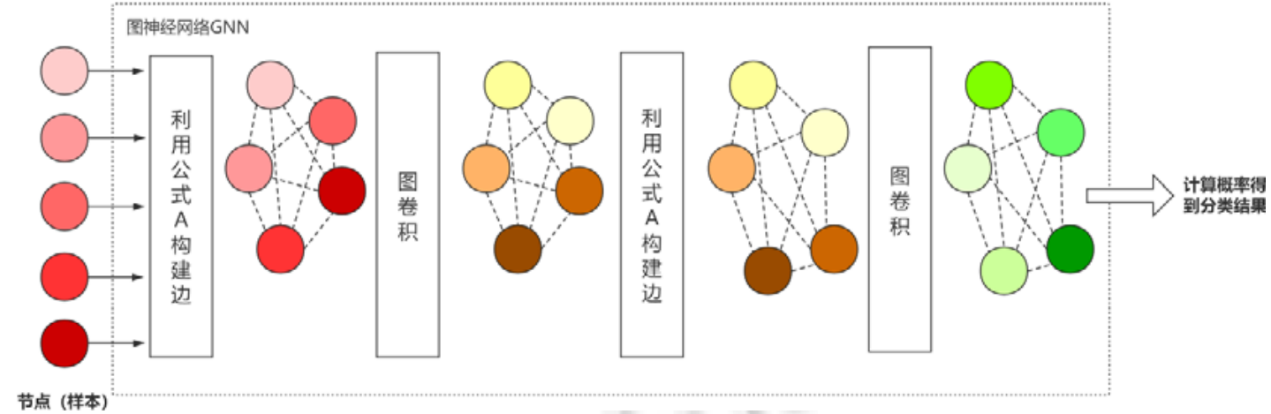


图12.11 GNN模型流程图

## 12.3 零样本学习算法

小样本学习试图仅从一个或少量几个有标签训练样本中解决目标学习，而当目标任务的训练数据完全缺失时，如何解决该学习问题，则是零样本学习（zero-shot learning）的学习目标。

简单来说，零样本学习就是让计算机具备人类的推理能力，使其可以识别出一个从未见过的新事物。比如，我们告诉一个从没见过斑马的小朋友：“斑马长得像马，身上有像老虎一样的条纹，并且像熊猫一样是黑白色的”，他就可以很轻松地在动物园里找出哪个是斑马。

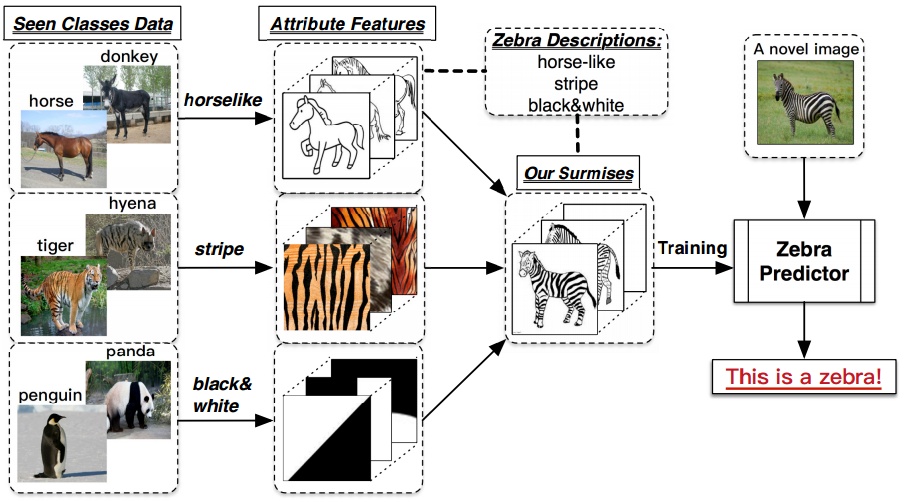


图 12.12 斑马案例

零样本学习就是将上述的推理过程抽象为通过已知信息加上辅助信息进而推断出新出现对象的类别的过程。因此，推理过程中已知的信息（马，老虎，熊猫）为训练集，辅助信息（马的外形、老虎的条纹、熊猫的颜色）为训练集与测试集相关联的语义信息，新出现对象（斑马）为测试集。

