# Sense Discovery via Co-Clustering on Images and Text阅读笔记

## 介绍

不同于传统的聚类方法假定聚类中文字特征和视觉特征是一对一的映射，考虑到每个语义感知可能对应于多个视觉感知（视觉感知的视角和外观多样性），作者假定两个域之间是一对多的映射。

歧义是现在面临的比较重要的问题，一个名词短语可能有多个语义和视觉含义。作者归回到了人工组建的知识库，这些知识库覆盖面广但是细粒度过高，且存在对命名实体的资料较少（人、地点、组织），这些知识库也缺乏对视觉感知的数据。

作者提出了一个联合聚类算法，对于给定的名词短语，联合寻找语义和视觉感知。

单独使用图片，难以区分是视角的变化还是观念的变化，而单独使用文字则难以区分近似的语义，因此作者联合两者，进行联合聚类。

## 联合聚类方法

给定两个域，目标是联合聚类两个域中的实例。传统方法通过增加特征空间，在联合空间中进行聚类，其他方法都预先假定了两个域存在一一对应的映射。然而域存在不同的粒度，比如一个语义聚类可能关联了不同的视觉聚类（考虑视角、外观区别）和语音聚类（考虑发音的不同）。因此作者打破这种限制。

#### 公式

首先在域中，假定存在一对多的映射。算法输入是个数据点，表示为，其中表示域中的第个数据点。输出是各域的一系列聚类簇，以及一系列聚类簇和两个域间的一对多映射。

一对多的映射表示为一个无向图，是域中的某一个聚类，是图中的边集，表示域的聚类对应域的聚类。为了约束一对多的关系，。每个在域的聚类节点都对应了模型参数。数据点的聚类成员关系被表示为，表示数据点属于域中的聚类。

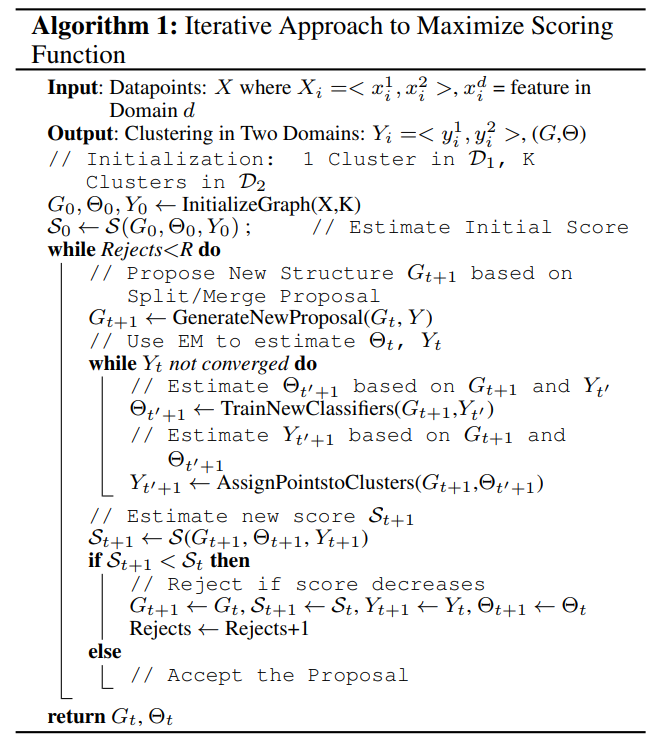
因此给定数据点，目标就是推断出，以最大化得分函数

第一项对应了两个域的数据似然得分，其更青睐被转化后，更相近的聚类簇。第二项是平滑项，其尝试保证如果在域中是相似的，那么应该被分配到相同的聚类簇中。这个项是一方面为了正则化第一项，避免最终每个数据点被分配一个聚类簇，另一方面提供了类内和类间簇距离的度量，使得聚类结构的步步转化是合适的。第三项是一个跨域项，如果两个实例在域相似，那么它们应该在域中处于同一个聚类簇。注意这是非对称的，其目的是为了约束一对多的映射关系。

#### 使用迭代方法优化

给定图的结构（固定聚类簇数量和映射数量），迭代的求解和。给定，通过估计从属关系，进而将数据点分配到图节点中。强制约束一对多的映射，那些不满足当前图中映射的数据点会被丢弃，低得分的数据点会被视为噪声并被忽略。一旦得到了关系，则用这个关系去估计新的。

