

文章编号:1001-9081(2010)05-1273-04

基于二次 Renyi 熵的正则化互信息特征选择方法

洪智勇, 刘灿涛, 邓宝林

(五邑大学 计算机学院, 广东 江门 520290)

(hongmr@163.com)

摘要:提出了一种基于二次 Renyi's 熵的正则化互信息特征选择方法,该方法能高效地对互信息进行估计从而使计算复杂度大大降低。同时把正则化互信息特征选择方法与嵌入式方法相结合得到一个两段式特征选择算法,该算法可以找出更具特征的特征子集。通过实验比较了该方法与其他基于互信息的特征选择算法的效率与分类精度,结果表明该方法能够有效改善计算复杂度。

关键词:特征选择;互信息;Renyi 熵;熵估计

中图分类号: TP182; TP311 **文献标志码:** A

Normalized mutual information feature selection method based on Renyi's quadratic entropy

HONG Zhi-yong, LIU Can-tao, DENG Bao-lin

(School of Computer Science, Wuyi University, Jiangmen Guangdong 529020, China)

Abstract: Normalized mutual information feature selection method was proposed based on quadratic Renyi's entropy. This method could efficiently estimate the mutual information, so that the computational complexity was greatly reduced. At the same time, the normalized mutual information feature selection method and embedding method were combined to get a two-stage feature selection algorithm, which could find more characteristic feature subsets. The experimental results compared with the other similar algorithms show that the proposed method can effectively reduce the computational complexity.

Key words: feature selection; mutual information; Renyi entropy; entropy estimation

0 引言

特征选择在模式识别、数据挖掘以及更多的通用机器学习任务中起着关键性的作用,例如:可视化、分类、检测与校验。它是从原先可能毫无相关或是具有冗余的特征中提出具有高度预测性特征的过程。特征选择方法大致可分为两类:过滤器方法(Filter)与嵌入式方法(Wrapper)。过滤器方法主要是通过训练集中数据的特性提取出一个特征子集,并且单独地作为机器学习算法的一个步骤,因此它不区分任何学习算法。而嵌入式方法则是根据某一特殊分类算法与应用领域来选择最优特征,因此它要求一个预知的算法并且根据算法的性能来评估和决定选择哪些特征。评价特征的标准有很多例如:类间和类内距离测量法、互信息、相关测量法、相容测量法和分类错误测量法。

本文提出一种基于互信息的特征选择方法,以互信息来衡量特征中相关性及冗余度。Battiti 在文献[1]中提出 MIFS 方法,其思想是在已经得到一组已选择的特征情况下,下一个选择的特征要求与对应的类别具有最大的互信息并且与已选择的特征具有最小的互信息,文中还给出了一个相应的贪心选择算法。然而该方法通常很难对多元密度给出一个准确的估计,因此要计算类别与所选特征的 Shannon 互信息是不可能的。Kwak 等人在文献[2]中提出一种贪心选择方法 MIFS-U 来克服 MIFS 的这些限制。虽然通常状况下 MIFS-U 性能比 MIFS 要好,但当待选特征与已选择的特征相近时计算性能就出现下降。Peng 等人在文献[3]中提出最小冗余与最大相关性标准

mRMR,使用 Parzen Gaussian 窗^[4]来估计连续变量间的互信息,避免了多元密度估计的困难。Estévez 等人在文献[5]中提出两种对 MIFS、MIFS-U 和 mRMR 的改善方案,这种方法称为正则化互信息特征选择(NMIFS),并且文中还给出由互信息导向的遗传算法(GAMIFS),解决了 MIFS-U 由于一组相关特征而导致的性能下降问题。

另一方面,Hild 等人在文献[6]中基于一种类似 Renyi 熵互信息通过有监督学习来提取特征,应用 Parzen 窗与高斯核来估计二次 Renyi 熵而不是估计 Shannon 熵,因此降低了计算的复杂度。Bonev 等人在文献[7]中借助于 Renyi 熵的图相似性及 Shannon 熵相似性避开了概率密度函数估计,并基于上述工作提出一种新的用于监督学习中的特征选择过滤器方法。本文提出基于二次 Renyi 熵特征选择的正则化互信息,它依靠对互信估计效率而减少了 NMIFS 计算复杂度;再把 NMIFS 与嵌入式方法相结合提出一个两段式特征选择算法,该算法能找出较好的特征子集;最后通过实验比较了新方法与其他基于互信息特征选择算法的分类结果。

