

基于 Kullback-Leibler 与 PCA 的概率 密度比值估计

兰远东^{1,2}, 邓辉舫¹

(1. 华南理工大学 计算机科学与工程学院, 广东 广州 510000;

2. 惠州学院 计算机科学系, 广东 惠州 516007)

摘 要:为了更好地解决在机器学习和数据挖掘等领域中经常遇到的两个概率密度函数的比值估计问题,文中提出了一种新的概率密度比值估计算法。该算法基于 Kullback-Leibler 距离,综合混合高斯模型和主成分分析的概率密度比值估计方法,使用混合概率主成分分析为两个概率密度比值函数建模。在概率密度比值估计的过程中,不是分别估计比值函数的分子和分母,而是对整个比值函数进行混合组成建模。算法避免了分别对分子分母的概率密度估计,降低了估计的误差。实验表明该算法能够获得较好的估计结果。

关键词:概率密度;机器学习;主成分分析;样本空间

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2012)06-0107-04

Ratio Estimation of Probability Density Based on Kullback-Leibler and PCA

LAN Yuan-dong^{1,2}, DENG Hui-fang¹

(1. School of Computer Science and Engineering, South China University of Technology,

Guangzhou 510000, China;

2. Department of Computer Science, Huizhou University, Huizhou 516007, China)

Abstract: In order to solve the estimation problems for the ratio of two probability density functions, which is often encountered in machine learning and data mining area, propose a new estimation algorithm for the ratio of the probability density. The algorithm is based on Kullback-Leibler distance, integrated Gaussian mixture model and principal component analysis of the estimation methods, use mixed probabilistic principal component analysis for the modeling. In the estimation process for the ratio of the probability density, not separately estimate the numerator and denominator of the ratio function, but modeling the function in the same time. In this way, the algorithm can reduce the estimated error. Experiments show that the algorithm can obtain better result.

Key words: probability density; machine learning; principal component analysis; sample space

0 引 言

在解决诸如迁移学习 (transfer learning)^[1]、多任务学习 (multi-task learning)^[2]、离群点检测 (outlier detection)^[3]、特征选择 (feature selection)^[4]、特征提取 (feature extraction)^[5,6]、条件密度估计以及概率分类等实际问题的过程中,通常需要估计两个概率密度函数的比值。因此,概率密度比值估计受到广泛关注。一种比较直观的做法是分别估计分子和分母的概率密

度函数值,进而得出比值。但是,概率密度估计本身就是一个非常困难的问题,如果分别对分子分母的概率密度估计更容易加大估计的误差。为了解决这个问题,文献[7]提出了一种直接估计概率密度比值的方法—KLIEP (Kullback-Leibler Importance Estimation Procedure),用球形高斯核的线性组合来为概率密度比值函数建模,通过交叉验证选择高斯宽度。

虽然实验显示 KLIEP 能够获得较好的结果,但是在某些情况下使用椭球的高斯内核会更适合。鉴于此,文献[8]在 KLIEP 的基础上提出了混合高斯模型的 GM-KLIEP (Gaussian-Mixture KLIEP) 算法。在 GM-KLIEP 算法中,使用期望最大化算法 (Expectation Maximization Algorithm) 来训练高斯混合模型,训练过

收稿日期:2011-11-09;修回日期:2012-02-13

基金项目:国家自然科学基金(61170193)

作者简介:兰远东(1975-),男,博士研究生,研究方向为模式识别与机器学习;邓辉舫,教授,博士生导师,研究方向为模式识别与机器学习。

程中,高斯内核的协方差矩阵做自适应调整。在实验中 GM-KLIEP 能取得非常好的估计结果,但是算法的训练过程中需要计算(估计)样本的逆协方差矩阵,这就使得在样本存在局部秩亏的时候,GM-KLIEP 的数值不稳定。