### 12.3.1 算法说明

零样本学习主要涉及三类数据：已知类、未知类和辅助信息。其中已知类是模型训练时用到的带类别标签的图像。未知类是模型测试、训练时不知道类别标签的图像。辅助信息是对已知类和未知类图像的描述、语义属性或者词嵌入等信息，该信息充当了已知类和未知类之间的桥梁。

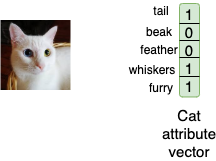


图12.13 数据集示意图

如图12.13所示，数据的集合由图像x、标签y和辅助信息组成，辅助信息存储为一个由 5 个元素 (尾巴，喙，羽毛，胡须，毛茸茸) 组成的向量。猫有尾巴，而且毛茸茸，因此在向量中，对应于尾巴和毛茸茸的位子为 1，而在其他位置为 0。

S表示已知类的数据集，U表示未知类的数据集。集合表示已知类的图像集，集合表示已知类的类别标签集，集合已知类的属性集。而集合表示未知类的图像集，集合表示未知类的类别标签集，集合未知类的属性集。

基于嵌入的方法的主要目标是使用投影函数将图像特征和语义属性映射到一个公共的嵌入空间，该投影函数是使用深度网络学习的。共同的嵌入空间可以是视觉特征空间、语义空间或重新学习的中介空间。大多数基于嵌入的方法都使用语义空间作为公共嵌入空间。

在训练期间，目标是使用来自已知类的数据学习从视觉空间（即图像特征）到语义空间（即词向量/语义嵌入）的投影函数。由于神经网络可以用作函数逼近器，因此投影函数自然可以选用深度网络来学习。

在测试阶段，将未知类图像特征输入到前面训练好的网络里，获得相应的语义嵌入。然后在语义属性空间中进行最近邻搜索，以找到与网络输出最接近的匹配项。最后，将与最接近的语义嵌入相对应的标签预测为输入图像特征的输出标签。

图12.14展示的是基于嵌入的零样本学习方法的流程框图。首先将输入图像通过特征提取器网络（通常是深度神经网络），以获取图像的 N 维特征向量。此特征向量充当我们主投影网络的输入，而主投影网络又输出 D 维输出向量。目的是学习投影网络的权重/参数，以便从视觉空间中的 N 维输入映射到语义空间中的 D 维输出。损失函数用于度量 D 维输出与真实语义属性之间的误差。然后就是训练网络的权重，使 D 维输出尽可能接近真实语义属性。

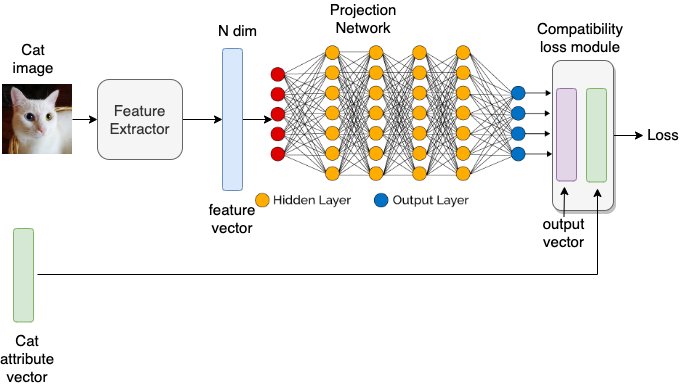


图12.14 基于嵌入的零样本学习方法示意图

Top-1 准确率是图像识别模型常用评估指标。但零样本识别模型通常会使用每类 Top-1 的平均准确率，即分别找到每个类别的识别准确率，然后在所有类别上取平均值。

### 12.3.2 理论基础与基本模型

在零样本学习问题中，测试类与训练类之间相互互斥，我们不能简单地从有标签训练数据集中学习测试类的分类器。为了对这些未知测试类进行预测，我们必须在未知类和已知类之间引入耦合关系，从而建立两者之间的语义关联，这样我们就可以从已知类的训练数据集中抽取到相关知识，并使用这些知识进行未知类的预测。因此，我们可以在图片特征空间与物体标签空间中引入共享的中间层语义嵌入空间，从而将共享的概念或知识从已知类迁移到未知类中。

