基于贝叶斯原理的多维 Spike Train 分类预测模型

樊一娜[①] 郎 波*[②] 危 辉[②]
[①](青海大学电力学院 西宁 810016)
[②](复旦大学认知算法模型实验室 上海 201203)

摘 要：神经元集群编码和 spike train 分析是神经信息处理的关键问题。该文介绍了一种利用高阶多维泊松模型对 spike train 进行分类预测的方法，并从 spike 的强度分布、匹配准确性和集成策略上进行了数学论证。最后利用该方法在大鼠 U 迷宫实验中选取 20 组作为训练集进行分类测试，实验结果表明，利用该方法得到的分类准确率在 97%左右。

关键词：信息处理；多维 spike train； 高阶多维泊松模型；贝叶斯原理；预测分类模型
中图分类号：TP18 **文献标识码**： A **文章编号**：1009-5896(2013)07-1619-05
DOI: 10.3724/SP.J.1146.2012.01453

A Classification and Prediction Model of Multi Spike Train Based on Bayes Theory

Fan Yi-na[①] Lang Bo[②] Wei Hui[②]
[①](*Institute of Electric Power, Qinghai University, Xi'ning* 810016, *China*)
[②](*Laboratory of Cognitive Algorithm, Fudan University, Shanghai* 201203, *China*)

Abstract: Neural population encoding and analysis of spike train play an important role in the field of neural inforamtion processing. In this study, a classification method of spike train is proposed based on high-order multiple Possion model, and a mathematic deduction is made in the spike instensity distribution, accuracy of matching and integration strategy, respectively. Finally, 20 trails, as a traing set, are applied to experiment of U maze of mouse. The result demonstrates that the accuracy rate of the classification method is about 97%.

Key words: Information processing; Multi spike train; High-order multiple Possion model; Bayes theory; Prediction classification model

1 引言

神经解码主要研究神经元的时空特性与动物所经历的外部事件之间的关系。该研究的基本框架是首先选择对某类任务敏感的神经元，它们在不同任务或者刺激中表现出明显的发放序列的差异，然后应用隐马尔科夫模型或者贝叶斯原理获得新观察到的神经序列的后验概率[1,2]，spike train 数据的各种分析方法目前已经得到了成熟的发展[3]。一直以来，贝叶斯法则都被认为是处理大脑推断的最优法则，在神经计算领域已经有了许多成功的应用[4]。此外，神经元发放的序列可以被看作是离散的点序列，所以利用多通道记录的 spike train 可以看作是一种多维的点过程，这些点过程是动态随机的，文献[5,6]详细研究了单个神经元的点过程概率模型。另一方面的研究是量化神经元聚群编码信息，例如使用信息熵等方法研究动物下颞叶视皮层中神经元编码的信息量[7−9]。

动物行为预测是神经解码领域非常重要的一个问题[10,11]。在动物自主行为实验中，发现某些神经元(尤其是在额叶皮层和海马区)在每次实验中发放的稳定性很低。这种不稳定性一方面来自于神经元发放的随机机制，另一方面也是由于不同实验环境中大脑内部认知过程的变化导致的。因此，如何找到一种既能够保持发放稳定性又能在较小训练集上有较好输出效果的方法是具有挑战性的。目前对于神经解码的处理主要是借助概率分析的方法[12,13]。这种方法的特点是对神经元的发放规则建立概率模型，然后通过最大似然估计、贝叶斯等理论来预测和分类动物的行为。本文提出的方法就是按照这种思路进行的：首先根据一定的假设建立随机过程的概率模型，得出先验概率和条件概率的推导，当新的 spike train 到来的时候计算后验概率，并通过最

2012-11-12 收到，2013-03-15 改回
国家自然科学基金(30990263)和十二五国家科技支撑计划项目(2012BAI37B06)资助课题
*通信作者：郎波 langbo666@126.com

大似然估计得出后验概率的最大值，并在高阶泊松过程基础上提出了一种多维 spike train 分类、变化及相似性的度量方法。

2 神经元脉冲发放的非齐次泊松分布概率模型

目前对于神经元发放的研究都是基于这样的一种假设：神经元的脉冲发放具有一定的编码规律，而这种规律可以被模型数学语言描述[14−17]。由于在大脑皮层中动作电位到达的时刻是不规则的，这表现为在表达某些信息的时候，神经元会以不同的频率来发放。然而对于单个 spike 来说其产生的时刻又是完全随机的[18]。利用泊松分布可以描述这种情况的分布，即给定发放频率条件下完全随机。我们假设每个动作电位是随时间变换的连续信号，反映了动作电位的瞬时发放频率。同时假设每一个动作电位相互独立，则 spike train 可以用非齐次泊松过程数学模型来进行描述[19−21]。

