# 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：**针对基于朴素贝叶斯模型在点模式数据分类表现不佳问题，本文提出了一种基于MLE和BIC的随机点模式模型参数学习算法。该方法基于随机点模式模型框架，根据训练集构建多个复杂度不同的随机点模式模型，将点模式模型参数的极大似然估计（MLE）转化为对模型基数分布和特征分布做参数估计，并在对特征分布参数估计时采用结合模糊C均值聚类（FCM）和EM的方法，最后计算每个模型关于训练集的Bayes信息准则（BIC）指标，作为模型评价指标确定备选模型中最佳模型。仿真实验通过两组点模式数据分类实验，结果验证了随机点模式模型相对于传统模型的优势和本文算法良好的学习性能。

**关 键 词：点模式；FCM；EM；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,**

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

## 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，针对基于朴素贝叶斯模型在点模式数据分类表现不佳问题，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，完善点模式模型参数学习算法。基本思想为使用最大似然估计学习基数分布参数，结合FCM和EM的方法学习特征分布参数，借助贝叶斯信息准则作为判断标准，确定数据集对应的最佳模型。

## 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，同时保证训练所得随机点模式模型的数据拟合能力和泛化能力，一定程度上提升最终所得模型对于点模式数据的分类性能。

### 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 EM算法

最大期望算法是一种通过迭代优化估计模型参数的方法，由Dempster等[25]于1977年正式提出，常常用于概率模型参数关于不完全数据的极大似然估计。实际情况中，由于数据观测手段存在一定限制，必然会导致观测数据存在一定程度上的信息缺失，成为不完全数据。EM算法引入合适的隐变量，结合原始的观测数据组成完全数据，并极大化概率模型参数关于完全数据的似然函数的条件期望，由此便能得到模型参数的极大似然估计。

EM算法的每次迭代由E步（求期望）和M步（求极大）组成。一般，用表示观测数据，用表示隐变量数据，和组成完全数据；其中，表示模型参数， 那么对于模型参数的估计流程如下：

1. 设置模型参数的初值，开始迭代；
2. E步：用表示第次迭代所得的模型参数估计值，第次E步，计算

 （2-4）

这里，表示观测数据和隐变量数据关于模型参数的联合概率分布，表示观测数据和当前模型参数下隐变量数据的条件概率分布；

1. M步：通过对求极大化，确定第次迭代的模型参数估计值

 （2-5）

1. 给定停止迭代条件，重复步骤2）和3），直到满足

 （2-6）

或达到指定迭代次数，则停止迭代，一般一般为较小的正数。

EM算法提供了一种近似计算含有隐变量概率模型的极大似然估计的方法，算法所得的估计序列的收敛性；以及收敛到全局最大值或者局部极大值已由文献[28]证明。

### 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。虽然当训练数据充足的情况下，可以通过不断提高模型精度的方式增大目标函数，但必须以牺牲模型复杂度为代价，随之而来的就是机器学习中过拟合问题。根据信息论准则（ITC）作为模型的判断标准是选择最佳模型的有效手段，也是衡量模型拟合程度的一种标准，因此广泛应用于模型选择。目前，科研人员已经提出多种用于最佳模型选择的ITC准则，主要包括赤池信息准则（AIC）[27]，贝叶斯信息准则（BIC）[26]等，

贝叶斯信息准则由Schwarz[26]在1978年提出，与赤池信息准则相似，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，两者都是从降低模型复杂度的角度出发实现模型的拟合，但在定义方式有存在些许差异。AIC与BIC的定义如下：

 （2-7）

 （2-8）

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均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，且；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）

#### 3.2.1 高斯混合模型参数估计的EM算法

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，称为隐变量，表示形式如下：



 （3-12）

EM算法通过迭代生成参数的估计序列来最大化特征分布似然函数（3-11），具体步骤如下：

E步：计算，定义如下

 （3-13）

记，表示在当前模型参数下第个元素来自第个分布元的概率，称为分布元对元素的响应度。将代入式（3-13）得

 （3-14）

M步：求解关于的极大值，即求新一轮迭代的模型参数

 （3-15）

求，只要对式（3-15）分布关于，求偏导并令其为零，即可得；求需要在条件下关于求偏导为零得。结果如下：

 （3-16）

 （3-17）

 （3-18）

重复E步和M步，直到收敛。

#### 3.2.2 FCM算法确定EM算法的初始参数

EM算法只能保证参数估计序列收敛到对数似然函数的稳定点，并不能保证收敛到极大值点。所以在实际应用中，初值的选择显得尤为重要，合适的初始值不仅可以缩短算法收敛时间，还能避免收敛到局部最优点。基于此思想，使用模糊C均值聚类（FCM）算法确定EM算法的初始值，并且使用势函数法[30]来初始化聚类中心，提升算法性能。