这样，每个物体类不再仅仅被视为一个互相之间相互独立的原子标签，而是用语义嵌入空间中的一个语义向量表示（也称类原型“class prototype”），这些语义向量刻画了类与类之间的语义关联关系。基于语义嵌入空间，我们可以首先使用已知类的训练数据学习图片特征与类原型之间的关系，然后使用该关系对测试数据进行两步法预测，即首先预测测试数据对应的语义向量，其次找到与该语义向量最匹配的类，这个类就是我们要预测的未知类。正式地，我们使用表示m维语义嵌入空间，分类器可表示为以下形式:

其中函数通常由训练数据集学习而来，函数可以是空间中的最近邻算法，也可以是基于其他更复杂的相似度的分类方法。

基于语义嵌入空间，根据未知类数据在零样本学习模型中的使用情况，我们主要有两种类型的零样本学习方法，即归纳式（inductive）零样本学习方法和直推式（transductive）零样本学习方法。

### 12.3.3 归纳式零样本学习

归纳式零样本学习方法假设在模型训练过程中函数的学习只使用训练数据，而且在测试过程中，函数以并行方式对所有测试数据进行类标签预测，即所有测试数据的标签预测过程是相互独立的。任意测试数据的标签预测为：

由于其具有良好的灵活性和可扩展性，归纳式学习方法是目前零样本学习问题中应用最广泛的一种解决思路。其中，最经典的学习模型是Christoph等人在2009年提出的直接属性预测模型（direct attribute prediction，简称为DAP）和间接属性预测模型（indirect attribute prediction，简称为IAP），其分别采用不同的方式构造图片特征空间与语义嵌入空间的关系函数。作为零样本学习问题的基本解决方案，DAP和IAP在很大程度上启发了后续的基于语义嵌入空间的零样本学习方法。

（1） DAP模型

DAP模型使用训练数据集直接学习图片特征到属性的映射，如图12.15所示。

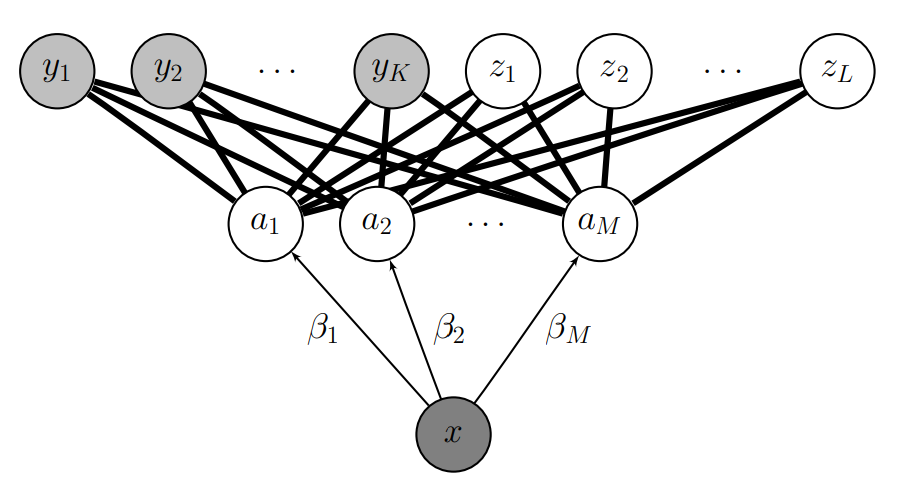


图12.15 直接属性预测（DAP）示例图

假设我们有*K*个已知类，*L*个未知类和M个语义属性，我们使用0/1值表示属性与物体类之间的关系，0表示物体类不具备该属性，1表示物体类具有该属性，每个类可表示为属性空间中的二值向量和。这样，我们可以使用分类器建模图片特征与属性之间的关系。对于每个属性，我们使用训练数据学习概率分类器。在测试时，我们使用公式计算每个未知类的后验概率，

其中为每个未知类的先验概率。我们将数据分类为具有最大后验概率的类，即

此外，我们也可以使用连续值的属性向量来表示每个类中具有多大强度的属性，在这种情况下，我们使用回归模型来预测函数。

（2） IAP模型

与DAP不同，IAP模型通过使用已知类的类标签间接学习图片特征到属性的映射，如图12.16所示。首先，我们使用传统的有监督学习方法为每个已知类学习概率分类器。其次，我们使用已知的属性向量来预测每个测试数据的属性，即