2.1 估计非齐次泊松过程的强度分布

对于单个神经元，设 $\{T_1,T_2,\cdots,T_M\}$ 表示同一个任务的 spike train 集合，M 为实验次数，T_i 表示第 i 个 spike train。$\{\lambda_i(t)\}$ 表示每次实验的泊松强度函数，则 $\lambda(t)$ 服从参数为 $k(t)$ 和 $\theta(t)$ 的伽玛分布，即

$$p(\lambda(t))=\frac{\lambda(t)^{k(t)-1}e^{-\lambda(t)/\theta(t)}}{\theta(t)^{k(t)}\Gamma(k(t))} \tag{1}$$

其中，$k(t)$ 和 $\theta(t)$ 恒大于 0，表示在 t 时刻 λ 的概率分布，λ 表示泊松过程的强度参数。

我们用 $\overline{\lambda}(t)$ 表示该伽玛分布的均值，在 Δt 足够小的时候，非齐次泊松过程可以看成是齐次泊松过程，且 $N_i(t)\sim \text{Possion}(\lambda_i(t)\Delta t)$，其中 $N_i(t)$ 表示在 Δt 时间内第 i 个神经元产生 spike 的个数。因为泊松分布具有累加性，所以 Δt 内所有 spike train 的动作电位数量 $\sum N_i(t)$ 服从泊松分布的和，即 $\sum N_i(t)\sim \text{Possion}\left(\Delta t\sum\lambda_i(t)\right)$。$\sum N_i(t)$ 的期望值为 $\Delta t\cdot\sum\lambda_i(t)$，因此，$\overline{\lambda}(t)$ 是各个 spike train 在 Δt 内产生的动作电位个数的均值，即

$$\overline{\lambda}(t)=\frac{\sum_{i=1}^{M}N_i(t)}{M\cdot\Delta t} \tag{2}$$

既然引入伽玛分布作为泊松过程的参数的先验概率，我们希望得到 $\lambda(t)$ 的分布的估计，而不仅仅是 $\overline{\lambda}(t)$ 的估计。伽玛分布的方差为 $k(t)\theta(t)^2$，其方差值可以通过若干个样本来估计。设 $\lambda_i(t)=N_i(t)/\Delta t$，则 t 时刻伽玛分布的方差可以通过统计样本方差来估计，其无偏估计量为

$$\delta^2=\frac{(\lambda_i(t)-\overline{\lambda}(t))^2}{n-1} \tag{3}$$

设 $\theta(t)=\delta^2(t)/\overline{\lambda}(t)$，$k(t)=\overline{\lambda}(t)^2/\delta^2(t)$。$f(\lambda)$ 为 λ 的先验分布的概率密度函数。假设在 Δt 时间内观察到 k 个 spike，则

$$P(k|\lambda)=\frac{1}{k!}(\lambda\Delta t)^k e^{-\Delta t\lambda}\propto\lambda^k e^{-\Delta t\lambda} \tag{4}$$

因此，我们得到后验概率的表达式为

$$f(\lambda\mid k)\propto\lambda^k e^{-\lambda t}f(\lambda) \tag{5}$$

接下来的问题就是如何选取先验概率 $f(\lambda)$。在泊松过程中，spike train 的第 k 次发放和第 k+1 次发放出现的时间间隔服从指数分布，根据实验特点，为了计算方便，我们选取 $f(\lambda)$ 为参数是 1 的指数分布，则

$$f(\lambda\mid k)\propto\lambda^k e^{-\lambda(t+1)} \tag{6}$$

2.2 基于贝叶斯原理的 Spike Train 匹配值

设 $\{\lambda_i^j(t)\}$ 表示在第 j 个任务中，第 i 号神经元的非齐次泊松过程的强度参数，其中 $i\in[1,N]$，$j\in[1,K]$，N 是记录到的神经元的个数，K 是研究的任务的数量。通过单个神经元 i 来对一次实验的数据进行分类，所以要找到所有 j 中最大的后验概率 $p(\lambda_i^j(t)\mid T_i)$。设 T_i 表示新到达的实验的第 i 号神经元的 spike train。一个 spike train 可以被分解成一系列 ISI(Inter-Spike Interval)的序列，即 $T_i=(\text{ISI}_1,\text{ISI}_2,\cdots,\text{ISI}_S)$，其中 S 表示该 spike train 中动作电位的个数。根据贝叶斯法则得到

$$p\left(\lambda_i^j(t)|T\right)=\frac{p\left(T_i\mid\lambda_i^j(t)\cdot p(\lambda_i^j(t)\right)}{p(T)} \tag{7}$$

