# 基于MLE和BIC的随机点模式模型学习算法

杨豪杰，刘伟峰

杭州电子科技大学自动化学院，杭州，310018

**摘 要：**针对基于模型的点模式数据分类问题，本文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。仿真实验首先验证了本文方法所得模型对训练集数据具有良好的数据拟合能力，和对测试数据具有优秀的泛化能力；并在对点模式数据分类测试中，取得了不俗的表现。

**关 键 词：点模式；Gibbs采样；BIC准则；多示例学习；分类问题**

**Abstract: Aiming at the problem of model-based point pattern data classification, this paper proposes a learning method combined with model complexity estimation. Under the framework of point pattern model, this method constructs multiple point pattern models with different complexity, and estimates the parameters of each model through maximum likelihood estimation algorithm and Gibbs sampling algorithm, and further calculations The distribution parameters are estimated. The Bayes Information Criterion index of each model on the training data determines the optimal model among the candidate models. The simulation experiment verifies that the optimal model has good data fitting ability to the training set data and excellent generalization ability to the test data; and it has achieved good performance in the point pattern data classification test.**

**Key words: point pattern; Gibbs sampling; BIC criterion; multi-instance learning; classification**

## 1 引言

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。该问题起源对药物分子的活性预测[1]，通过分析一组已知的药物分子，预测某类药物分子是否适合具备特定的药物活性，而问题的主要难点在于每个药物分子存在大量的低能形状（同分异构体），其中只有一种或者几种特定的低能形状适合制药，并且专家对已知的药物分子并不确定具体哪些形状发挥决定性作用。因此Dietterich等[1]将每一个药物分子看做一个包，适合制药的分子为正包，药物分子的每一种低能形状为包中的一个示例，使分子适合制药的低能形状视为正例，并把这类学习问题称为多示例学习问题。自此以后，有关多示例学习问题的研究受到广泛关注，目前多示例学习方法已经广泛应用于图像分类[2-4]、文本分类[5-6]、图像检索[7-9]、医疗图像辅助识别[10-11]等领域。

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[12]，发展至今，已有很多实用的多示例学习算法被提出。根据Amores[13]等提出的观点,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。常见的算法有mi-SVM[6]、MIBoosting[14]、Clustering MIL[15]、SMILE[16]。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。常见的算法有Citation-kNN[17]、MI-SVM[6]、MI-Kernel[18]。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。常见的算法有DD-SVM[19]、Simple MI[20]、MILES[21]、miFV[22]。

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，关于基于统计的点模式模型并没有得到足够的关注。传统的朴素贝叶斯模型（Naïve Bayesian Model）只能描述点模式数据的特征信息，而无法描述点模式的示例数目信息。于是Vo等[23]借助点过程理论和随机有限集的思想提出了一种基于模型的点模式学习框架。本文在此框架的基础上，针对基于模型的点模式数据分类问题，研究多示例学习与其他机器学习方法之间的联系，利用已有的监督学习、无监督学习方法，对点模式模型参数学习算法进行扩展，基本思想为使用最大似然估计对基数分布参数进行估计，Gibbs参数学习特征分布的参数，借助贝叶斯信息准则，最终确定最优模型复杂度下的点模式模型的参数。

## 2 背景与问题描述

在基于模型的多示例分类问题中，由于朴素贝叶斯模型不足以充分描述点模式数据的信息，导致类间特征相似度较高时，分类性能明显下降。因此文献[23]借助点过程理论，提出一种包含点模式基数分布信息的随机点模式模型，结合点模式的基数信息和特征信息，有效弥补了朴素贝叶斯模型的不足。在此基础上，参考文献[24]中对有限混合模型分布元估计的方法，本文提出一种结合模型复杂度优化指标的参数学习方法，同时保证模型的数据拟合能力和泛化能力。

