**说 明 书**

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━

**基于MLE和BIC的随机点模式参数估计方法**

**技术领域**

本发明属于模式识别技术领域，具体设计基于最大似然估计（MLE）和贝叶斯信息准则（BIC）的随机点模式参数估计方法。

**背景技术**

点模式是无序点或者特征的集合形式，在自然界中许多现象都可以用点模式的形式描述，而在许多数据分析问题中，点模式通常被称为包（多示例），用来描述对象的特征信息，针对点模式的学习问题本质上就是多示例学习问题。多示例学习作为有别于监督学习、无监督学习和强化学习的第4种机器学习框架，发展至今，已有很多实用的多示例学习算法被提出。目前主要的多示例学习方法大致可以分为三类：（1）基于示例空间的多示例学习算法；（2）基于包空间的多示例学习算法；（3）基于嵌入空间的多示例学习算法。基于示例空间的这类算法的核心是训练一个示例水平的分类器，使之能够区分来自正负示例的包，然后对于新的包，由示例的标签去推断包的标签；基于包空间的这类算法的核心思想是定义一个度量包之间距离的函数，然后把该距离函数嵌入标准的基于距离的分类器,由已知包的标签推断未知包的标签；基于嵌入空间的方法的核心思想是定义一个距离映射函数或者核函数，将每个包映射为一个单一的特征向量，用来描述和对应包相关的整体信息，这样原始的包空间就被映射为一个向量化的嵌入空间，并在这个空间进行分类器训练，把多示例问题转化为标准的监督学习问题。一般用似然函数描述的统计数据模型，是基于模型的数据分析方法的必要前提。但是针对点模式数据的机器学习算法研究中，关于基于统计点模式模型的算法并没有得到足够的关注。

**发明内容**

本发明的目的就是提供一种基于最大似然估计（MLE）和贝叶斯信息准则（BIC）的随机点模式参数估计方法。

针对基于模型的点模式数据分类问题，本发明提出了一种结合模型复杂度估计的点模式模型参数学习方法。该方法在基于随机点模式模型的框架下，构建多个复杂度不同的随机点模式模型，通过最大似然估计算法（MLE）和吉布斯采样算法分别对每个模型的基数分布参数和特征分布参数进行估计，进一步计算每个模型关于训练数据的Bayes信息准则（BIC）指标，确定备选模型中最优模型。

本发明方法具体时间：

1.基于MLE和BIC的随机点模式参数估计方法，其特征在于，该方法具体是：

步骤（1）.构建随机点模式模型：

；表示第个随机点模式的观测数据，其中，表示随机点模式的观测数据总个数；表示中元素的个数；表示点模式模型的基数分布，表示基数分布的模型参数；表示点模式模型的特征分布，表示特征分布的模型参数；，表示中的元素；为超参数，用于取消各元素间的单位不统一。

步骤（2）.构建观测数据集关于点模式模型参数的似然函数：

；表示观测数据集关于模型参数的似然函数。一般情况下，模型基数分布为泊松分布，参数,表示泊松系数；模型特征分布为高斯混合分布，参数；,,分别表示第个分布元的混合权重，均值和协方差。

步骤（3）.对步骤（2）的似然函数最大化可以转化为分别关于基数分布参数和特征分布参数进行极大似然估计，，表示基数分布参数的估计值；，表示特征分布参数的估计值。

步骤（4）.对基数分布参数进行极大似然估计，首先构建关于基数分布参数的对数似然函数：；由对数似然函数关于基数分布参数求导为零得，在泊松分布的情况下，，其中，为基数分布参数的估计值。

步骤（5）.明确隐变量，写出完全数据关于特征分布参数的对数似然函数：

反映点模式观测数据中元素来自第个分布元的数据是未知的，,其中,;,以隐变量表示，是的随机变量。那么完全数据为，其中，；于是可以写出完全数据关于特征分布参数的对数似然函数：，式中，；表示第个高斯分布函数。