解决秩亏的一个常用方法就是对样本空间进行降维,如使用主成分分析(Principal Component Analysis, PCA)。基于这个思想,文中提出了一种综合 GM-KLIEP 和 PCA 的概率密度比值估计算法,该算法使用混合概率主成分分析模型来为概率密度比值函数建模。在模拟实验中显示,该算法具有较好的估计精度。

1 背景介绍

1.1 问题描述

设数据域 $D \in R^d$, 给出独立同分布的样本集 $\{x_i^{de}\}_{i=1}^{n_{de}}$, 该分布的概率密度函数为 $p_{de}(x)$; 另外给出独立同分布的样本集 $\{x_j^{nu}\}_{j=1}^{n_{nu}}$, 该分布的概率密度函数为 $p_{nu}(x)$ 。假定对所有数据域 D 中的样本 x 有 $p_{de}(x) > 0$, 文中的目标就是设计一种算法, 从给出的样本集来估计上面两个概率密度函数的比值:

$$w(x) = \frac{p_{nu}(x)}{p_{de}(x)} \quad (1)$$

并且在估计 $w(x)$ 的过程中, 不能直接估计 $p_{de}(x)$ 和 $p_{nu}(x)$ 。

1.2 KLIEP 算法

KLIEP 可以不通过估计概率密度函数来估计 $w(x)$, 它采用下面的线性组合模型来估计 $w(x)$:

$$\hat{w}(x) = \sum_{i=1}^b \alpha_i \varphi_i(x) \quad (2)$$

其中 $\{\alpha_i\}_{i=1}^b$ 是参数, b 是参数的个数, $\varphi_i(x)$ 是高斯基函数, 具体如下:

$$\varphi_i(x) = \exp\left(-\frac{\|x - c_i\|^2}{2\tau^2}\right) \quad (3)$$

其中 $\tau > 0$ 是高斯宽度, c_i 是从 $\{x_j^{nu}\}_{j=1}^{n_{nu}}$ 中随机选择的高斯中心。通过使用高斯线性组合模型 $\hat{w}(x)$, 可以这样来估计 $p_{nu}(x)$:

$$\hat{p}_{nu}(x) = \hat{w}(x)p_{de}(x) \quad (4)$$

我们自然希望 $p_{nu}(x)$ 和 $\hat{p}_{nu}(x)$ 的差异越小越好, 因此希望 $p_{nu}(x)$ 和 $\hat{p}_{nu}(x)$ 之间的 Kullback-Leibler 距离最小化, 即:

$$\begin{aligned} & KL[p_{nu}(x) || \hat{p}_{nu}(x)] \\ &= \int p_{nu}(x) \ln \frac{p_{nu}(x)}{p_{de}(x)\hat{w}(x)} dx \\ &= \int p_{nu}(x) \ln \frac{p_{nu}(x)}{p_{de}(x)} dx - \int p_{nu}(x) \ln \hat{w}(x) dx \end{aligned} \quad (5)$$

可以看出, 在式(5)中, 最后一个等式的第一项与

$\{\alpha_i\}_{i=1}^b$ 无关, 因此可以忽略不记。将上面的第二项定义为 KL' :

$$KL' = \int p_{nu}(x) \ln \hat{w}(x) dx \approx \frac{1}{n_{nu}} \sum_{j=1}^{n_{nu}} \ln \hat{w}(x_j^{nu}) \quad (6)$$

在上式中用独立同分布的样本集 $\{x_j^{nu}\}_{j=1}^{n_{nu}}$ 的均值来近似 $p_{nu}(x)$ 的期望。因为 $\hat{p}_{nu}(x)$ 是概率密度函数的估计, 所以可以得到下式:

$$\begin{aligned} 1 &= \int \hat{p}_{nu}(x) dx = \int p_{de}(x) \hat{w}(x) dx \\ &\approx \frac{1}{n_{de}} \sum_{i=1}^{n_{de}} \hat{w}(x_i^{de}) \end{aligned} \quad (7)$$