### 2.1 随机点模式模型

通常情况下点过程的概率密度函数可能不存在，为了保证点过程的概率密度函数有效性，本文建模只针对简单的有限点过程，使得点过程生成的点模式可以等价于一个随机有限集，随机有限集的概率函数可以表示为[23]

 （2-1）

其中表示点模式的基数分布；是一个对称函数，表示关于的联合概率分布；是特征空间上的超参数，的作用是将概率密度函数中各点的单位消除，防止各个点之间单位不匹配。

在点模式的特征点之间引入独立性假设，则式（2-1）的模型可以转化为模型[23]

 （2-2）

其中表示点集中元素的个数；表示点模式在特征空间上的特征分布；，且。当模型的基数分布为泊松分布时，且泊松分布的参数为，对应的模型即为泊松点过程模型

 （2-3）

### 2.2 吉布斯采样

吉布斯采样（Gibbs Sample）用于多元变量联合分布的抽样和估计，可以认为是Metropolis-Hastings（MH）算法的特殊情况，是更简单、使用更广泛的马尔科夫蒙特卡洛法（MCMC）。采样过程通过从联合概率分布定义满条件概率分布，依次对满条件概率分布进行抽样，从而产生样本的序列。可以证明这样的过程是在一个马尔科夫链上随机游走，每一个样本对应着马尔科夫链的状态，平稳分布就是目标的联合分布。

假设目标分布是多元概率分布，其中为维随机变量。吉布斯抽样从一个初始分布样本出发，不断进行迭代，每一次迭代得到联合分布的一个样本。迭代结束后，最终得到的样本序列。

每次迭代中，依次对个随机变量中的一维变量进行随机抽样。假设对第个样本的第维进行随机抽样，那么抽样的分布是满条件概率分布，其中，表示第次迭代中，除第维变量的其他条件变量。吉布斯采样的算法如下[25]：

（1）初始化参数，给出一个初始样本

（2）对进行循环执行

假设第次迭代结束得到样本，则第次迭代进行如下操作：



通过第次迭代得到样本。

（3）循环结束得到样本序列。

### 2.3 贝叶斯信息准则

大部分模型参数估计问题采用似然函数作为目标函数，以极大似然估计方法求解形式已知，参数未知的参数估计问题。虽然当训练数据足够充足的情况下，可以不断提高模型精度，但必须以不断提高提高模型的复杂度为代价，随之而来就是机器学习中过拟合的问题。

贝叶斯信息准则（BIC）由Schwarz[26]在1978年提出，与赤池信息准则（AIC）[27]相似，是衡量统计模型拟合优良性的一种标准，建立在熵的概念，提供了衡量权衡复杂度和拟合数据优良性的标准，借助信息论提出确定模型阶次（分布元个数）的方法，常常用于模型的选择。AIC与BIC的定义如下：

 （2-4）

 （2-5）

其中表示模型参数个数，是样本数据关于模型参数的似然函数，为样本的数量。AIC与BIC的公式都有两项组成，前半部分是惩罚项，后半部分是似然函数项，当两个模型之间存在较大差异时，AIC与BIC值主要受似然函数项影响；当两个模型差异较小的情况下，似然函数项差异对AIC与BIC值影响不明显，主要由惩罚项起作用，即受模型复杂度影响，因为AIC与BIC通过最小值来选择模型，从而会选择参数个数少的模型。一般而言，当模型复杂度提高，即增大时，似然函数也会随机增大，导致AIC与BIC值减下，但过大时，似然函数增速减缓，反而导致AIC与BIC值增大。BIC相比AIC引入样本数量的概念，AIC在大量样本数据时通常会因为似然函数值过大，淹没了模型复杂度的影响，导致AIC作为判断标准的结果不理想，而BIC相比AIC的惩罚项会随着样本数据量增加而加大，导致BIC更倾向于选择模型复杂度低的相对简单的模型。