对于任意 k_1，k_2 $(k_1\neq k_2)$，$p(\text{ISI}_{k1}\mid\lambda_i^j(t))$ 与 $p(\text{ISI}_{k2}\mid\lambda_i^j(t))$ 是相互独立的，同时，对于分类任务来说，$p(T_i)$ 对每个任务的值是一样的，由式(7)得到

$$p\left(\lambda_i^j(t)|T_i\right)=p\left(\lambda_i^j(t)\right)\cdot\prod_{k=1}^{S}p\left(\text{ISI}_k\mid\lambda_i^j(t)\right) \tag{8}$$

$p(\text{ISI}_K\mid\lambda_i^j(t))$ 表示非齐次泊松过程的一阶等待时间的概率密度函数

$$p\left(\text{ISI}_k|\lambda_i^j(t)\right)=\lambda_i^j(t_k)\cdot e^{-\int_{t_{k-1}}^{t_k}\lambda_i^j(\tau)\mathrm{d}\tau} \tag{9}$$

其中 t_k 是第 k 个 spike 的产生时间。$p(\lambda_i^j(t))$ 是这个任务的先验概率，为了表达符号的简洁，我们使用 p^j 来表示第 j 个任务的概率。将式(9)代入式(8)，得到

$$p\left(\lambda_i^j(t)|T_i\right)=p(\lambda_i^j(t))\cdot\prod_{k=1}^{S}\lambda_i^j(t_k)\cdot e^{-\int_{t_{k-1}}^{t_k}\lambda_i^j(\tau)\mathrm{d}\tau} \tag{10}$$

式(10)是 spike train 后验概率的完整形式，为了求

得该式的最大值，对其取指数形式，并取最大值，得到

$$\max\arg_j\ p\left(\lambda_i^j(t)|T\right)=\log p^j+\sum_{k=1}^{S}\log\lambda_i^j(t)-\sum_{k=1}^{S}\int_{t_{k-1}}^{t_k}\lambda_i^j(\tau)\mathrm{d}\tau \qquad (11)$$

式(11)表示一个 spike train 和一个泊松过程的匹配值的度量方法，或者称为 spike train 到泊松过程的距离。

我们利用 3 组模拟数据来验证式(11)的有效性。首先，按照泊松强度函数为 $\lambda(t)=10\sin(t)+10$ 产生 300 个动作电位，如图 1 所示。将该 spike train 和几个不同的泊松曲线做比较，使用式(11)计算它们的匹配值，如表 1 所示。可以看到对于非 $\lambda(t)=10\sin(t)+10$ 的泊松过程，其匹配值都会比较小。而 λ_2 的匹配值最高，这在某种程度上验证了可以将式(11)作为 spike train 和泊松过程的度量值。

图 1 按照泊松强度函数 $\lambda(t)=10\sin(t)+10$ 产生的 300 个动作电位的 spike train

表 1 spike train 在不同泊松强度 λ 下的匹配值

λ 函数	匹配值
$\lambda_1(t)=20$	362.4487
$\lambda_2(t)=10\sin(t)+10$	520.8132
$\lambda_3(t)=10\sin(t+\pi/2)+10$	217.6754
$\lambda_4(t)=t$	297.6714

另外，设定 $\lambda(t)=20$ 来产生一个 spike train，如图 2 所示。

对该 spike train 和 $\lambda(t)=c,(c=1,2,\cdots,80)$ 进行匹配计算，使用到的 ISI 的个数 S=1,2,…,300，从图 3 可以看出，使用的 ISI 个数越多，其匹配值越高；当 c=20 的时候匹配值达到峰值，也就是说式(11)可以找到与某个 spike train 产生的泊松过程最接近的部分。

