文章编号: 1002-0411(2000)02-0152-05

模式识别的最大熵方法

张九龙　潘　泉　戴冠中
(西北工业大学自动控制系　西安　710072)

摘　要: 本文提出了模式识别的最大熵方法, 其基本思想是求出最大熵概率分布, 再求出条件概率分布, 进而作出二值分类. 它的特点是能最大限度的利用已有信息作出最合理的推测. 和其它方法相比较, 该方法的突出优点是在小样本情况下仍能保持很好的识别率.

关键词: 模式识别, 最大熵, 概率分布
中图分类号: TP13　　文献标识码: B

1　引言

在模式识别问题中, 经常需要求解概率分布. 即在已知一些样本的情况下, 要求出概率分布. 现考虑这样的二值分类问题, 设有一对象, 它可以分为两类, 可以量测到该对象的 n 个属性, $x_1, x_2, \cdots, x_n$, 要求根据这 n 个属性将对象进行分类, 已有很多方法可以解决这类问题, 例如用前向神经网络及 BP 算法, 文[1]中使用了该方法. 根据已知的样本对网络进行训练, 然后用训练好的网络工作, 但训练样本需要大量的样本, 当样本较少时, 如医学上的疾病诊断、复杂工业过程的故障诊断等, 该方法的效果不很理想. 另外也可以使用概率神经网络(PNN)来解决该问题, 但同样对小样本情况效果不佳. 本文中使用熵方法来解决这类模式识别问题, 结果表明熵方法不仅有很高的识别率, 而且对小样本情况尤其有效.

熵是源于物理学的基本概念, 后来 Shannon 在信息论中引入了信息熵的概念, 它在统计物理中的成功使人们对熵的理论和应用有了广泛和高度的重视. 在信息论中, 最大熵的含义是最大的不确定性, 它解决的一大类问题是在先验知识不充分的条件下进行决策或推断等. 熵方法在谱估计、图象滤波、图象重建、天文信号处理、专家系统等中都有广泛的应用[1-2].

2　最大熵模式识别的原理

设 $S=\{0, 1, 2, \cdots, N\}$ 为一有限集, $\{p_0, p_1, p_2, \cdots, p_n\}$ 为该集合上的概率分布, 即 发生的概率. 那么有:

$$p_j > 0, \quad \forall\ j, \quad \text{且} \sum_{j=0}^{N} p_j = 1$$

该概率分布的熵为:

$$H(P) = -\sum_{j=0}^{N} p_j \ln(p_j) \tag{1}$$

设 f_k, $1=1, 2, \cdots, C$ 为定义在 S 上的函数, 那么这些函数的均值为:

$$\overline{f}_k = \sum_{j=0}^{N} p_j f_k(j) \tag{2}$$

收稿日期: 1998-06-16

我们的目的是求出使 $\overline{f}_k$ 取特定值的概率分布. 一般, 如果 $C< N$, 那么这样的概率分布有很多, 为了从这些分布中选择"最合理"的分布, 就必须有一个挑选准则. E. T. Jaynes 在 1957 年提出了最大熵原理, 即从全部相容的分布中挑选出这样的分布, 它是在某些约束条件下(通常为矩约束)使信息熵达到最大值的分布[4]. 现在的约束是 $\overline{f}_k$ 取特定值, 求使熵最大的概率分布.

理论上可以证明, 当信息熵取极大值时对应的一组概率分布出现的几率占绝对优势. 在 Shannon 引入熵这个概念时, 它的含义是最大的不确定性, 它是用来表示一个概率分布的不确定性的, 选择最大熵分布就意味着选择在给定约束下具有最大不确定性的分布. 从其含有的不确定性来看, 这种分布是最随机的, 是主观成分最少, 把不确定性看作最大的分布.