特征空间中的观测数据集以GMM的分布元个数作为聚类中心个数，FCM算法的目标函数为：

 （3-19）

其中，隶属度为样本属于第类的概率，为隶属度矩阵；表示第类的聚类中心，为聚类中心向量；表示模糊度的控制参数，其值越大算法的模糊度越高，文献[29]研究表明的最佳取值范围应在，一般应用中取2。

FCM算法通过迭代实现，具体步骤如下：

Step1:给定的值，使用势函数法初始化类中心向量，令隶属度矩阵；

Step2:最小化目标函数，，更新隶属度矩阵:

 （3-20）

Step3:最小化目标函数，，更新聚类中心向量:

 （3-21）

Step4:重复执行步骤Step2-3，直到目标函数收敛。

通过FCM算法的聚类结果计算GMM的初始参数，结果如下：

 （3-22）

 （3-23）

 （3-24）

将结算结果作为3.2.1节EM算法的初始值，便可学习得特征分布的模型参数。

### 3.3 基于BIC的最佳模型选择

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区间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区间，其中，许多研究者经常使用的经验规则有或者[30] ，防止因为主观判断误差作为先验信息，导致无法取得BIC值最小时所对应的分布元个数。对比文献[23]的参数学习方法，给定一个形式确定的模型先验信息，本节要求的两种模型先验信息相对模糊，明显具有更好的信息容错率，且更容易获取。

备选模型的BIC计算：

 （3-25）

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

 （3-26）

最终输出通过训练集数据所得的最佳模型。

## 4 仿真及实验结果分析

本节由两组点模式数据分类实验构成，第一组实验采用虚拟数据集，由计算机随机生成，第二组实验则采用纹理数据集[30]。训练阶段，使用本文提出方法训练泊松点过程模式和朴素贝叶斯模型，两种模型在特征分布上是一致的，并且为了简化测试阶段，假设每个数据集的各类先验概率符合均分分布。

### 4.1 虚拟数据集分类实验

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个用作模型参数学习，给定特征参数元区间为，剩余40个用作分类任务测试学习效果。



图1. 数据集的分布信息

从图1（a）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1（b）中可以看出，类二和类三在特征分布上重叠，类一可以通过特征信息与其他两类作区分。Poisson模型的学习结果如图2（a）、（b）所示，BIC准则，从图2（c）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模型在分类任务中的确有着更加出色的分类表现。



图2. 模型的学习结果

### 4.2 纹理数据集分类实验

选取纹理数据集[30]中的三类：“T14 brick1”，“T15 brick2”，“T20 upholstery”，进行分类实验。其中，每一类均包含40张图片，首先采用尺度不变特征变换（SIFT）算法提取每张图片的关键点，如图3所示，获得每张图片对应的128维点模式数据，然后使用主成分分析法（PCA）进行特征提取，将128维的特征降成2维，得到2维的点模式数据。

采取4折交叉验证对分类性能进行验证，每次测试中，每类使用30张图像的点模式数据作为训练集，对泊松点过程模型和NB模型进行训练，

## 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，再使用结合FCM和EM学习基数分布参数，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曲线的最优准则，从备选模型中选出最佳模型。该方法对比文献[23]，需要给定一个形式确定的先验特征分布，一定程度上放宽了特征先验信息的条件，尽可能避免因先验信息存在误差，对分类性能带来的负面影响。然而，在实际应用中，如何合理地选择先验信息变得至关重要；且随机点过程模型引入了独立性假设并不能描述示例间的结构关系，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. Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1): 1-22.Michalak K P. Estimating correlation dimension of high-dimensional signals-quick algorithm[J]. AIP Advances, 2018, 8(10): 105201.
26. 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.
27. 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.
28. McLachlan G J, Krishnan T. The EM algorithm and extensions[M]. John Wiley & Sons, 2007.
29. Pal N R, Bezdek J C. On cluster validity for the fuzzy c-means model[J]. IEEE Transactions on Fuzzy systems, 1995, 3(3): 370-379.