另外使用同一形式参数不同的泊松过程来做匹配。由 $\lambda(t)=10\sin(t)+10$ 来产生一个 spike train，

图 2 根据 $\lambda(t)$=20 产生的 spike train

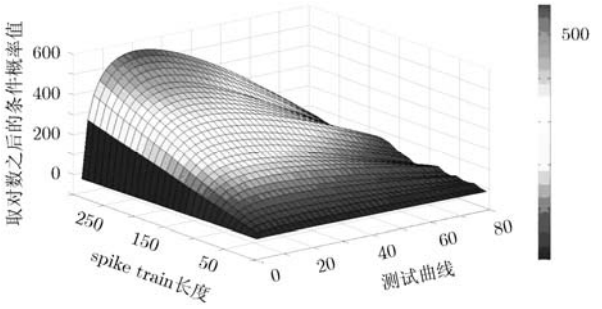

图 3 Spike train 匹配值统计图

如图 4 所示。利用式(11)计算该 spike train 的不同相位的正弦函数在不同 ISI 个数的情况下的匹配值，结果如图 5 所示。

从图 5 中可以看到，将图 4 所示的 spike train 和 $\lambda(t)=10(\sin(t+c)+1),0\le c\le 2\pi$ 做匹配值计算，使用到的 ISI 的个数 S=1,2,…,300。使用的 ISI 个数越多，其匹配程度越高，当 $c=0$ 和 2π，Y=300 的时候，匹配程度最高。这再一次验证了利用式(11)是可以找到与多个 spike train 产生的泊松过程是最接近的值。

此外，利用式(11)我们还可以通过单个神经元来对每次新到达的 spike train 进行分类，将其归类到 $C=(C_1,C_2,\cdots,C_N)$，N 为任务的个数。对于神经元 i，$v_i=(v_{i1},v_{i2},\cdots,v_{iK})$ 表示该神经元对应 K 个任务的泊松曲线和新到来的神经元的匹配值，该匹配值由式(11)计算得到。设 $\overline{v_i}$ 表示 v_i 的平均值，该神经元对本次分类的置信度定义为

$$\mathrm{conf}_i=\frac{1}{1+\exp(-\max(v_i)+\overline{v}_i)} \qquad (12)$$

$\max(v_i)$ 表示 v_i 的最大值。通过 sigmoid 单元的

图 4 泊松强度 $\lambda(t)=10\sin(t)+10$ 产生的一个长度为 300 个 ISI 的 spike train

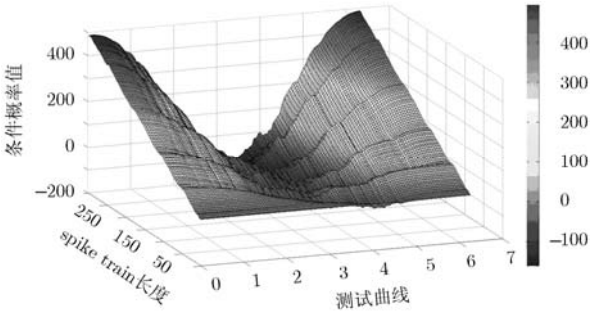

图 5 spike train 匹配值统计图

归一化，每个神经元的分类置信度都是在 0.5 和 1 之间。设 $\mathrm{conf}=(\mathrm{conf}_1,\mathrm{conf}_2,\cdots,\mathrm{conf}_N)$，一个 trail 被分类到任务 i 中，存在

$$i=\max \arg_i \delta_i(C)\cdot \mathrm{conf} \tag{13}$$

其中 δ_i 函数将等于 i 的值变为 1，不等于 i 的函数变为 0，即

$$\delta_i(x)=\begin{cases}1, & x=i\\ 0, & x\neq i\end{cases} \tag{14}$$

式(13)表明，如果每个神经元在本次任务判断中对不同任务的区分度越大，则该神经元所占权重越大。如果一个神经元对不同任务的区分度是大致相同的，则该神经元所占的权重比较小。

3 实验结果

3.1 大鼠 U 迷宫实验

为了检验算法的效果，我们使用大鼠活体动物实验中得到的 multi spike train 来进行分类和预测。在 U 迷宫实验中，训练老鼠在 U 型迷宫里顺时针和逆时针交替饮水。当大鼠跑到位置 1 或位置 2 时，大鼠海马区神经元所发放的 multi spike train 被微电极阵列同步记录下来。在 3 天时间里对同一只大鼠一共进行了 5 组实验，每组实验大约持续 30 min。实验设计如图 6 所示。