1 互信息与 Renyi 熵的估计

1.1 二次 Renyi 熵及估计

对于连续变量 X , 其概率密度函数(Probability Density Function, PDF)为 $f_X(x)$, Shannon 微分熵定义为:

$$H_S(X) = - \int_S f_X(x) \log f_X(x) dx \quad (1)$$

Renyi 微分熵定义为:

收稿日期:2009-12-03;修回日期:2010-02-04。

作者简介:洪智勇(1978-),男,江西上饶人,讲师,博士研究生,主要研究方向:智能信息处理、数据挖掘;刘灿涛(1976-),男,湖北武穴人,讲师,博士研究生,主要研究方向:模式识别、机器学习、数据挖掘;邓宝林(1983-),男,广东肇庆人,工程师,主要研究方向:数据挖掘。

$$H_{R_a}(X) = \frac{1}{1-a} \log \left[\int_X f_X^a(x) dx \right] \quad (2)$$

其中 $a > 0, a \neq 1$ 。

Shannon 熵与 Renyi 熵有以下关系:

$$H_{R_a} \geq H_S \geq H_{R_b}; 0 < a < 1 \text{ 且 } b > 1 \quad (3)$$

$$\lim_{a \rightarrow 1} H_{R_a}(X) = - \int f_X(x) \log f_X(x) dx \quad (4)$$

当 $a \rightarrow 1$ 时, a 阶 Renyi 熵收敛于 Shannon 熵。

当 $a = 2$ 时:

$$H_{R_2} = - \log \int f_X^2(x) dx \quad (5)$$

式(5)称为二次 Renyi 熵,因为概率密度函数是二次式,其对应于二项概率分布。文献[8]指出这种交替定义的 Renyi 熵等价于基于最大化熵的 Shannon 熵。

文献[9]指出连续变量的二次 Renyi 熵可以通过非参数法即带核函数的 Parzen 窗来估计;文献[10]中证明了当二次 Renyi 度量结合使用高期核的 Parzen 窗时能大大节省计算量。

假设一数据集 $\chi = \{x_i\}_{i=1}^N$, 高斯核密度估计为:

$$\hat{f}(x) \propto \frac{1}{N} \sum_{i=1}^N G(x - x_i, \sigma I) \quad (6)$$

其中 $G(x, \sigma I)$ 是在 X 上计算的一个高斯核,并且协方差矩阵是对角方阵,根据二次 Renyi 熵的定义(5)得:

$$\begin{aligned} \hat{H}_{R_2}(X) &= - \int \log f_X^2(x) dx = \\ &= - \log \int_X \left(\sum_{k=1}^N \sum_{j=1}^N G(x - x_k, \sigma I) G(x - x_j, \sigma I) \right) dx = \\ &= - \log \sum_{k=1}^N \sum_{j=1}^N G(x_k - x_j, \sqrt{2}\sigma I) + \text{const} \end{aligned} \quad (7)$$

其中 $\hat{\cdot}$ 代表估值。

因此二次 Renyi 熵可以用各个局部自交叉和来估计,就如核定义的那样。又因为 Renyi 熵具有对称性,所以实际应用中只要估计其一半就行了。

1.2 互信息及其估计的背景知识

两个随机变量 X 与 Y , 它们的互信息可根据它们的概率密度函数 $f(x)$, $f(y)$ 和 $f(x, y)$ 来定义:

$$I(X; Y) = \iint f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy \quad (8)$$

互信息与熵的关系如下:

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) = \\ &= H(X) + H(Y) - H(X, Y) \end{aligned} \quad (9)$$

其中 $H(\cdot)$ 为各个变量的熵,可以是一维变量也可是多元变量, $H(\cdot|\cdot)$ 为条件熵其定义如下:

$$H(X|Y) = \int f(y) H(X|Y=y) dy \quad (10)$$

图1给出了互信息与熵的关系^[11]