# 3 结合模型复杂度估计的模型参数学习算法

学习点过程模型的计算高效算法很重要，因为机器学习与空间统计应用相比通常涉及大数据集，由于学习一般的点过程模型在计算上是非常困难的，而泊松点过程模型通过忽略点之间的相关性，在模型的通用性和模型参数的学习难度之间取得了一个良好的平衡。

泊松点过程模型由它的基数分布和特征分布共同决定，模型的参数表示形式如下

 （3-1）

其中和分别由参数和所决定的基数分布和特征分布，的函数形式为泊松分布，由于高斯混合模型在数据拟合方面有着不俗的表现，所以的函数形式为高斯混合分布，对学习泊松点过程模型的学习过程相当于通过训练集估计参数。

从泊松点过程模型的似然函数形式中我们不难发现，对整体模型参数极大似然估计等价于分别对基数分布参数和特征分布参数进行极大似然估计，估计过程如下：

假设，……，是个独立且由一个泊松点过程模型生成的样本数据集，因此样本数据集关于参数的似然函数如下：

 （3-2）

从似然函数的形式中可以得出，最大化等价于分别最大化似然函数的第二与第三部分，即

 （3-3）

 （3-4）

### 3.1 基数分布的参数估计

基数分布关于参数的似然函数如下：

 （3-5）

为了便于分析与计算，我们定义基数分布关于参数的对数似然函数，

 （3-6）

 （3-7）

容易证明，使对数函数似然值最大的值也是似然函数最大。对关于求导为零得



 （3-8）

### 3.2 特征分布的参数估计

泊松点过程模型的特征分布为有限高斯混合模型可以表示为如下形式：

 （3-9）

其中，维度为维，令，表示集合中元素个数；是混合权重且，；是高斯分布密度，，

 （3-10）

称为第个分模型；特征分布的参数。那么特征分布似然函数可以表示为：

 （3-11）

根据式（3-11），高斯混合模型的后验分布可以描述为：

 （3-12）

 （3-13）

其中是正则常量。对于每一个训练样本中的元素,需要定义一个维的指示变量，表示该元素由哪个分布元产生，称为缺失变量，表示形式如下：



 （3-14）

#### 3.2.1 高斯混合模型的参数先验分布

实现公式（3-12），优先需要考虑如何获取先验分布，在高斯混合分布的情况下，模型参数为，那么先验分布如下：

 （3-15）

由于混合权重反应各分布元所生成的观测数据占整个观测数据的比重，如果各分布元的比重位置，则关于混合权重的先验分布可以认为是等比重的Dirichlet分布，

 （3-16）

均值的先验分布采用高斯分布，

 （3-17）

表示第个分布元的统计均值，协方差的逆矩阵的先验分布一般符合Wishart分布，

 （3-18）

其中表示自由度；表示正定矩阵，一般取单位矩阵。

#### 3.2.2 高斯混合模型的参数后验分布

结合先验分布中式（3-15）、（3-16）、（3-17）、（3-18），以及特征分布似然函数式（3-11），根据Bayes定理更新公式（3-12）得：

 （3-19）

由式（3-19）可得各参数的后验分布：后验混合权重符合Dirichlet分布，后验均值符合正态分布，后验方差符合Wishart分布[28]，参数的具体分析如下

（1）混合权重，混合权重满足下面的Dirichlet分布：

 （3-20）

其中是大于零的常数，是属于第个分布元的观测数据的个数。

（2）缺失变量，缺失变量可以根据如下的Bayes公式估计：

 （3-21）

 （3-22）

（3）协方差，协方差的逆服从Wishart分布[28]

