模式识别期末报告

171240510 马宇恒

1.问题提出

在图像识别与检索领域，FV和VLAD等方法取得了很好的效果。这些方法通过对局部特征（如SIFT）等进行编码，将原始的N维数据映射到更低的K维特征空间中，记录下映射的码本，然后在特征空间上执行任务。然而随着数据量和数据维度的增加，算力与内存逐渐缺乏。于是需要探索数据降维、简化数据更新过程的手段。

现有的解决这类问题的通用办法是feature compression的手段。但是这类方法的特点是对所有的feature维度做一个投射，所以任何维度都会影响压缩后的数据。Introduction中介绍了这类手段的三个类别，即Product Quantization（分割+编码），哈希类型手段和数据降维技巧（PCA等）。

而另一类手段，feature selection，则是选择所有feature的一个子集，使得子集里的feature能够提供大多数信息（其实也是特殊的feature compression，但系数为0-1，类似于lasso）。这类办法同样能够降低feature维度，但是大量的维度提供了0信息。

总体来说，人们的目标是找到符合以下几点的数据降维办法：有效信息损失少（任务精确度高），运算量小，内存占用少，调整超参数（压缩率）方便，可解释性强。

2. 文章观点与方法

本文提出的观点是：feature selection比feature compression在以上几个维度做的要好。其中可解释性文章并未提到，但我认为参考lasso相比岭回归的例子可以看出，选择一些变量比选择一些变量的线性组合更容易解释。文章用以下方法来证明这一观点：

1. 首先经验性的证明在得到的fisher vector或者VLAD中，不同feature之间几乎没有强相关性。所以没有理由去寻找他们的线性组合来降维，而是更适合单独考虑每个feature，删去其中是噪音的部分。这一事实佐证了feature selection的合理性。
2. 通过对两者计算方式的比较说明，FS只需要计算一次重要性排行就可以应付所有的压缩率，而FC每次都需要计算新的系数。这一事实佐证了feature selection调整超参数的方便。
3. 进行实验。图像分类任务使用基于互信息（MI）的重要性排行执行feature selection，并在此基础上分类；图像查找任务使用基于熵的重要性排行执行feature selection。并将以上两组实验的结果与feature compression的结果对比。分类任务中feture selection的正确率较高，查找任务中也给出了不俗的正确率。同时feature selection计算时间显著低于feature compression。这一事实说明了feature selection的任务精确度高、运算量小。

3. 文章的创新点

1. 文章总体的考虑了feature selection作为一类方法与feature compression的优劣。
2. 文章通过讨论FV中各个feature的相关性来经验性的反驳了feature selection的做法。
3. 提出了两个使用FS的标准化步骤。一是通过基于MI的重要性排行来执行FS，二是基于熵的重要性排行来执行FS。二者在直觉上都是合理的。其形式化如下：
   1. 图像识别：计算FV中feature与标签的互信息，记录互信息（重要性）排序，从高到低选取适当多的feature，执行任务。
   2. 图像查找：计算FV中feature的熵，记录熵（重要性）排序，从高到低选取适当多的feature，计算距离（离散化的使用对称距离，连续的使用非对称距离，而且显然非对称距离保留信息较多，准确率较高）。

4. 讨论

1. 有关feature离散化、讨论储存空间和密码本的部分，对于没有修过数据结构的同学并没有很直观，需要花时间思考这些做法的意义。
2. 证明FV的非相关性为经验证明，没有理论证明，并且只使用了一个数据集作为例子，证据不够充分。
3. Feature selection优于FC的结论是否能够单纯的由MI、Entropy两种方法得到？我认为文章只尝试了这两种模型，在特定情况下的结论有些武断。
4. 离散化中N-bins的划分方法是否可以尝试使用线性划分之外的方法？例如采用直方图+分布函数的方法，求出N个大致等可能的区间。该方法计算略复杂于原来，但直觉上能提供更多的信息。