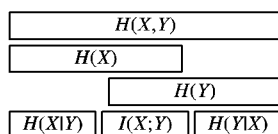


图1 互信息与熵的关系

Shannon 熵的估计计算量比较大,如果采用非参数法^[12]计算复杂度为 $O(N^2)$,其中 N 为样本的数量。

在文献[10]中,式(9)二次 Renyi 熵替代了 Shannon 熵,这种替代产生了以下互信息的估计:

$$I(X; Y) \approx H_{R_2}(X) + H_{R_2}(Y) - H_{R_2}(X, Y) \quad (11)$$

由于二次 Renyi 熵的非参数估计复杂度为 $O(N)$,所以使得互信息计算复杂度由原来的 $O(N^2)$ 下降到了 $O(N)$ 。

2 特征选择算法

基于互信息的特征选择标准有很多,例如:MIFS、mRMR 和 NMIFS。然而这些互信息的估计都是基于 Shannon 熵的,这就导致了非线性的计算复杂度。因此有必要采用更加合适的估计方法。而 NMIFS 能够避免因为局部的一些大值而影响计算的精度。

本文基于 NMIFS 提出一种基于二次 Renyi 熵的正则化互信息方法来对互信息进行估计,在每个样本上使用一个带核函数的 Parzen 窗,然后结合嵌入式方法得到一个两阶段的特征选择算法。

2.1 基于二次 Renyi 熵特征选择的互信息

给定输入数据 χ , 含 N 个样本,每个样本有 M 个属性 $F = \{f_i, i = 1, \dots, M\}$, 目标类别变量 c , 特征选择问题就是要找出最能刻画 c 的一个含有 m 个特征 $\{f_i\}$ 的特征子集 S 。

(NMIFS)^[13]第 m 个特征选择公式如下:

$$\max_{x_j \in F - S_{m-1}} \left[NI(f_j, c) - \frac{1}{m} \sum_{f_i \in S_{m-1}} NI(f_j, f_i) \right] \quad (12)$$

其中:

$$NI(f_j, f_i) = \frac{I(f_j, f_i)}{\min\{H(f_i), H(f_j)\}} \quad (13)$$

结合式(6)和(7)可得:

$$\hat{H}_{R_2}(X, Y) = - \log \sum_{i=1}^N \sum_{j=1}^N G(x_i - y_j, \sqrt{2}\sigma I) \quad (14)$$

从式(7), (11), (14)式可得:

$$\begin{aligned} \hat{I}(X; Y) &= - \log \sum_{i=1}^N \sum_{j=1}^N G(x_i - x_j, \sqrt{2}\sigma I) - \\ &= \log \sum_{i=1}^N \sum_{j=1}^N G(y_i - y_j, \sqrt{2}\sigma I) + \\ &= \log \sum_{i=1}^N \sum_{j=1}^N G(x_i - y_j, \sqrt{2}\sigma I) \end{aligned} \quad (15)$$

结合式(12)~(15)得到本文所提出的特征选择的互信息方法,称为基于二次 Renyi 熵特征选择的正则化互信息。

2.2 两阶段特征选择方法

为了决定最佳的特征个数,本文提出一个两段式特征选择算法:第一阶段使用基于二次 Renyi 熵的 NMIFS 方法找到一个候选特征集;第二阶段用更加先进的方案从候选集中找出一个精选子集。

首先选择一个贪心选择算法,在原始的特征集上找到一个错误相对小并且范围稳定的子集作为候选特征集合 Ω , 整个过程包括以下步骤。

1) 初始化: 给出特征集 $F = \{f_i, i = 1, 2, \dots, M\}$, $S = \emptyset$ 为空集。

2) 选出第一个特征 $\hat{f}_i = \max_{i=1, \dots, M} I(f_i, c)$:

a) 对每个特征 $f_i \in F$, 计算 $I(f_i, c)$;

b) 找出最大的 \hat{f}_i , 使得 $S \leftarrow \hat{f}_i, F \leftarrow F \setminus \{\hat{f}_i\}$ 。

3) 进行贪心选择: 重复直到 $|S| = k$, 其中 $|S|$ 为候选集合 S 的维度。

a) 对所有的 (f_i, f_s) 计算 $I(f_i, f_s)$, 其中 $f_i \in F, f_s \in S$ 且这一对互信息没有被计算过;

b) 从 F 中选择满足式(12)的 \hat{f}_i 并设置 $S \leftarrow \hat{f}_i, F \leftarrow F \setminus \{\hat{f}_i\}$ 。