其中, $p_{de}(x)$ 的期望可以用独立同分布的样本集 $\{x_i^{de}\}_{i=1}^{n_{de}}$ 的均值来近似。那么 KLIEP 的优化问题, 就可以用下式给出:

$$\max_{\{\alpha_i\}_{i=1}^b} \left[\sum_{j=1}^{n_{nu}} \ln \left(\sum_{i=1}^b \alpha_i \varphi_i(x_j^{nu}) \right) \right] \quad (8)$$

$$\text{s. t. } \frac{1}{n_{de}} \sum_{i=1}^{n_{de}} \sum_{j=1}^b \alpha_i \varphi_i(x_i^{de}) = 1$$

$$\text{and } \alpha_1, \dots, \alpha_b \geq 0 \quad (9)$$

1.3 通过交叉验证选择模型

在 KLIEP 算法中, 高斯宽度 τ 的选择至关重要。因为 KLIEP 算法是要最小化 $p_{nu}(x)$ 和 $\hat{p}_{nu}(x)$ 之间的 Kullback-Leibler 距离 $KL[p_{nu}(x) || \hat{p}_{nu}(x)]$, 也就是需要最大化 KL' , 所以选择能使 KL' 最大化的高斯宽度。

KL' 中 $p_{nu}(x)$ 的期望, 可以通过交叉验证的方法来近似。先将 $\{x_j^{nu}\}_{j=1}^{n_{nu}}$ 分为 K 个近似相等的不相交子集 $\{x_j^{nu}\}_{j=1}^K$, 然后用其中的 $K-1$ 个子集, 即 $\{x_j^{nu}\}_{j \neq k}$ (没有 x_k^{nu}) 来估计 KL' 。

$$\hat{KL}'_k = \frac{1}{|\mathcal{X}_k^{nu}|} \sum_{x \in \mathcal{X}_k^{nu}} \ln \hat{w}_k(x) \quad (10)$$

将上面的过程重复 k 次 ($k=1, \dots, K$), 用 k 次估计的平均值来作为 KL' 的估计, 即:

$$\hat{KL}'_k = \frac{1}{K} \sum_{k=1}^K \hat{KL}'_k \quad (11)$$

2 混合概率主成分分析的 KLIEP

文中提出一种新的估计概率密度比值的算法 PP-CA-mixture KLIEP (PM-KLIEP)。文中的算法不使用式(1)所示的线性组合模型, 而使用式(12)所示的模型来估计概率密度的比值。

$$w(x) = \sum_{i=1}^b \pi_i p_i(x) \quad (12)$$

其中, b 是混合成分的数量, $\{\pi_i\}_{i=1}^b$ 是混合系数, $\{p_i(x)\}_{i=1}^b$ 是概率主成分分析器, 定义如下:

$$\begin{aligned} p_i(x) &= (2\pi\sigma_i^2)^{-\frac{d}{2}} \det(C_i)^{-\frac{1}{2}} \\ &\exp\left(-\frac{1}{2}(x - \mu_i)^T TC_i^{-1}(x - \mu_i)\right) \end{aligned} \quad (13)$$

其中, \det 是行列式, $C_l = \sigma_l^2 I_d + W_l W_l^T$, I_d 是 d 阶单位矩阵, W_l^T 是 W_l 的转置, $\{W_l \in R^{d \times m}, \mu_l \in R^d, \sigma_l > 0\}_{l=1}^b$ 是相应的参数, d 是输入空间的维度, $m \leq d$ 是潜在空间的维度。这样 PM-KLIEP 的优化问题, 就转化为:

$$\max_{\{\pi_l, W_l, \mu_l, \sigma_l\}_{l=1}^b} \left[\sum_{j=1}^{n_{nu}} \ln \left(\sum_{l=1}^b \pi_l p_l(x_j^{nu}) \right) \right] \quad (14)$$

$$\text{s. t. } \frac{1}{n_{de}} \sum_{i=1}^{n_{de}} \sum_{l=1}^b \pi_l p_l(x_i^{de}) = 1$$

and $\pi_1, \dots, \pi_b \geq 0$ (15)