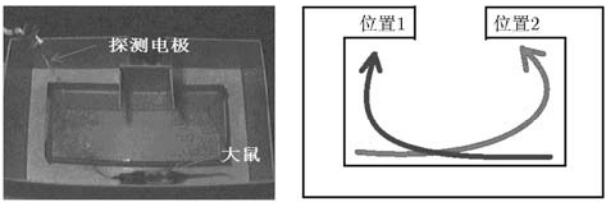

图 6 大鼠 U 迷宫实验示意图

由于在不同组实验中插入大鼠脑中的微电极会有位置上的偏移，同时每组实验通过 spike sorting 得到的结果也不一样。所以在这 5 组实验中被排序出来的神经元的个数也是不一样的。同时大鼠在进行实验时也会因为一些未知因素会被打断，比如受到外部环境的刺激在某几次实验中没有顺时针和逆时针的交替跑动，又或者大鼠在实验过程中停下来等特殊情况。将这些失败的实验次数删除掉，因此顺时针的实验次数和逆时针的实验次数在每组实验中也会有些差异。

我们在每组实验中随机选取 10 次顺时针和 10 次逆时针的实验结果作为训练集，用来估计神经元的对应任务的泊松过程，其余的实验作为测试集，用来测试算法的准确度。上述过程重复进行 100 次，每次都随机选择训练集和测试集，算法平均的准确率作为最终的预测准确率。表 2 列出了 5 组实验的预测准确率，实验结果表明，分类的准确率一直在 97%以上。除此以外，这 5 组测试的结果也显示了算法的鲁棒性，它能在很小的训练数据集上得到较高的准确率。同时，该算法还显示出另一个优势，即不需要人工干预就能得到很好的分类效果。之前的方法大多需要实验人员手工挑选出对任务敏感并且区分度大的神经元来做预测和分类，而本文提出的算法能够自动降低不敏感神经元的权重，提高主导神经元的权重，这样无需实验人员的干预就可以得到非常好的分类效果。

表 2 大鼠 U 迷宫实验参数和结果

组号	神经元个数	顺时针次数	逆时针次数	总持续时间(s)	预测准确率(%)
1	25	77	82	1680	97.94
2	37	43	47	1798	98.53
3	33	60	60	1789	99.82
4	25	24	23	1080	98.55
5	30	24	23	1364	99.10

3.2 算法性能的分析

为了详细研究算法的性能和分类结果，我们分析了每个神经元在训练集大小为 10 和 20 的情况下的分类准确率，结果见图 7。可以看出训练集大小的增加不会使单个神经元的分类准确率产生明显的提高(分类准确率在训练集为 10 的情况下已经接近或者达到 98%)。这个现象说明训练集在 10 左右的时候其分类准确率已经稳定下来并能够得到很好的分类效果。因此，分类准确率的提高主要是由于多通道的集成策略，这种现象更加说明神经系统集群编码的重要性。

4 结论

对于多维 spike train 的分类预测算法，一般有

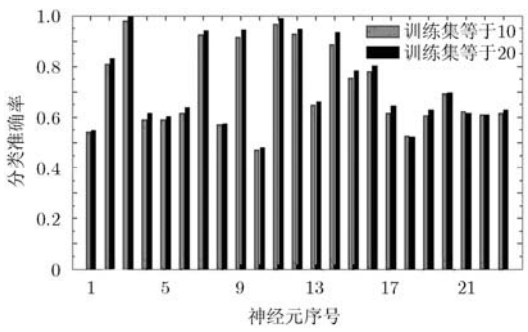

图 7 神经元的分类准确率统计图

两种方案：第 1 种是建立多维的概率模型，该模型包括了神经元之间的相互作用关系。第 2 种方案是将单个神经元的预测结果进行集成和整合。从理论上讲，第 1 种方案更加可靠，并且能够真实地揭示神经元之间的内在关系，但是目前学术界对于多维神经元的理论模型还是比较有争议，因为多维的模型必然涉及到大量的待估计参数，而估计这些参数则需要大量的实验数据。此外，参数的增加将导致训练集呈指数级增长。考虑到实验数据的稀缺，在实际应用中第 2 种方案是比较可行的。本文提出的算法在真实的大鼠行为实验中得到了很好的验证，说明了之前建立的模型是成功的。

参 考 文 献

[1] Ahmadian Y, Pillow J W, and Paninski L. Efficient Markov chain Monte Carlo methods for decoding neural spike trains[J]. *Neural Computation*, 2011, 23(1): 46–96.

