关于论文《Robust continuous clustering》的学习报告

1、 均值偏移算法：均值漂移算法的核心思想是多个随机中心点向着密度最大的方向移动，最终能达到多个极大密度中心，也就是要找的聚类中心。均值漂移算法进行聚类的过程如下：首先在未标记的数据点中随机选择一个点作为中心 center，然后找到该点的距离不超过 threshold 的点记为集合 M，认定这些点属于聚类 C，同时把这些点属于这个类的概率加一；然后以 center 为中心点，计算 center 到集合 M 中所有点的向量，把这些向量相加，于是就形成了一个指向点数分布最多的方向的向量 v，然后把 center 往这个方向移动，移动的距离为|v|，重复上述红字标注的步骤直到移动的聚类小于设定的阈值(迭代收敛)，此时 M 的所有点都最终归到 C 类，此时判断是否有一个已存在的聚类中心与 C 的聚类中心很近，如果是则把这两类合并成一类，接着重复上述步骤直到所有点都被标记。RCC 目标函数第二项用的就是均值漂移算法的思想。

2、 AP（Affinity Propagation）聚类算法：算法一开始把所有的数据点假定为一个潜在的聚类中心，然后通过节点间的通信去找出最合适的聚类中心，并将其他节点划分到最合适的聚类中心中去。节点的通信过程是基于两个变量的，分别是吸引度和归属度。吸引度是 k 点作为 i 点的聚类中心的合适长度以及 i 点对 k 点的认可能力的综合考虑。；归属度是某个节点选择其他的一个节点作为它的聚类中心的合适程度。最终，把每个点的吸引度和归属度还有该点不适合作为其他点的聚类中心的程度做个均衡得到该点最为其他点的聚类中心的可能性。

3、 抗差估计（robust estimation）：（不理解）

4、 算法有很多很巧妙的地方：比如增加了一个辅助变量 L 把原始 RCC-Objective 转化成新的目标函数 C(U,L)，然后再利用目标函数的 biconvex 性质，固定一个变量从而单独求解另一个变量。而这里的 biconvex 性质是通过初始设定μ值得到的，在迭代的过程中，μ值逐渐自动减少，慢慢地把非凸的性质引入到目标函数中，最终求到一个与最优值很接近的解。（虽然不了解这样做的原理，但是觉得这种思想本身就很巧妙）

5、 稀疏编码：研究表明：初级视觉皮层 V1 区第四层有 5000 万个（相当于基函数），而负责视觉感知的视网膜和外侧膝状体的神经细胞只有 100 万个左右（理解为输出神经元）。因此，在神经网络学习中，可以通过稀疏编码的方式实现降维技术，把多维的冗余数据集变成含有重要信息的少维数据集。常用的学习方法为 LASSO，与岭回归方法相似，在回归算法中，LASSO 通过限定回归参数的大小，最终求解出一组对应于各个维度的回归系数，而 LASSO 算法的约束条件限定了某些回归参数最终被迫降至 0，得到的这组回归参数可以用来“理解”各个维度的重要性，因此最终可以只保留重要性较高的几个维度，从而在不失去重要信息的前提下实现降维效果。

6、 PCA：主成份分析法，PCA 技术通过计算出协方差矩阵的特征值和特征向量，再保留特征值较大的几个特征向量，得到的就是原始数据集中的主要成分，把原始数据集转换到这些特征向量构成的新空间中，从而实现降维效果。

7、 这篇论文学了较长的时间，最终也还有很多部分没看懂，没看懂的都是些公式以及公式的推导。比如说如何确定目标函数以及对目标函数进行转化，如何分析得出这是一个凸优化问题以及如何使用最小二乘法解决，以及各个变量的更新公式，所以看不懂的东西的一个很大的共同点就是利用数学方法来进行算法思想的验证和推导。由于不是数学专业出身的，希望能加强这些方面的训练。我想做聚类方面的研究，所以首先想系统地学习凸优化算法这些经常遇到的数学处理方法（网上找过很

多都没有很详细地介绍，只是简单的介绍而没有深入地了解），因此希望老师能推荐一些这方面的论文。