使用期望最大化算法(expectation-maximization algorithm)来求解上面的优化问题。

初始化: 对参数 $\{\pi_l, W_l, \mu_l, \sigma_l\}_{l=1}^b$ 进行初始化。

E-step: 使用当前参数 $\{\pi_l, W_l, \mu_l, \sigma_l\}_{l=1}^b$ 计算下式:

$$\gamma_{lj} = \frac{\pi_l p_l(x_j^{nu})}{\sum_{l=1}^b \pi_l p_l(x_j^{nu})}, \text{ 其中 } l=1, \dots, b; j=1, \dots, n_{nu}$$

n_{nu}

M-step: 使用当前得到的 $\{\gamma_{lj}\}_{l=1, j=1}^{b, n_{nu}}$ 重新估计参数 $\{\pi_l, W_l, \mu_l, \sigma_l\}_{l=1}^b$

$$\pi_l = \frac{n_{de} \sum_{j=1}^{n_{nu}} \gamma_{lj}}{n_{nu} \sum_{l=1}^b \sum_{j=1}^{n_{nu}} \gamma_{lj}}$$

$$W_l = \left(\sum_{j=1}^{n_{nu}} \gamma_{lj} (x_j^{nu} - \mu_l) (x_j^{nu} - \mu_l)^T \right) \left(\sum_{j=1}^{n_{nu}} \gamma_{lj} C_{lj} \right)^{-1}$$

$$\sigma_l^2 = \frac{1}{d \sum_{j=1}^{n_{nu}} \gamma_{lj}} \sum_{j=1}^{n_{nu}} \left(\gamma_{lj} \|x_j^{nu} - \mu_l\|^2 - 2 \gamma_{lj} z_{lj}^T W_l^T (x_j^{nu} - \mu_l) + \gamma_{lj} \text{tr}(C_{lj} W_l^T W_l) \right)$$

其中, z_{lj} 和 C_{lj} 是潜变量的期望和协方差^[9,10], 分别定义如下:

$$z_{lj} = M_l^{-1} W_l (x_j^{nu} - \mu_l)$$

$$C_{lj} = \sigma_l^2 M_l^{-1} + z_{lj} z_{lj}^T$$

$$M_l = \sigma_l^2 I + W_l^T W_l$$

评估: 重复 E-step 和 M-step 直到下面的似然函数收敛,

$$\sum_{j=1}^{n_{nu}} \ln \left(\sum_{l=1}^b \pi_l p_l(x_j^{nu}) \right) \quad (16)$$

3 实验

在这一小节, 通过实验来对比 PM-KLIEP, KLIEP 和 GM-KLIEP 的性能。首先考虑一个局部秩亏的二维概率密度比值估计问题。

分母 ($p_{de}(x)$) 和分子 ($p_{nu}(x)$) 的密度分别定义如下:

$$p_{de}(x) = \frac{1}{2} N \left(x; \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & \varepsilon \end{bmatrix} \right) +$$

$$\frac{1}{2} N \left(x; \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} \varepsilon & 0 \\ 0 & 2 \end{bmatrix} \right) \quad (17)$$

$$p_{nu}(x) = \frac{1}{2} N \left(x; \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix} \right) +$$

$$\frac{1}{2} N \left(x; \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} \varepsilon & 0 \\ 0 & 1 \end{bmatrix} \right) \quad (18)$$

其中, $N(\cdot; \mu, \Sigma)$ 是高斯密度函数, 均值 μ , 协方差矩阵为 Σ , $\varepsilon = 10^{-15}$ 。设定 $n_{de} = 100$, $n_{nu} = 1000$ 。在 KLIEP 中, 设 $b = 100$, 并且使用高斯核作为基函数, 高斯核的宽度使用 5 重交叉验证确定。在 GM-KLIEP 和 PM-KLIEP 中, 使用 k-means 聚类算法对参数初始化^[11,12], 混合成分的数目也使用 5 重交叉验证确定。