与DAP模型一样，我们使用公式来预测未知类标签。

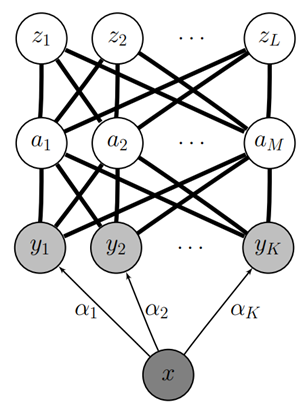


图12.16 直接属性预测(DAP)示例图

### 12.3.4 直推式零样本学习

与归纳式方法不同，直推式零样本学习方法允许在分类器*l*的学习过程中同时使用有标签训练数据和无标签测试数据。在基于语义嵌入空间的零样本学习方法中，我们有以下两种应用方式:

（1）与上文中的半监督学习类似，一方面，测试数据可以单独用在函数*f*的学习过程中，从而提高*f*在未知类上的泛化能力或迁移能力；另一方面，测试数据也可用来联合优化*l*通过在训练过程中迭代优化*f*和，我们可以一次性得到所有测试数据的标签，即

其中是训练数据在空间*K*中的语义嵌入向量，是的类标签。

（2）在测试过程中，所有测试数据同时用于函数中，通过使用整个测试数据集的结构信息，从而一次性预测所有数据的类标签。也就是说，测试数据的类标签预测过程之间不再是相互独立的，而是彼此之间相互影响。所有测试数据的标签预测为：

## 12.4 主动学习算法

机器学习中的有监督学习和半监督学习都需要大量的标注数据，但是在实际的业务场景或生产环境中，标记样本的获得是一个困难、耗时且昂贵的过程。因此如何通过较少的成本来获得较大价值的标注数据，并进一步地提升算法地效果成为一个值得思考的问题。主动学习（Active Learning）就是为了解决此类问题而提出的一种算法，在统计学领域也叫查询学习、最优实验设计。

在没有使用主动学习（Active Learning）的时候，通常来说系统会从样本中随机选择或者使用一些人工规则的方法来提供待标记的样本供人工进行标记。这样虽然也能够带来一定的效果提升，但是其标注成本总是相对大的。

主动学习是解决小样本问题的关键技术之一。相对于被动学习而言，被动学习基于已经标注好的数据集上进行模型训练，这些标记数据集是人工随机选取，造成有限标记训练样本中的信息量严重匮乏；主动学习通过人机交互方式，基于查询函数，有选择性的从大量未标记数据中选取信息量最丰富的样本进行人工标记，使得小样本数据中的信息尽可能的代表未标记数据中信息量，从而实现小样本下对大数据的分析处理，同时降低耗费。

例如，一个高中生通过做高考的模拟试题以希望提升自己的考试成绩，那么在做题的过程中就有几种选择。一种是随机地从历年高考和模拟试卷中随机选择一批题目来做，以此来提升考试成绩。但是这样做的话所需要的时间也比较长，针对性也不够强；另一种方法是每个学生建立自己的错题本，用来记录自己容易做错的习题，反复地巩固自己做错的题目，通过多次复习自己做错的题目来巩固自己的易错知识点，逐步提升自己的考试成绩。主动学习就是选择一批容易被错分的样本数据，让人工进行标注，再让机器学习模型训练的过程。

### 12.4.1 算法说明

主动学习可以使用尽可能少的标注数据集训练一个模型，使得模型性能可以达到一个由大量的标注数据集按照普通方法（随机选择训练数据）训练得到的模型的性能。以数据集Cifar10为例，该数据集共有60000张彩色图像，这些图像大小是32\*32，分为10个类，每类6000张图。 主动学习的过程如下。

（1）数据：，；。

（2）初始化模型，随机权重得到最初的模型，记为

；。

（3）分别对 目录下的 张数据进行预测，得到10个类别对应的10个概率值；

（4）重点关注每个样本预测结果的最大概率值：。我们初步认为 的情况表示当前模型对该样本有个确定的分类结果（此处分类结果的正确与否不重要）；反之，当前模型对该样本的判断结果模棱两可，标记为；比如：模型进行第一次预测，得到10个概率值，取其最大的概率；