步骤（6）.EM算法的E步，确定函数：

,其中，，表示分布元对观测数据的响应度。

步骤（7）.确定EM算法的M步

，用及表示的各参数。求只需要将函数分别对求偏导数并令其为0，即可得到；求是在条件下求偏导数并令其为0得到的。结果如下：；；。

步骤（8）.计算估计模型的贝叶斯信息值：

，其中，，表示模型特征分布的分布元个数，通过贝叶斯信息值作为模型评价标准。

2.如权利要求1所述的基于MLE和BIC的随机点模式参数估计方法，其特征在于，方法具体流程是：

设置特征分布元个数的范围；

根据步骤（1）、（2）构建模型似然函数；

根据步骤（3）、（4），学习得泊松分布参数；

初始化特征分布参数；

根据步骤（5），依据当前模型参数，计算分布元对观测数据的响应度；

根据步骤（6）、（7），计算新一轮的模型参数及，得到；

重复，直到收敛，得到特征分布参数；

保存模型参数，；

重复，直到超出预设的范围；

根据步骤（7），选取贝叶斯信息值最小的备选模型最优最后的输出模型

本发明的有益效果包括：为了准确估计点模式模型的复杂度和参数，本发明提出了一种结合模型复杂度估计的点模式模型参数学习方法。该方法在基于随机点模式模型的框架下，构建多个复杂度不同的随机点模式模型，通过最大似然估计算法（MLE）和吉布斯采样算法分别对每个模型的基数分布参数和特征分布参数进行估计，进一步计算每个模型关于训练数据的Bayes信息准则（BIC）指标，确定备选模型中最优模型。

**具体实施方式**

1.基于MLE和BIC的随机点模式参数估计方法，其特征在于，该方法具体是：

步骤（1）.构建随机点模式模型：

；表示第个随机点模式的观测数据，其中，表示随机点模式的观测数据总个数；表示中元素的个数；表示点模式模型的基数分布，表示基数分布的模型参数；表示点模式模型的特征分布，表示特征分布的模型参数；，表示中的元素；为超参数，用于取消各元素间的单位不统一。其中图（一）、图（二）分别为观测数据的基数分布信息和特征分布信息。

步骤（2）.构建观测数据集关于点模式模型参数的似然函数：

；表示观测数据集关于模型参数的似然函数。一般情况下，模型基数分布为泊松分布，参数,表示泊松系数；模型特征分布为高斯混合分布，参数；,,分别表示第个分布元的混合权重，均值和协方差。

步骤（3）.对步骤（2）的似然函数最大化可以转化为分别关于基数分布参数和特征分布参数进行极大似然估计，，表示基数分布参数的估计值；，表示特征分布参数的估计值。

步骤（4）.对基数分布参数进行极大似然估计，首先构建关于基数分布参数的对数似然函数：；由对数似然函数关于基数分布参数求导为零得，在泊松分布的情况下，，其中，为基数分布参数的估计值。

步骤（5）.明确隐变量，写出完全数据关于特征分布参数的对数似然函数：

反映点模式观测数据中元素来自第个分布元的数据是未知的，,其中,;,以隐变量表示，是的随机变量。那么完全数据为，其中，；于是可以写出完全数据关于特征分布参数的对数似然函数：，式中，；表示第个高斯分布函数。

步骤（6）.EM算法的E步，确定函数：

,其中，，表示分布元对观测数据的响应度。

步骤（7）.确定EM算法的M步

，用及表示的各参数。求只需要将函数分别对求偏导数并令其为0，即可得到；求是在条件下求偏导数并令其为0得到的。结果如下：；；。

步骤（8）.计算估计模型的贝叶斯信息值：

，其中，，表示模型特征分布的分布元个数，通过贝叶斯信息值作为模型评价标准。图（三）为贝叶斯值关于模型复杂度的曲线，图（四）为模型特征分布学习结果，图（五）为模型基数分布学习结果，表一为真实模型的参数，表二为输出模型的参数。

2.如权利要求1所述的基于MLE和BIC的随机点模式参数估计方法，其特征在于，方法具体流程是：

设置特征分布元个数的范围；

根据步骤（1）、（2）构建模型似然函数；

根据步骤（3）、（4），学习得泊松分布参数；

初始化特征分布参数；

根据步骤（5），依据当前模型参数，计算分布元对观测数据的响应度；

根据步骤（6）、（7），计算新一轮的模型参数及，得到；

重复，直到收敛，得到特征分布参数；

保存模型参数，；

重复，直到超出预设的范围；

根据步骤（7），选取贝叶斯信息值最小的备选模型最优最后的输出模型。