图 1 显示的实验结果表明 PM-KLIEP 比 KLIEP 和 GM-KLIEP 相比能够获得更好的性能。

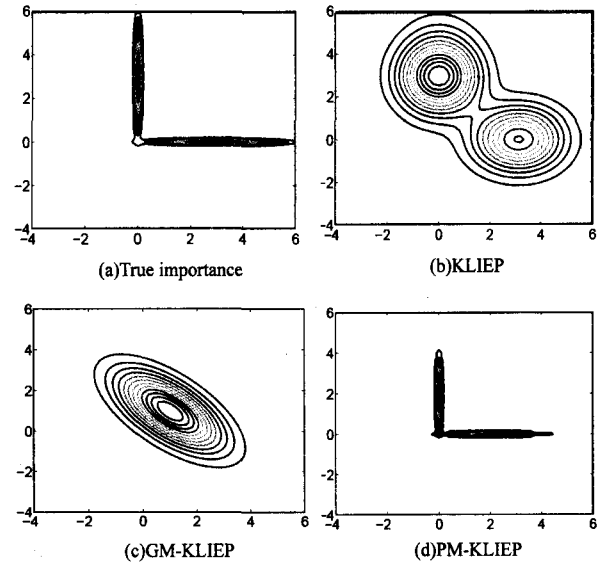


图 1 各算法对比值估计的效果

接下来, 将文中提出的算法 PM-KLIEP 与 KLIEP 和 GM-KLIEP 在离群点检测中的性能进行对比。离群点检测问题, 已经引起人们的广泛关注。离群点检测在网络入侵检测、信用卡欺诈、电子商务犯罪、医疗诊断以及反恐等诸多领域都具有十分重要的作用。离群点检测的目的是为了发现数据集中的一小部分对象, 与数据集中其余的大部分对象相比, 这一小部分对象有着特殊的行为或者具有反常的属性。现存的很多离群点检测算法的实现过程中都会涉及到概率密度函数的比值计算, 因此可以用文中提出的概率密度比值估计算法来替代常规的概率密度函数比值计算。

考虑两个数据集的密度比值, 非离群点倾向于两个不同的中心, 而离群点则会显著地偏离数据集的中心。使用 IDA 基准数据集来评价算法的性能。初始数据集是一个二分类的数据集, 每一个数据子集中包

含有正样本和负样本以及训练集和测试集。将正样本作为内部结点,而将负样本作为离群点。将 95% 的正样本作为训练集,其余 5% 作为测试集。

实验结果(离群点检测精度)如表 1 所示,可以看出 PM-KLIEP 的性能要优于 KLIEP 和 GM-KLIEP。

表 1 各个算法的检测精度

Datasets	KLIEP	GM-KLIEP	PM-KLIEP
banana	56.9	70	60.4
diabetes	64	53.1	67.4
heart	71	73.1	73.6

4 结束语

文中提出了一种新的概率密度比值估计算法,该算法使用混合概率密度主成分分析的方法来估计概率密度比值。对算法的优化,可以通过期望最大化算法有效实现。实验证明该算法相比其他算法更有效。

参考文献:

- [1] 梅灿华,张玉红,胡学钢,等.一种基于最大熵模型的加权归纳迁移学习[J].计算机研究与发展,2011,48(9):1722-1728.
- [2] 许棣华,王志坚.基于多任务学习的邮件过滤系统的研究[J].计算机技术与发展,2010,20(10):137-141.
- [3] 施冬冬,贾瑞玉,黄义堂.基于遗传算法的高维离群点检测

算法的改进[J].计算机技术与发展,1999,19(3):141-147.