对于上面所说的模式识别问题, 以 $x_0= 0$ 或 1 代表分类结果, 设 $x_1, x_2, \cdots, x_n$ 的取值都是有限个数的离散值, 则构造如下积空间:

$$S = x_0 \times x_1 \times \cdots \times x_n$$

S 就是样本空间, 利用最大熵原理求出样本空间中每个元素的概率 $p(x_0, x_1, \cdots, x_n)$, 然后再求出 $p(0\mid x_1, \cdots, x_n)> p(1\mid x_1, \cdots, x_n)$, 则模式识别的结果如下:

$$x_0 = \begin{cases} 0, & p(0\mid x_1, \cdots, x_n) > p(1\mid x_1, \cdots, x_n) \\ 1, & p(0\mid x_1, \cdots, x_n) < p(1\mid x_1, \cdots, x_n) \end{cases} \tag{3}$$

下面以一个简单的例子来说明该方法.

设 $x_0\in \{0, 1\}$, $x_1\in \{0, 1, 2\}$, 则样本空间如下:

$$S = \{0, 1\} \times \{0, 1, 2\} = \{(0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2)\}$$

记 $p_{ij}= p(i, j)$,　$i= 0, 1$;　$j= 0, 1, 2$

定义如下矩约束函数:

$$f_1(x_0, x_1) = x_0$$
$$f_2(x_0, x_1) = x_1$$
$$f_2(x_0, x_1) = x_0x_1$$

它们的均值如下:

$$\overline{f}_1 = \sum_{x_0= 0}^{1}\sum_{x_1= 0}^{2} p_{ij}f_1(i, j) = \sum_{x_0= 0}^{1}\sum_{x_1= 0}^{2} p_{ij}x_0$$
$$\overline{f}_2 = \sum_{x_0= 0}^{1}\sum_{x_1= 0}^{2} p_{ij}f_2(i, j) = \sum_{x_0= 0}^{1}\sum_{x_1= 0}^{2} p_{ij}x_1$$
$$\overline{f}_3 = \sum_{x_0= 0}^{1}\sum_{x_1= 0}^{2} p_{ij}f_3(i, j) = \sum_{x_0= 0}^{1}\sum_{x_1= 0}^{2} p_{ij}x_0x_1$$

假设有 M 个样本, 构成样本集$\{(x_{01}, x_{11}), (x_{02}, x_{12}), \cdots, (x_{0M}, x_{1M})\}$, 那么利用该样本集就可以求出 x_1, x_0, x_1, x_1^2 的均值, 即

$$m_1 = \frac{1}{M}\sum_{p= 1}^{M} x_{0p}$$
$$m_2 = \frac{1}{M}\sum_{p= 1}^{M} x_{1p}$$
$$m_3 = \frac{1}{M}\sum_{p= 1}^{M} x_{0p}x_{1p}$$

这样, 三个矩约束条件为:

$$\overline{f}_k = m_k, \qquad k = 1, 2, 3$$

在加上概率归一化条件, 形成四个等式约束, 整理为以下形式:

$$\begin{vmatrix} 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{vmatrix} \begin{vmatrix} p_{00} \\ p_{01} \\ p_{02} \\ p_{10} \\ p_{11} \\ p_{12} \end{vmatrix} = \begin{vmatrix} m_1 \\ m_2 \\ m_3 \\ 1 \end{vmatrix}$$

这个方程有很多组解, 但熵最大的解只有一个, 这时的分布就是最大熵分布.

3 最大熵分布的解

求解最大熵分布相当与一个约束极值问题, 可以用 Lagrange 法解决. 定义

$$G(p_j, \lambda_k) = -\sum_{j=0}^{N} p_j \log p_j - \sum_{k=0}^{C} \lambda_k (\overline{f}_k - m_k) \tag{4}$$

这样, 总共有 $N + C + 2$ 个方程式, 分别为:

$$\frac{\partial G}{\partial p_j} = 0, \qquad j = 0, 1, \cdots, N \tag{5}$$

即 $\log p_j + 1 - \sum_{k=0}^{C} \lambda_k f_k(j) = 0$

和 $\overline{f}_k = m_k$, $k = 0, 1, \cdots, C$. 其中 f_0 是概率归一化条件.