接着使用嵌入式方法来找到一个更优的子集。一个嵌入式方法就是包含一个特殊分类器的特征选择器。通常嵌入式方法以高计算复杂度代价来得到一个高精度的分类。第一阶段使用基于二次 Renyi 熵互信息的特征选择方法来获得一个较小的候选特征集,在此基础上可以花费较小的代价来实现嵌入式方法。其算法如下:

1) 从 S 中选择一个特征 H_1 , H_1 应是能够得到最佳分类精度的 S 中的单个特征,设置 $P \leftarrow \{H_1\}, S \leftarrow S \setminus \{H_1\}$;

2) 重复 1) 直到分类精度 $a_k > a_{k+1}$ 。

从 S 中选出特征 H_{k+1} , 使得 $P + \{H_{k+1}\}$ 能够获得最佳分类精度。

3 实验结果及分析

通过仿真实验测试了新的特征选择方法性能(计算速度与分类准度)。实验分为两个部分:一个在 UCI 森林覆盖类型数据集^[13]上比较了 MIFS、mRMR、NMIFS 与本文方法的性能差别;另一个在 Alizadeh 主页上的 Lymphoma (LYM) 数据集上做了相同的比较实验^[14]。森林覆盖类型是一个超复杂的分类问题,它也是 UCI 数据集中最大的一个数据集,该数据集集中有 7 种森林类型,每组数据表示一个 30 m × 30 m 森林单元,如表 1 所示。森林覆盖类型数据从美国森林管理局 (USFS) 的资源信息系统 (RIS) 中获得,总共有 581012 个观察实例,每个实例有 54 个属性,在 54 个属性中 10 个连续变量,44 个离散变量。本实验是基于 44 个离散属性的。LYM 数据集有 96 个样本,4026 个基因特征,9 个类型,特征均为连续变量,类型分布如表 2 所示。

表 1 森林覆盖数据类型分布

样本数	比例/%	类别
211840	36.4	1(Spruce-Fir)
283301	48.8	2(Lodge pole)
35754	6.1	3(Ponderosa)
2747	0.5	4(Cottonwood)
9493	1.6	5(Aspen)
17367	2.9	6(Douglas-Fir)
20510	3.5	7(Krummholz)

表 2 Lymphoma 数据类型分布

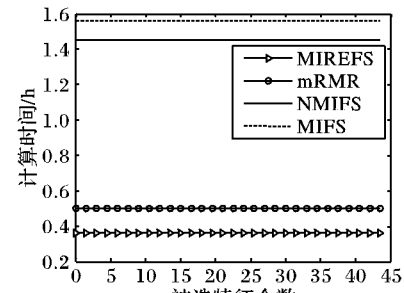
样本数	比例/%	类别
46	47.9	1(DLBCL)
11	11.5	2(CLL)
10	10.4	3(ABB)
9	9.4	4(FL)
6	6.2	5(RAT)
6	6.2	6(TCL)
4	4.2	7(RBB)
2	2.1	8(GCB)
2	2.1	9(LNT)

3.1 计算复杂度分析

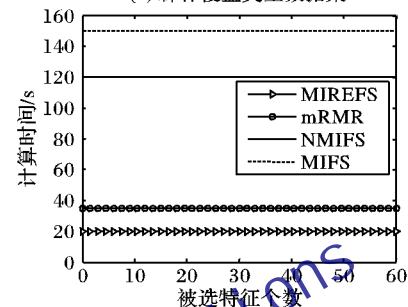
通过实验比较了这几种方法在这两个数据集上特征选择平均计算时间。

图 2 结果表明,对于两个不同的数据集,四种方法所用时间都不随被选择的特征个数而变化,但每种方法所花费的时

间不同,MIREFS 最少,mRMR 其次,NMIFS 与 MIFS 相近,实验表明 MIREFS 计算效率要高于其他几种方法。图 2(a)的横坐标对应着离散属性,0 对应着所有特征均为连续特征。



