Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach

1.引言

这是一篇比较老的文章，方法可能不适用于现在，但还是有一定的借鉴价值。

2.背景

在很多机器学习任务中，模型的训练和测试时所采用的样本分布是一致的，但在实际中我们模型训练时候的样本分布和实际使用时候的样本分布在很多情况下是不一致的，这就导致了领域适应性问题（Problem of Domain Adaptation）。如果为每一个数据集都训练一个自己的模型，则需要耗费大量的资源。Domain Adaptation 尝试去建立一个在 training 和 Testing 都适用的模型，用概率统计表示为 P(X)不等于 P(X');P(Y|X)约等于 P(Y|X')。

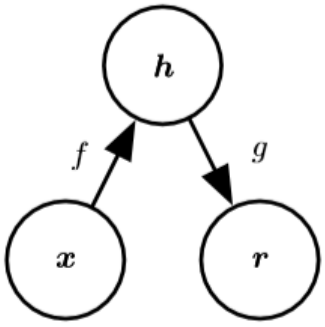
本文基于的假设是不同的分布，其提取出的抽象特征的分布更相似，并在这些抽象特征上进行分类。本文最大的贡献就是采用无监督深度学习技术（Stacked Denoising Auto-encoders）提取出了这些有相似分布的抽象特征。

3.Stacked Denoising Auto-encoders

3.1 Auto-encoders

自编码器是一个三层神经网络，该网络可以看做两个部分，一个由函数 h=f(x)表示的编码器（encoder）和一个由 r=g(h)表示的解码器（decoder）。其中 x 为 input data，h 表示 hidden data，r 表示 output data。一般用平方误差作为自编码器的误差函数。

当 f，g 都是线性变换函数时，这个自编码器的功能等同于 PCA。



3.2 Denoising Auto-encoders

去噪自编码器是在 input data 中加入一些噪声，目的是为了提高模型的泛化能力。本文采用的去噪自编码器是在 input data 中每一维的数据以 p 的概率取值为 0（类似于 dropout 技术）。本文的去噪自编码器在 encoder 中采用的非线性函数是 relu，在 decoder 中采用的非线性激活函数是 sigmoid。

3.3 Stacked Denoising Auto-encoders

Stacked Denoising Auto-encoders 即在 Denoising Auto-encoders 上添加一/多层 Auto-encoders，输入为前一个自编码器的隐藏层数据，并且本文在输入数据中加入高斯噪声。

3.4 无监督深度学习

自编码器和深度信念网等无监督深度学习技术，在前些年的作用在于降维、深度神经网

络预训练。目前有人运用无监督深度学习技术做半监督任务。

传统深度学习网络得训练方法是，初始化参数，然后用反向传播算法训练，但随着神经网络层数得增加，会有梯度消失或梯度爆炸得风险。用深度无监督技术，逐层贪心预训练。预训练结束后，在大部分情况下神经网络的参数值到了局部最优值的附近，然后通过最顶层的监督训练对整个网络进行微调，可以快速收敛。近些年又有了基于监督学习的贪心预训练方法。

4.分类

在 Stacked Denoising Auto-encoders 的顶层提取出的数据中采取 SVM 算法分类。

5.不足

这篇文章基于的假设是，不同域的数据集用无监督技术提取出的抽象特征具有相似性，于是在一个数据集上训练的模型直接拿到第二个数据集上实验。

首选我们要弄清楚一点我们的学习器在不同的分布上学习到的模型的参数是不同的。比如在高斯分布的数据上学习到的神经网络，直接拿来用于均值分布的数据上测试，效果会差很多，这也是 Problem of Domain Adaptation。作者认为两个数据集的抽象特征的分布更相近，那作者应该不仅仅通过最后分类实验的结论来证明这个相似性。比如，作者可以在原始数据集上计算两者的 KL 散度 a，再在抽象特征上计算两者的 KL 散度 b，再进行比较两者大小，这样会更加有说服力。

再者，如果目标数据集是有标签的，作者可以把源数据集上训练的模型在目标数据集上进行微调以更好的适应新的数据集（具体的方式没有想好）。

思考

自编码器的作用是无监督的提取数据的高级抽象特征，也看作对数据的降维。对于 PCA，最初学习的时候，资料上给的解释是需要降维后保持原本数据更多的信息，从方差上考虑，有的资料可能考虑的更加全面一些，但如果从编码、解码的方向思考，可能更符合它的本质。

Domain Adaptation 是迁移学习中的一种，虽然在这方面已经有了不少工作，但我觉得它还是有研究价值，并没有哪个工作可以普适性的解决各种问题。现在的 ML/DL 都是数据驱动的，并且绝大部分都是监督学习，工作需要大量的标签数据，需要大量的计算资源。而实际中的标签数据是少数，并且很多存在域适应性的问题。在数据驱动的研究期间，迁移学习应该一直有可做点。

参考文献

1. Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: A deep learning approach[C]//Proceedings of the 28th international conference on machine learning (ICML-11). 2011: 513-520.