- [4] 张家柏,王小玲.基于聚类和二进制 PSO 的特征选择[J].计算机技术与发展,2010,20(6):25-28.
- [5] 姜 鹤,陈丽亚.SVM 文本分类中一种新的特征提取方法[J].计算机技术与发展,2010,20(3):17-23.
- [6] 袁 健,姚明海.基于简化局部二元法的人脸特征提取[J].计算机技术与发展,2009,19(6):84-90.
- [7] Suzuki T, Sugiyama M, Kanamori T, et al. Mutual information estimation reveals global associations between stimuli and biological processes[J]. BMC Bioinformatics, 2009, 10(1): S52.
- [8] Suzuki T, Sugiyama M. Sufficient dimension reduction via squared-loss mutual information estimation[C]//Intl. Conf. on Artificial Intelligence and Statistics. [s. l.]: [s. n.], 2010: 804-811.
- [9] Sugiyama M, Suzuki T, Nakajima S, et al. Direct importance estimation for covariate shift adaptation[J]. Annals of the Institute of Statistical Mathematics, 2008, 60(4): 699-746.
- [10] Yamada M, Sugiyama M. Direct importance estimation with Gaussian mixture models[J]. IEICE Trans. on Information and Systems, 2009, E92-D(10): 2159-2162.
- [11] Tipping M E, Bishop C M. Mixtures of probabilistic principal component analyzers[J]. Neural Computation, 1999, 11(2): 443-482.
- [12] Bishop C M. Pattern recognition and machine learning[M]. New York, USA: Springer-Verlag, 2006.

(上接第 106 页)

4 结束语

分析了传统 Meanshift 算法的局限性,针对快速运动目标容易跟踪丢失,提出了基于目标质心的 Mean-shift 目标跟踪算法,经实验仿真该方法对快速运动目标具有良好的跟踪性能,有较高的应用价值。

参考文献:

- [1] 雷 云,王夏黎,孙 华.基于视频的交通目标跟踪方法研究[J].计算机技术与发展,2010,20(7):44-47.
- [2] 刘卫光,李广鑫.一种通用的视频目标跟踪系统设计[J].计算机技术与发展,2009,19(10):110-113.
- [3] Comaniciu D, Ramesh V, Meer P. Kernel-based object tracking[J]. IEEE transaction on pattern analysis and machine intelligence, 2003, 25(5): 564-577.
- [4] Luo Cheng, Cai Xiongcai, Zhang Jian. Robust object tracking using the particle filtering and level set methods: a comparative experiment[C]//Proceedings of the 2008 IEEE 10th Workshop on Multimedia Signal Processing. Cairns, Australia: [s. n.], 2008: 359-364.
- [5] 朱胜利. Meanshift 及相关算法在视频跟踪中的研究[D].

杭州:浙江大学,2006.

- [6] Cheng Y. Meanshift, Mode Seeking and Clustering[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1995, 17(8): 790-799.
- [7] 施 华,李翠华.图像中的运动目标跟踪[J].计算机工程与应用,2005,41(10):56-58.
- [8] Collins R. Meanshift blob tracking through scale space[C]//IEEE Conference on Computer Vision and Pattern Recognition. [s. l.]: [s. n.], 2003: 234-240.
- [9] McFarlane N J B, Schofield C P. Segmentation and tracking of piglets in images[J]. Machine Vision and Applications, 1995(8): 187-193.
- [10] Hang Z, Faugeras O D. Three dimensional motion computation and object segmentation in along sequence of stereo frames[J]. International Journal on Computer Vision, 1992(3): 211-241.
- [11] Liu Y, Huang T S. Determining straight line correspondences from intensity images[J]. Pattern Recognition, 1991, 24(6): 119-216.
- [12] Ferruz J, Ollero A. Integrated real time vision system for vehicle control in non-structured environments[J]. Engineering Applications of Artificial Intelligence, 2000(13): 215-236.