(a) 森林覆盖类型数据集

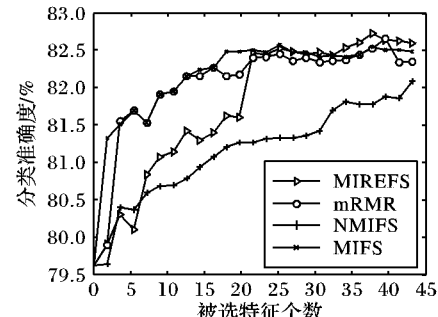


(b) Lymphoma 数据集

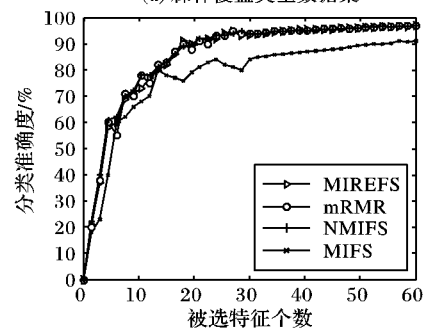
图 2 特征选择时间与特征数之间的关系

3.2 分类精度分析

本文所提出的特征选择方法不指定某一种特殊的分类器,该方法在任何类型分类器上都能获得好性能。在此鉴于篇幅有限只考虑以支持向量机^[15]作为分类器。本实验使用支持两个分类或多个分类的 LIBSVM 包^[16]作为分类器,结果如图 3 所示。



(a) 森林覆盖类型数据集



(b) Lymphoma 数据集

图 3 分类准确度与被选特征数间的关系

图 3(a) 的横坐标对应着离散属性,0 对应着所有特征均为连续特征。它表明对于森林覆盖类型数据集而言,连续属性对分类准确度起决定性作用。从图中可看到,当所选离散特征数在 1~20, NMIFS 和 MIFS 的分类准确度要略高于 MIREFS;当所选特征数更多时(例如:大于 30 时) MIREFS 能

获得更高的分类准确度。仅有 mRMR 始终分类准确度最低,但所有方法的分类准确度都非常接近,偏差不超过 2%。图 3(b)中,四种方法中只有 MIFS 比其他三种方法稍差,而 MIREFS、mRMR、MIFS 三种方法分类准确度非常接近。因此对各种特征选择方法来说计算时间就变得更加重要了。

4 结语

本文提出一种基于二次 Renyi 熵互信息特征选择方法,该方法由于在互信息估计上的高效性而减少了 NMIFS 计算复杂度。同时由于各种分类方法的分类精度非常接近,差别不大,因此时间复杂度成为衡量各种方法的重要指标。实验表明 MIREFS 相比其他方法具有更高的效率。

参考文献:

- [1] BATITTI R. Using mutual information for selecting features in supervised neural net learning [J]. IEEE Transactions on Neural Networks, 1994, 5(4): 537 - 550.
- [2] KWAK N, CHOI C H. Input feature selection for classification problems [J]. IEEE Transactions on Neural Networks, 2002, 3(1): 143 - 159.
- [3] PENG H, LONG F, DING C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226 - 1238.
- [4] KWAK N, CHOI C H. Input feature selection for classification problems [J]. IEEE Transactions on Neural Networks, 2002, 3(1): 143 - 159.
- [5] ESTÉVEZ P A, TESMER M, PEREZ C A. Normalized mutual information feature selection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 20(2): 189 - 201.
- [6] HILD K E, ERDOGMUS D, TORRKOLA K, *et al.* Feature extraction using information theoretic learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(9): 1385 - 1392.
- [7] BONEV B, ESCALANO F, CAZORLA M. Feature selection, mutual information, and the classification of high-dimensional patterns [J]. Pattern Analysis and Applications, 2008, 11(3/4): 309 - 319.
- [8] KAPUR J N. Measures of information and their applications [M]. [S. l.]: Wiley-Interscience, 1994.
- [9] PRICIPE J C, XU D, FISHER J W. Information theoretic learning [M]// Unsupervised Adaptive Filtering. New York: Wiley, 2000: 265 - 319.
- [10] RENYI A. Probability theory [M]. Amsterdam: North - Holland Publishing Company, 1970.
- [11] HU BAO-GANG, WANG YONG. Evaluation criteria based on mutual information for classifications including rejected class [J]. 自动化学报, 2008, 34(11): 1396 - 1403.
- [12] HILD K E, II, ERDOGMUS D, PRICIPE J C. An analysis of entropy estimators for blind source separation [J]. Signal Processing, 2006, 86(1): 182 - 194.
- [13] BLAKE C, MERZ C. UCI repository of machine learning databases [EB/OL]. [2009 - 09 - 22]. <http://archive.ics.uci.edu/ml/datasets/Coverttype>.
- [14] ALIZADEH A A. Lymphoma/Leukemia molecular profiling project [EB/OL]. [2009 - 09 - 22]. <http://llmpp.nih.gov/lymphoma/>.
- [15] VAPNIK V N. The nature of statistical learning theory [M]. Berlin: Springer-Verlag, 1995.
- [16] HSU C W, LIN C J. A comparison of methods for multi-class support vector machines [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415 - 425.