（5）对（此处的10分类任务取）的样本进行排序，取前个样本加入集合中；

（6）基于当前的训练数据集对模型进行微调，得到新的模型记为 ；

（7）重复（3）到（6）步骤，直到 样本数为 0 或者当前模型 已经达到理想效果

实验结果如表12.1所示，实验结果表明引入active learning 不仅能够得到减少样本标注代价，还能够提升分类的准确率。验证集准确率val\_acc就能够达到99.04%，将剩余的训练样本扔到训练好的模型进行预测，剩余训练样本的准确率Acc\_left\_active\_samples能够达到99.39%。

表12.1 Cifar10分类实验结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method | Train Set | Val Set | Model | Val-acc | Acc\_left\_active\_samples |
| Just Train | 50000 | 10000 | AlexNet | 90.00% | - |
| Random select | 44000 | 10000 | AlexNet | 90.06% | - |
| Active learning | 17750 | 10000 | AlexNet | 90.16% | 99.39% |

### 12.4.2 理论基础与基本模型

主动学习的大致思路就是：通过机器学习的方法获取到那些比较“难”分类的样本数据，让人工再次确认和审核，然后将人工标注得到的数据再次使用有监督学习模型或者半监督学习模型进行训练，逐步提升模型的效果，将人工经验融入机器学习的模型中。

在实际应用中，很多问题是动态变化的，需要不断修正数据，否则很难保证训练模型的准确性。比如，在微博文本分类中，用户随着兴趣爱好的改变，微博状态也会随之改变，以前用户关注的是生活、旅游和科技，可能在某个时期对武器、军事等产生兴趣，微博转发内容发生改变，这时候就需要人工标记少许文本和更新模型，从而提供更加个性化的服务。主动学习的适用场景较为灵活和宽泛，仅仅需要少量的标记样本，就可以逐步建立可靠数据集。

主动学习技术主要通过提升训练样本的质量来提升分类模型的性能。通过样本数量进行训练集信息量的扩充，同样可以达到提升分类器性能的目的。其模型如下: A=(C,Q,S,L,U)，其中 C 为一组或者一个分类器，L是用于训练已标注的样本。Q 是查询函数，用于从未标注样本池U中查询信息量大的信息，S是督导者，可以为U中样本标注正确的标签。学习者通过少量初始标记样本L开始学习，通过一定的查询函数Q选择出一个或一批最有用的样本，并向督导者询问标签，然后利用获得的新知识来训练分类器和进行下一轮查询。主动学习是一个循环的过程，直至达到某一停止准则为止。

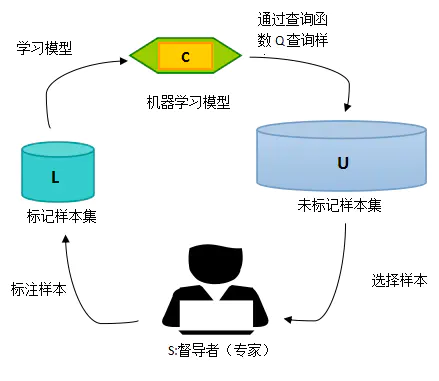


图12.17主动学习示意图

主动具体流程如下：

（1）选取合适的分类器（网络模型）记为 current\_model 、主动选择策略、数据划分为 train\_sample（带标注的样本，用于训练模型）、validation\_sample（带标注的样本，用于验证当前模型的性能）、active\_sample（未标注的数据集，对应于ublabeled pool）；

（2）初始化：随机初始化或者通过迁移学习（source domain）初始化；如果有target domain的标注样本，就通过这些标注样本对模型进行训练；

（3）使用当前模型 current\_model 对 active\_sample 中的样本进行逐一预测（预测不需要标签），得到每个样本的预测结果。此时可以选择 Uncertainty Strategy 衡量样本的标注价值，预测结果越接近0.5的样本表示当前模型对于该样本具有较高的不确定性，即样本需要进行标注的价值越高。