 （3-23）

 （3-24）

其中，为常数，且；起调节作用；为二阶矩值。

（4）均值，满足参数为的高斯分布，我们可以根据式中进行采样获取均值

 （3-25）

 （3-26）

#### 3.2.3 Gibbs采样的模型参数学习算法

基于以上的式（3-15）～（3-26），我们可以通过Gibbs采样算法，迭代产生高斯混合模型的参数。算法流程如下：



### 3.3 基于BIC的最优模型选择

根据观测数据在特征空间上的分布，通过特征点的分布情况大致估计分布元个数范围，设置分布元个数在一个合理且较小范围内可以有效减少算法的空间和时间复杂度；当然可以将设置为一个较大的范围，防止因为主观判断，导致对分布元个数范围的错误估计，无法取得BIC值最小时所对应的分布元个数。

备选模型的BIC计算：

 （3-27）

其中表示特征分布为个分布元的备选泊松点过程模型对应的模型参数；。参考曲线，根据下式获得最优模型参数

 （3-28）

## 4 仿真及实验结果分析

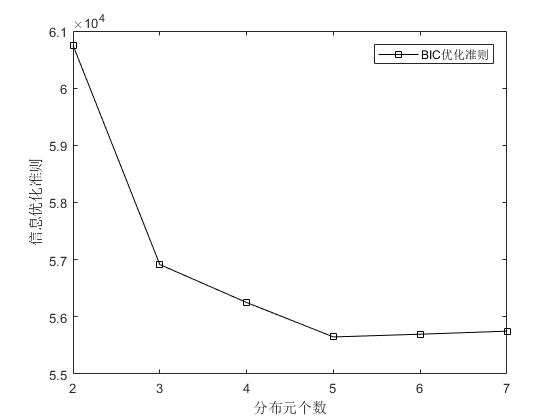
本节由两部分实验组成，第一部分验证由本文提出算法所得模型的数据拟合能力和泛化能力；第二部分比较泊松点过程模型和朴素贝叶斯模型在点模式数据分类问题上的性能差异，并比较点过程模型分布元（模型复杂度）差异对分类效果的影响。说明本文学习算法的可靠性。

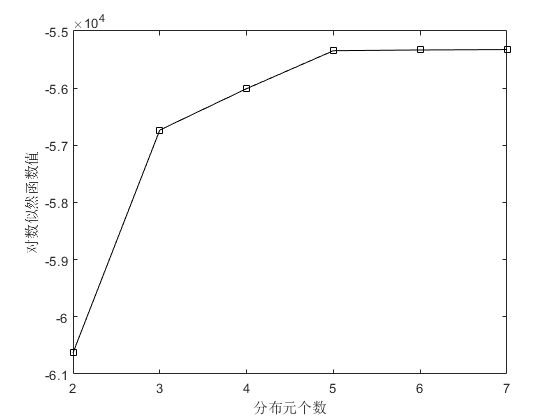
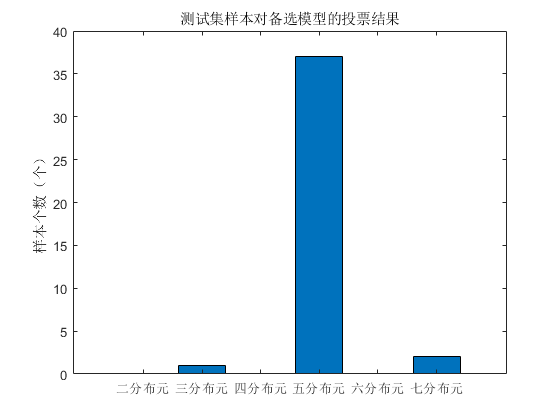
### 4.1 关于最优模型的数据拟合和泛化能力分析

假设某个点模式模型的基数分布服从泊松分布，特征分布服从混合高斯分布，参数为;;,,;,,。通过该模型生成140组点模式数据样本，将其中的100组数据样本作为训练集，剩余的40组样本作为测试集。

图（一）为训练集数据样本的真实特征分布，我们通过本文的参数学习算法对100组训练数据进行参数学习，分布元个数从2个递增至7个分别进行学习得到6个备选模型，并且以Bayes信息准则作为最优模型选择依据。根据图（二）的BIC曲线，分布元为5时取得最小BIC值，选取备选模型中特征分布为5元高斯混合的泊松点过程模型为最优模型。