对给定估计后，下一步优化网络结构。首先提出结构的修改手段，将节点一分为二或者合并两个节点。通过逼近方法贪心地选择图结构作为。利用新的结构，重新估计分数，如果得分更高则接受结构，否则拒绝并让。初始化时，在上有一个节点，在上有个节点（使用较大的使得结构步骤更加平滑），在中偏向分裂节点，在中偏向合并节点。



## 探索语义和视觉感知

运用算法到现在这个任务上，是文字域，是视觉域。输入的数据点是一系列名词短语以及利用Google查询这些短语得到的图片，并下载前1000张网页。

#### 文字域

给定一个名词短语和包含此短语的网页，可用以下方法提取出：

首先用Stanford parser对句子做句法分析。对于每个提及的短语，提取的特征包含距离短语头一步或两步的依赖路径。此外在网页上提取词袋模型特征。在很多case中，相关联的文本可能主题是和短语无关的，为解决这个问题，词袋模型表示被约束语句需要出现当前短语。此外还使用了Part-of-speech tagging（词性标注）来构建词袋模型，同一个单词在各语境下的不同词性就可以区分开了。这样会得到非常高维的特征向量，为解决这个问题，使用LDA主题模型（不是线性判别模型，LDA主题模型目的是从文章中找到指定数量的主题）从网页中训练，将BOW投影到主题上，构成最后的。

为了表示文本聚类和联系参数，作者考虑将每个聚类表示为所有聚类成员向量的均值，似然项就可以对输入特征向量和均值的直方图交叉核（评估两个特征集的相似度）。平滑项被定义为

其中是直方图交叉核，是indicator function（指示函数，用于将断言或条件的成立与否转化为1或0的输出）。这个平滑项为那些具有相同标签且相似的数据点提供奖励。奖励和两个数据点之间的点积成比例。

#### 图像域

因为图像中目标的位置是未知的，建立似然模型很困难。为克服这一点，作者利用Enriching Visual Knowledge Bases via Object Discovery and Segmentation一文中所使用的算法，为图像寻找检测边界，确定目标位置。同时将图像聚类到个聚类簇中（和前文初始化图时在域构建个节点对应）。然后使用HOG特征表示。

每个聚类簇用一个线性SVM表示，参数化为。训练时将属于聚类簇的数据点作为正样本，随机边界框的图像作为负样本。

每个属于聚类的数据点的似然分数可以定义为。为了平滑，额外添加一个相似度项，利用在HOG特征上的点乘运算表示相似度。最终平滑项可以表示为

#### 跨域项

跨域项用来保证在视觉域相似的数据点被分配到相同的文字域聚类簇中（即一对多的映射关系）。其被表示为

利用在HOG特征上的点乘运算表示相似度。

#### 优化

上文说到使用EM结构的方法优化这个模型，需要交替迭代分割文字节点和合并视觉节点，这里详细描述如何分割和合并节点。

###### 分割

给定文本节点，在每个交替迭代中，通过将每个文字节点分裂成两个来生成新的。首先假定文字节点连接了视觉节点，选择一对视觉节点，将文字节点中属于first视觉节点的实例作为正样本，属于second视觉节点的实例作为负样本，训练一个文本分类器。这样一共能选出个分割方案，根据训练分类器的正则化经验风险（也就是训练误差）来选择最佳的方案。

###### 合并

如果属于视觉聚类的实例在聚类上也具有很高的似然分数，那么这两个视觉聚类簇就可以被合并。

#### 实现细节

为了处理噪音数据，在视觉域创建了一些额外的节点。不同于在整个特征空间上进行聚类分割，本文的方法侧重于从数据子集中提取语义和视觉感知信息，使得这在各个域都具备高可信度。低可信度的数据被分配到噪声聚类簇，不参与最终的得分函数。

对于在某个域缺失信息的数据（比如只有图像没有文字的数据），其依旧通过其特征被分配到簇中，在很多case中，视觉聚类簇存在数据饥饿问题，那么这一操作就很有必要，能帮助学习到更好的视觉分类器。

## 结论

本文提出了在两个域做联合聚类的方法，其创新点在于舍弃了两个域之间存在一一对应的关系这一传统假定，宽松为存在一对多的关系，这一改动在当两个域粒度不同时非常有用。作者在文字域和视觉域上应用了这一方法，探索给定名词短语的语义和视觉感知信息。此方法不仅能正确找到语义和视觉感知信息，还建立了语义感知和视觉感知之间的映射关系。