（4）专家对选择的样本进行标注，并将标注后的样本放至train\_sapmle目录下。

（5）使用当前所有标注样本 train\_sample对当前模型current\_model 进行fine-tuning，更新 current\_model；

（6）使用 current\_model 对validation\_sample进行验证，如果当前模型的性能得到目标或者已不能再继续标注新的样本（没有专家或者没有钱），则结束迭代过程。否则，循环执行步骤（3）-（6）。

### 12.4.3 查询函数设计准则

在各种主动学习方法中，查询函数的设计最常用的策略是：不确定性准则、代表性准则以及两者的联合准则。

不确定性准则是主动学习中最受关注的准则，其研究比较宽泛，包括边缘采样，专家委员会，期望误差减小等。对于不确定性，我们可以借助信息熵的概念来进行理解。我们知道信息熵是衡量信息量的概念，也是衡量不确定性的概念。信息熵越大，就代表不确定性越大，包含的信息量也就越丰富。

事实上，有些基于不确定性的主动学习查询函数就是使用了信息熵来设计的，比如熵值装袋查询（Entropy query-by-bagging）。所以，不确定性策略就是要想方设法地找出不确定性高的样本，因为这些样本所包含的丰富信息量，对我们训练模型来说就是有用的。

查询函数每次迭代中查询一个或者一批样本。我们当然希望所查询的样本提供的信息是全面的，各个样本提供的信息不重复不冗余，即样本之间具有一定的差异性。在每轮迭代抽取单个信息量最大的样本加入训练集的情况下，每一轮迭代中模型都被重新训练，以新获得的知识去参与对样本不确定性的评估可以有效地避免数据冗余。但是如果每次迭代查询一批样本，那么就应该想办法来保证样本的差异性，避免数据冗余。

代表性准则，在去除冗余的同时，通过分布估计使得标记数据集和未标记数据集之间具有相似的分布，挖掘数据内部结构信息。不同于一般的差异性策略，代表性准则不需要预先依据不确定性准则进行样本的预筛选，而是可以直接看作是主动学习查询函数，其目标是保证选择的样本与未标记数据集具有独立同分布的特性。

主动学习过程中准则的联合学习主要是将不确定性和代表性结合到同一个主动学习框架中，查询既具有不确定性又具有代表性的样本。例如，利用稀疏表达进行样本相似性表达，将不确定性的衡量作为稀疏表达中稀疏系数的权重，实现从不确定性高的样本中选择相似性表达能力强的样本。

### 12.4.4 样本选择策略

从主动学习的核心过程可以发现，主动学习过程中最重要的部分是查询函数 F的确定。上一节介绍了查询函数的设计准则，本节将具体介绍目前较为常用的样本选择策略，主要有不确定采样、专家委员会、期望模型变化和期望模型最小化四种策略。

**1.不确定采样**

不确定性采样是主动学习中最常用和最经典的方法，其查询的是在分类过程中分类结果可靠性最低的样本。比如，边缘采样策略，通过衡量点到面的距离，选择离分类平面最近的点作为查询的最具信息量的样本

边缘采样对于样本不确定性的描述较为直接，衡量值越小，说明样本可靠性越低。目前很多主动学习方法都是基于此选择策略进行主动学习算法的进一步优化和提升。

边缘采样仅考虑了一种标签的可能性，针对多类别问题可以采用最大的类别预测概率和第二大的类别预测概率进行不确定性的衡量，称为 BvSB(Bestvs Second Best)，表达式为

表示类别归属概率最大值，表示类别归属第二大概率值。越小则表明样本被判断为类别和的概率越相同，其主要查找位于类别和边界上的样本。

但是当类别数较多时，BvSB 则会忽视其它类别的分类信息，为此对所有的分类损失进行综合考虑，提出了基于熵的主动学习方法：

上述三种方法是目前主动学习算法中较为常用的策略。

**2.专家委员会**

专家委员会是主动学习中另一种样本选择机制，主要利用有限的训练样本构造多个弱分类模型，将这多个分类模型看成是专家委员会对未标记样本进行预测，选取委员会预测结果最不一致的样本, 作为最具有信息量的样本。常用的机制有投票熵：

其中表示的是专家委员会的个数，表示专家委员会中对预测标签为 的个数。

平均 KL(Kullback-Leibler) 散度也是常用的投票不一致性指标，表达式为:

其中：

表示专家委员会中第个分类模型将x判属于的概率，表示专家委员会中所有模型将 x 判属于的概率和的平均值。专家会员会相当于利用多个模型进行集成，具有集成学习的优点，在专家委员会个数较少时便可以查询到最具信息量的样本。对于专家委员会的不一致性度量，还有很多指标比如 Jensen-Shannon 散度等。

**3.期望模型变化**

期望模型变化策略与不确定性采样和专家委员会不同，其主要是选取能够使当前模型产生最大变化的样本。就是当选取的样本标记以后加入训练集时，训练模型能够产生最大变化。定义表示目标函数中模型参数 θ 的梯度，使表示训练集加入训练样本以后模型参数的新梯度。由于 是待查的样本，其标签是未知的，因此，必须用对所有标签进行梯度的计算作为期望来代替真实的模型变化衡量，其目标函数表达式为:

其中表示加入样本梯度向量的欧式范数。应该注意到，相对于加入的单个样本，训练集 L 的数量在梯度中作用更大，因此为了加快学习效率，可以在训练模型梯度上直接进行梯度求解，即 ≈ ，这样对于数目多的数据集可以进行并行计算。但是如果要真实的计算其期望模型变化，样本数量和标签数量大时，此方法的计算时间复杂度较高。事实上，模型期望变化就是衡量样本在无标签下对于模型改变的影响。

**4.期望误差减小**

期望误差减小不同于期望模型改变，其主要查询能够使模型泛化误差最大限度减小的样本。通过进行模型训练，在剩余的未标记样本集 U 上进行模型泛化误差的衡量，由于假定未标记集和测试集是独立同分布的，因此，在 U 上的测试可以表示模型的泛化能力。但是由于样本的标签是未知的，因此仍然采用当前模型参数对所有标签进行遍历，其目标函数就是期望减少预测错误标签的数量，可以表示为最小化期望的 log-loss：

除了熵值形式，还可以通过最大化样本的增益信息或者互信息，来选取最具信息量的样本。

## 习题

1. MNIST数据集是机器学习领域中非常经典的一个数据集,由60000个训练样本和10000个测试样本组成，每个样本都是一张28 \* 28像素的灰度手写数字图片。请尝试使用GAN生成MNIST数据。

2. IMDB-WIKI数据集包含超过 50 万张带有年龄和性别标签的人脸图像（<https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/> ）。请尝试对数据集进行微笑检测、情绪识别和性别分类的多任务学习，可以参考

（<https://github.com/truongnmt/multi-task-learning> ）中的方法。

3. MiniImageNet包含100类共60000张彩色图片，其中每类有600个样本。地址：<https://few-shot.yyliu.net/miniimagenet.htm> 跟踪了关于该数据集的前沿算法，请尝试使用元学习的方法对数据集进行分类。

4. Omniglot 数据集包含来自 5050 个不同字母的 16231623 个不同手写字符。每一个字符都是由 2020 个不同的人通过亚马逊的 Mechanical Turk 在线绘制的（数据集和样例地址：<https://github.com/brendenlake/omniglot> ）。请尝试使用小样本相关算法，尽可能提升模型分类的准确率。

5. AWA数据集包含50个动物类别，85个属性，数据集下载地址：<http://cvml.ist.ac.at/AwA2/AwA2-data.zip> ，请尝试使用零样本学习方法对数据集进行分类。

6. SUN数据集包含131067个图像，由908个场景类别和4479个物体类别组成，其中背景标注的物体有313884个。数据集下载地址：<http://groups.csail.mit.edu/vision/SUN/> ，请尝试使用零样本学习方法对数据集进行分类。

7. CIFAR-10 数据集共有60000张彩色图像，这些图像是32\*32，分为10个类，每类6000张图，请尝试使用主动学习的方法对数据集进行分类，提高分类准确率。

8.请尝试采用不同的主动学样本寻找策略（RS策略、LC策略、BT策略），对MNIST数据集进行分类。（RS策略：随机选择，作为对照。LC策略：寻找分类器最没有信心的预测样本，即预测的最可能类别的概率也很低的样本。BT策略：寻找分类器最“左右为难”的预测样本，即预测的最可能的两个类别的概率很接近的样本。）