分析图（三）的曲线可得，模型复杂度达到一定水平（分布元个数大于等于5时），模型的对数似然函数随着模型复杂度的提升而增大的现象变得不明显，说明模型复杂度达到一定程度即可保证模型对训练数据集的拟合能力，本文算法所得的最优模型通过图（三）可以说明对训练数据集具有良好的数据拟合能力。

  图（一） 图（二）

图（三） 图（四）

我们再通过40组测试集数据样本对BIC信息准则获得的最优模型进行泛化能力验证，首先构建6个备选模型的似然函数。将每个测试集样本分别代入6个似然函数中获得6个似然值，选择最大似然值所对应的备选模型作为投票对象。通过图（四）测试集的投票结果表明，40组测试数据中有37组在特征分布为5元高斯混合的泊松点过程模型（算法所得最优模型）处取到最值，说明该模型最符合测试集分布，验证了算法所得的最优模型为备选模型中对测试集数据泛化能力最强。

### 4.2 点模式数据集分类实验

存在三类点模式数据分别由以下三个随机点模式模型生成。第一类样本模型的基数分布服从泊松分布，参数为，特征分布为高斯混合模型，参数为；，,;, , 。第二类样本模型为基数分布服从泊松分布，参数为，特征分布为高斯混合模型，参数为；，,;, , 。第三类样本模型为基数分布服从泊松分布，参数为，特征分布为高斯混合模型，参数为, ；，,;, , 。每一类生成140个样本，其中100个用作模型学习，40个用作测试学习效果；从图（八）中可以看出，类一与类二在基数信息上重叠，类三可以借助基数分布信息与其他两类作区分；从图（九）中可以看出，类二和类三在特征分布信息上重叠，类一可以通过特征信息与其他两类作区分。从图（十）的分类结果中显示结合基数信息的Poisson模型相对于单纯使用特征分布信息的NB模型在分类任务中效果更好。图（十一）中，Test One的三类样本模型学习过程中，并没有进行对模型复杂度进行搜索，而是通过对训练集数据的特征分布直观判断分布元个数的方法；Test Two完全遵循本文提出的模型选择逻辑，当主观判断与基于BIC的最优准则不符是，即两次实验中描述三类样本的模型的分布元个数（模型复杂度）存在差异，从两次分类结果中可以看出，通过本文提出的模型学习方法所得到的模型具有更好的数据拟合能力和模型泛化能力，最大程度减轻主观判断对模型选择的影响，从而保证良好的分类性能。

图表, 直方图

描述已自动生成 图表, 散点图

描述已自动生成

图（八） 图（九）

图表, 条形图

描述已自动生成 图表, 条形图

描述已自动生成

图（十） 图（十一）

### 5 总结

针对基于模型的点模式数据分类问题，本文提出了一种结合模型复杂度估计的参数学习方法。算法首先构建不同复杂度的随机点模式模型，使用MLE估计基数分布参数，再使用Gibbs参数采样学习，分别学习点模式数据的包水平信息和示例水平信息，再结合BIC曲线与最优准则，从备选模型中选出最优模型。不过本文算法在对特征分布的分布元个数范围估计时，依然需要对训练数据进行主观判断；且随机点过程模型并不能描述示例间的结构性，基于模型的多示例算法研究可以从以上两方面进行深入研究