求解最大熵分布的另一种方法是使用 Gibbs 定义的分割函数法(Partition function). 分割函数也是一系列 Lagrange 乘子的函数, 其定义如下[3]:

$$Z(\lambda_1, \lambda_2, \cdots, \lambda_C) = \sum_{j=0}^{N} \exp\left(\sum_{k=0}^{C} \lambda_k f_k(j)\right) \tag{6}$$

记 $Q_j = \exp\left(-\sum_{k=0}^{C} \lambda_k f_k(j)\right)$, 那么 $Z(\lambda_1, \lambda_2, \cdots, \lambda_C) = \sum_{j=0}^{N} Q_j$

最大熵概率如下求得:

$$p_j = \exp\left(-\lambda_0 - \sum_{k=1}^{C} \lambda_k f_k(j)\right) = \frac{Q_j}{\sum_{j=0}^{N} Q_j} \tag{7}$$

其中, $\lambda_0 = \log Z$

λ_k 的求法如下:

$$\frac{\partial \log Z}{\partial \lambda_k} + m_k = 0 \tag{8}$$

针对上面的例子, 有

$$Z(\lambda_1, \lambda_2, \lambda_3) = \sum_{i=0}^{1} \sum_{j=0}^{2} \exp(-i\lambda_1 - j\lambda_2 - ij\lambda_3)$$

$$= 1 + e^{-\lambda_2} + e^{-2\lambda_2} + e^{-\lambda_1} + e^{-\lambda_1 - \lambda_2 - \lambda_3} + e^{-\lambda_1 - 2\lambda_2 - 2\lambda_3}$$

分割函数法的解总满足概率归一化条件, 所以该方法中并不需要单独地提出该约束. 另外三个约束对应的 Lagrange 乘子可如下求得:

$$m_1(1+e^{-\lambda_2}+e^{-2\lambda_2})+(m_1-1)(e^{-\lambda_1}e^{-\lambda_1-\lambda_2-\lambda_3}+e^{-\lambda_1-2\lambda_2-2\lambda_3})=0$$

$$m_2(1+e^{-\lambda_2})+(m_2-1)(e^{-\lambda_2}+e^{-\lambda_1-\lambda_2-\lambda_3}+(m_2-2)(e^{-2\lambda_2}+e^{-\lambda_1-2\lambda_2-2\lambda_3})=0$$

$$m_3(1+e^{-\lambda_2}+e^{-2\lambda_2}+e^{-\lambda_1})+(m_3-1)e^{-\lambda_1-\lambda_2-\lambda_3}+(m_3-2)e^{-\lambda_1-2\lambda_2-2\lambda_3})=0$$

定义:

$$x=e^{-\lambda_1}$$

$$y=e^{-\lambda_2}$$

$$z=e^{-\lambda_3}$$

则以上约束变为:

$$m_1(1+y+y^2)+(m_1-1)(x+xyz)+xy^2z^2)=0$$

$$m_2(1+x)+(m_2-1)(y+xyz)+(m_2-2)(y^2+xy^2z^2)=0$$

$$m_3(1+x+y+y^2)+(m_3-1)xyz+(m_3-2)xy^2z^2=0$$

求出 x, y, z 后, 就可以利用(7)式求出最大熵分布.

4　算例

4.1　算例 1

设研究的对象有两个属性, 记为 x_1, x_2, 它们的取值是 $x_1\in\{0,1\}$, $x_2\in\{0,1\}$, 它们可以将对象分为真假两类. 样本空间共有 8 个元素, 其概率分布如表 1 所示.

表 1　概率分布

(0 0 0)	(0 0 1)	(0 1 0)	(0 1 1)	(1 0 0)	(1 0 1)	(1 1 0)	(1 1 1)
0.01	0.24	0.01	0.24	0.24	0.01	0.24	0.01

在该算例中选取了 4 个矩约束, 分别为:

$$f_1(j)=x_1;\ f_2(j)=x_2;\ f_3(j)=x_0x_1;\ f_4(j)=x_0x_2$$

针对不同大小的样本集, 求出的概率分布如表 2 所示.