[2] Tetzlaff T, Rotter S, Stark E, *et al*.. Dependence of neuronal correlations on filter characteristics and marginal spike train statistics[J]. *Neural Computation*, 2008, 20(9): 2133–2184.

[3] Brown E N, Kass R E, and Mitra P P. Multiple neural spike train data analysis: state-of-the-art and future challenges[J]. *Nature Neuroscience*, 2004, 7(5): 456–461.

[4] Pouget A, Dayan P, and Zemel R S. Inference and computation with population codes[J]. *Annual Review of Neuroscience*, 2003, 26(1): 381–410.

[5] Brillinger D R. Time Series: Data Analysis and Theory[M]. San Francisco: Holden Day, Inc, 2001: 326–411.

[6] Rosenberg J R, Halliday D M, Breeze P, *et al*.. Identification of patterns of neuronal connectivity—partial spectra, partial coherence, and neuronal interactions[J]. *Journal of Neuroscience Methods*, 1998, 83(1): 57–72.

[7] Rolls E T, Aggelopoulos N C, Franco L, *et al*.. Information encoding in the inferior temporal visual cortex: contributions of the firing rates and the correlations between the firing of neurons[J]. *Biological Cybernetics*, 2004, 90(1): 19–32.

[8] Koyama S. On the relation between encoding and decoding of neuronal spikes[J]. *BMC Neuroscience*, 2011, 12(Suppl 1): 177–183.

[9] Ecker A S, Berens P, Keliris G A, *et al*.. Decorrelated neuronal firing in cortical microcircuits[J]. *Science*, 2010, 327(5965): 584–587.

[10] Bialek W, Rieke F, Van Steveninck R R R, *et al*.. Reading a neural code[J]. *Science*, 1991, 252(5014): 1854–1857.

[11] Pillow J W, Shlens J, Paninski L, *et al*.. Spatio-temporal correlations and visual signalling in a complete neuronal population[J]. *Nature*, 2008, 454(7207): 995–999.

[12] Bishop C M and Ligne S S E. Pattern Recognition and Machine Learning[M]. New York: Springer, 2006: 211–236.

[13] Ito H and Tsuji S. Model dependence in quantification of spike interdependence by joint peri-stimulus time histogram[J]. *Neural Computation*, 2000, 12(1): 195–217.

[14] Kostyukov A I, Ivanov Y N, and Kryzhanovsky M V. Probability of neuronal spike initiation as a curve-crossing problem for Gaussian stochastic processes[J]. *Biological Cybernetics*, 1981, 39(3): 157–163.

[15] Mahmud M, Bertoldo A, Girardi S, *et al*.. SigMate: a matlab-based automated tool for extracellular neuronal signal processing and analysis[J]. *Journal of Neuroscience Methods*, 2012, 207(1): 97–112.

[16] Cleanthous A and Christodoulou C. Learning optimisation by high firing irregularity[J]. *Brain Research*, 2011, 1434(3): 115–122.

[17] Kottas A, Behseta S, Moorman D E, *et al*.. Bayesian nonparametric analysis of neuronal intensity rates[J]. *Journal of Neuroscience Methods*, 2011, 203(1): 241–253.

[18] Litwin-Kumar A, Oswald A M M, Urban N N, *et al*.. Balanced synaptic input shapes the correlation between neural Spike trains[J]. *PLoS Computational Biology*, 2011, 7(12): e1002305.

[19] Kass R E and Ventura V. A spike-train probability model[J]. *Neural Computation*, 2001, 13(8): 1713–1720.

[20] Gabbiani F and Koch C. Principles of spike train analysis[J]. *Methods in Neuronal Modeling*, 1998, 12(4): 313–360.

[21] Heeger D. Poisson model of spike generation[Z]. Handout, University of Standford, 2000.

樊一娜：　女，1979 年生，讲师，研究方向为机器学习和模式识别.

郎　波：　男，1974 年生，副教授，研究方向为人工智能和数字媒体处理.

危　辉：　男，1971 年生，教授，博士生导师，研究方向为人工智能、神经信息处理和认知算法.