EM算法初探

周昊一[[1]](#footnote-1)\*

(南京大学 计算机科学与技术系, 南京 210093)

Introduce to EM Algorithm

Haoyi Zhou\*

(Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)

**Abstract**: Expectaion maximization algorithm is used to estimate the hidden parameters of a static model with the maximum likelihood. It can also be used to estimate the number of clusters.

**Key words**: data mining, clustering, expectation maximization

摘 要: 最大化期望算法是一种迭代算法,用来对模型进行最大似然估计.用最大化期望算法进行聚类时,还可以自动估计聚类的数量.

关键词: 数据挖掘，聚类，最大化期望

中图法分类号: TP301　　　文献标识码: A

# 问题分析

此次需要进行分析的数据共有1080项实例,每项实例有856个属性,没有训练数据,是一个无监督学习.

首先,我们无法得知每个属性具体代表的含义,因此也就无法准确得知最终的聚类数,因此所使用的算法需要能够估算聚类的个数.

其次,对所有的856个属性进行仔细观察后,发现对于每个实例来说,它的属性大部分取值都是0,非0的取值只存在于少部分的属性上,并且大部分取值是1,2出现的次数更少些,3总共就出现了24次,4总共出现了3次.似乎不同属性之间存在着某种关联,并且每个属性的取值又似乎呈现着某种分布(很像是正态分布).

最终,我决定使用基于模型的聚类方法来处理这些数据,而具体的算法则是选择了期望最大化算法(EM算法).

# 算法分析

## EM算法核心

EM算法的思想并不复杂,它将每个聚类都看成是一个概率分布,整个数据集合就可以看作是多个不同分布的混合.而算法的目标,就是估计这些概率分布的参数.

设X是观测到的数据的集合,Z是未观测到的数据的集合(隐含的数据),θ表示未知参数的向量.

定义似然函数.

EM算法的目标就是要最大化似然值.

算法分为两步,第一步是期望步(E),第二步是最大化步(M),具体描述如下:

1. 期望步:

在当前的参数估计下计算似然函数的期望值.

1. 最大化步:

最大化似然值,重新估计模型参数.

不停地重复进行E,M两个步骤,直到收敛.

## 算法的具体应用

对于此次的数据,我假设每个聚类对于每个属性都是一个正态分布.即聚类i上的属性j满足正态分布.则需要估计的参数.一个实例属于聚类的概率为每个属性在中该属性的分布下的概率的乘积.由于属性较多,概率又都较小,用乘法表示可能会导致精度问题,因此这里用概率的对数想加来表示.最终所属的聚类是概率最高的那一个.

## 聚类数的估计

用EM算法可以比较容易的估计聚类的个数,因为它有对模型进行评估的方法(似然函数).

具体的做法是:

1. 先假设只有一个聚类.
2. 将数据用10折交叉验证的方法变为10组.
3. 在每一组数据上用普通的EM算法进行聚类.
4. 求10组数据最终模型的平均似然值.
5. 如果平均似然值上升了,就将聚类数增加1,并回到(2)继续下一轮估计.

用此种方法可以估计出,本次实验数据的聚类数为7.

# 实验结果

## 模型参数

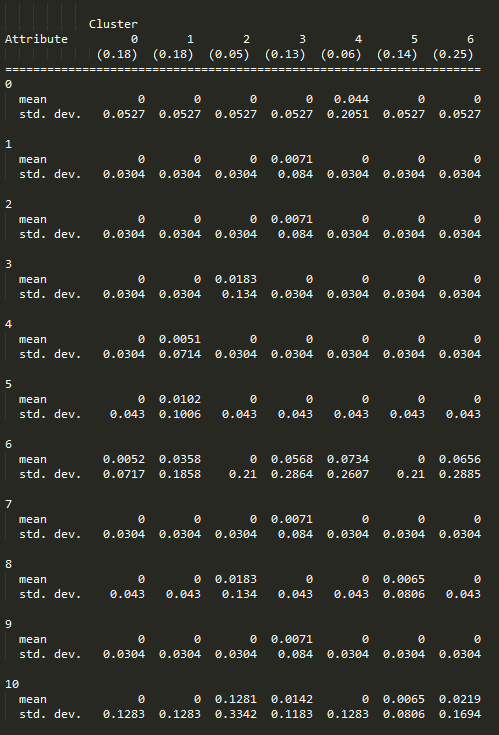


图1.估计出的模型参数(部分)

References:

1. Jiawei Han, Micheline Kamber 著范明 孟小峰 译 《数据挖掘——概念与技术(第二版)》机械工业出版社.
2. *Expectation–maximization algorithm*, http://<http://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm>.

1. \* 作者简介:周昊一，学号MF1233055。 [↑](#footnote-ref-1)