表 2　概率分布

200	0.1002	0.2936	0.0636	0.1864	0.1002	0.0060	0.2360	0.0140
150	0.1029	0.2968	0.0635	0.1832	0.1029	0.0040	0.2374	0.0093
100	0.1002	0.2936	0.0636	0.1864	0.1002	0.0060	0.2360	0.0140
90	0.1172	0.2990	0.0657	0.1676	0.1172	0.0000	0.2333	0.0000

可以发现, 对于任意样本数, 总有 $p_0=p_4$, 这是由于所选取的矩约束使得 $Q_0=Q_4$ 而导致的. 这说明针对该样本空间, 所选的矩约束是不合适的.

4.2　算例 2

同上, 设研究的对象有两个属性, 记为 x_1, x_2, 但它们的取值是 $x_1\in\{0,1\}$, $x_2\in\{1,2\}$, 样本空间的概率分布同表 1, 并使用同样的矩约束. 针对不同大小的样本集, 求出的概率分布如表 3 所示.

根据最大熵模式识别的原理, 由表 3 可以得出, 识别的准确率为 75%, 而且, 识别率不随样本数的减少而减小.

另外, 为了和其它方法进行比较, 文[1]中针对前向网络及 BP 算法进行了减少样本集的实验, 结果表明该方法的识别率随样本集的减小有明显下降.

表 3

200	0.1616	0.2131	0.1078	0.1422	0.0770	0.0483	0.1536	0.0964
150	0.1717	0.2097	0.1111	0.1356	0.0799	0.0454	0.1573	0.0894
100	0.1616	0.2131	0.1078	0.1422	0.0770	0.0483	0.1536	0.0964
90	0.2137	0.1885	0.1240	0.1093	0.0944	0.0368	0.1679	0.0654
80	0.2008	0.1970	0.1199	0.1176	0.0887	0.0385	0.1657	0.0718
70	0.2290	0.1783	0.1285	0.1001	0.1009	0.0346	0.1702	0.0584
60	0.2137	0.1885	0.1240	0.1093	0.0944	0.0368	0.1679	0.0654
50	0.1932	0.2021	0.1173	0.1227	0.0853	0.0394	0.1642	0.0758
40	0.2408	0.1705	0.1317	0.0933	0.1058	0.0329	0.1716	0.0534

造成上述两种方法差别结果的原因是, BP 网络利用单个样本进行训练, 而最大熵方法则样本整体的一些统计特性, 更能反映其整体信息. 而利用单个样本训练的 BP 网络在样本较少时则很难保证识别的一致性.

5 结论

本文讨论了最大熵原理, 当要求在先验知识不充分的条件下进行推断或决策时, 最大熵方法可以作出最合理的推测. 本文给出了它在模式识别中的应用. 尤其, 当样本数较少时, 该方法较其它方法更有效.

参 考 文 献

1 Poh Lian Choog *et al*. Entropy Maximization Networks: An Application to Breast Cancer Prognosis, IEEE Trans. Neural Networks, 1996, (3)

2 Dov Ingman *et al*. Maximum Entropy Signal Reconstruction with Neural Networks, IEEE Trans. Neural Networks, 1992, (2)

3 Robert L. Fry Observer-Participant Models of Neural Processing IEEE Trans. Neural Networks, 1995, (4)

4 王 彬. 熵与信息. 西北工业大学出版社, 1994

MAXIMUM ENTROPY METHOD FOR PATTERN RECOGNITION

ZHANG Jiu-long PAN Quan DAI Guan-zhong

(*Northwestern Polytechnical University, Xi'an* 710072, *P. R. China*)

Abstract: In this paper, maximum entropy method is applied to pattern recognition; the key point is to find the probability distribution that has maximum entropy measure, which ensures that the distribution is most reasonable. Compared with other methods, this method is especially useful when the sample set is small.

Keywords: maximum entropy method, pattern recognition, probability distribution

作者简介

张九龙(1974-), 博士生. 研究领域为神经网络控制系统、信息熵及应用, 虚拟现实系统等.

潘 泉(1961-). 研究领域为随机最优估计与控制, 数据融合, 多目标跟踪, 智能信息处理.

戴冠中(1937-), 校长, 教授, 博士生导师. 研究领域为大系统估计与控制, 智能控制, 控制系统中的并行处理理论、算法与并行计算机.