## 参考文献

1. Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles[J]. Artificial intelligence, 1997, 89(1-2): 31-71.
2. Csurka G, Dance C, Fan L, et al. Visual categorization with bags of keypoints[C]//Workshop on statistical learning in computer vision, ECCV. 2004, 1(1-22): 1-2.
3. Wang X, Wang B, Bai X, et al. Max-margin multiple-instance dictionary learning[C]//International conference on machine learning. 2013: 846-854.
4. Ramesh B, Xiang C, Lee T H. Shape classification using invariant features and contextual information in the bag-of-words model[J]. Pattern Recognition, 2015, 48(3): 894-906.
5. McCallum A, Nigam K. A comparison of event models for naive bayes text classification[C]//AAAI-98 workshop on learning for text categorization. 1998, 752(1): 41-48.
6. Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning[J]. Advances in neural information processing systems, 2002, 15: 577-584.
7. Xu Y Y, Shih C H. Content based Image retrieval using multiple instance decision based neural networks[C]//2012 IEEE International Conference on Computational Intelligence and Cybernetics (CyberneticsCom). IEEE, 2012: 175-179.
8. Chiang J Y, Cheng S R. Multiple-instance content-based image retrieval employing isometric embedded similarity measure[J]. Pattern Recognition, 2009, 42(1): 158-166.
9. Zhang C, Platt J, Viola P. Multiple instance boosting for object detection[J]. Advances in neural information processing systems, 2005, 18: 1417-1424.
10. Li W, Vasconcelos N. Multiple instance learning for soft bags via top instances[C]//Proceedings of the ieee conference on computer vision and pattern recognition. 2015: 4277-4285.
11. Kraus O Z, Ba J L, Frey B J. Classifying and segmenting microscopy images with deep multiple instance learning[J]. Bioinformatics, 2016, 32(12): i52-i59.
12. Zhou Z H. Multi-instance learning: A survey[J]. Department of Computer Science & Technology, Nanjing University, Tech. Rep, 2004, 1.
13. Amores J. Multiple instance classification: Review, taxonomy and comparative study[J]. Artificial intelligence, 2013, 201: 81-105.
14. Xu X, Frank E. Logistic regression and boosting for labeled bags of instances[C]//Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, Heidelberg, 2004: 272-281.
15. Tax D M J, Hendriks E, Valstar M F, et al. The detection of concept frames using clustering multi-instance learning[C]//2010 20th International Conference on Pattern Recognition. IEEE, 2010: 2917-2920.
16. Xiao Y, Liu B, Hao Z, et al. A similarity-based classification framework for multiple-instance learning[J]. IEEE transactions on cybernetics, 2013, 44(4): 500-515.
17. Wang J, Zucker J D. Solving multiple-instance problem: A lazy learning approach[J]. 2000.
18. Gärtner T, Flach P A, Kowalczyk A, et al. Multi-instance kernels[C]//ICML. 2002, 2(3): 7.
19. Chen Y, Wang J Z. Image categorization by learning and reasoning with regions[J]. Journal of machine learning Research, 2004, 5(Aug): 913-939.
20. Dong L. A comparison of multi-instance learning algorithms[D]. The University of Waikato, 2006.
21. Chen Y, Bi J, Wang J Z. MILES: Multiple-instance learning via embedded instance selection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12): 1931-1947.
22. Carbonneau M A, Granger E, Raymond A J, et al. Robust multiple-instance learning ensembles using random subspace instance selection[J]. Pattern recognition, 2016, 58: 83-99.
23. Vo B N, Dam N, Phung D, et al. Model-based learning for point pattern data[J]. Pattern Recognition, 2018, 84: 136-151.
24. 刘伟峰, 杨爱兰. 基于 BIC 准则和 Gibbs 采样的有限混合模型无监督学习算法[J]. 电子学报, 2011.
25. 李航. 统计学习方法[M]. Qing hua da xue chu ban she, 2012.
26. Michalak K P. Estimating correlation dimension of high-dimensional signals-quick algorithm[J]. AIP Advances, 2018, 8(10): 105201.
27. Akaike H. Information theory and an extension of the maximum likelihood principle[M]//Selected papers of hirotugu akaike. Springer, New York, NY, 1998: 199-213.
28. Diebolt J, Robert C P. Estimation of finite mixture distributions through Bayesian sampling[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1994, 56(2): 363-375.