(上接第 1258 页)

本文中的点的局部特征的计算还比较复杂,不能满足实时准确定位要求。在接下来的工作中,我们将致力于研究结合其他的几何信息和二维信息来降低运算量,并解决侧面情况下的定位问题,以实现任意姿态和表情下的更鲁棒快捷的人脸特征点定位。

参考文献:

- [1] XU C H, TAN T N, WANG Y H, *et al.* Combining local features for robust nose location in 3D facial data [J]. Pattern Recognition Letters, 2006, 27(13): 1487 - 1494.
- [2] SALAH A A, AKARUN L. 3D facial feature localization for registration [C]// International Workshop on Multimedia Content Representation, Classification and Security, LNCS 4105. Berlin: Springer-Verlag, 2006: 338 - 345.
- [3] SALAH A A, ÇINAR H, AKARUN L, *et al.* Robust facial landmarking for registration [J]. Annals of Telecommunications, 2007, 62(12): 1608 - 1633.
- [4] DİBEKLIOĞLU H, SALAH A A, AKARUN L. 3D facial landmarking under expression, pose, and occlusion variations [C/OL]// IEEE Second International Conference on Biometrics: Theory, Applications and Systems. Washington, DC: IEEE Press, 2008 [2009 - 05 - 04]. <http://www.cmpe.boun.edu.tr/~dibeklioglu/documents/Dibeklioglu2008btas.pdf>.
- [5] AKAKIN H Ç, SALAH A A, AKARUN L, *et al.* 2D/3D facial feature extraction [C]// Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning, SPIE 6064. San Jose, USA: SPIE, 2006: 441 - 452.
- [6] CONDE C, CIPOLLA R, RODRÍGUEZ-ARAGÓN L J. 3D facial feature location with spin images [C]// MVA 2005: IAPR Conference on Machine Vision Applications. Tsukuba Science City, Japan: [s. n.], 2005: 418 - 421.
- [7] WANG Y, CHUA C-S, HO Y-K. Facial feature detection and face recognition from 2D and 3D images [J]. Pattern Recognition Letters, 2002, 23(10): 1191 - 1202.
- [8] CANTZLER H, FISHER R B. Comparison of HK and SC curvature description methods [C]// 3DIM '01: Third International Conference on 3D Digital Imaging and Modeling. Washington, DC: IEEE Computer Society, 2001: 285 - 291.
- [9] KOENDERINK J J, van DOORN A J. Surface shape and curvature scales [J]. Image and Vision Computing, 1992, 10(8): 557 - 565.
- [10] WU ZHAOHUI, WANG YUEMING, PAN GANG. 3D face recognition using local shape map [C]// ICIP 2004: IEEE International Conference on Image Processing. Washington, DC: IEEE Press, 2004: 2003 - 2006.
- [11] COLBRY D, STOCKMAN G, JAIN A. Detection of anchor points for 3D face verification [C]// CVPR'05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2005: 118.
- [12] YAN Y, CHALLAPALI K. A system for the automatic extraction of 3D facial feature points for face model calibration [C]// IEEE International Conference on Image Processing. Washington, DC: IEEE Press, 2000: 223 - 226.
- [13] 三维人脸数据库 [DB/OL]. [2009 - 10 - 22]. <http://www.cbbsr.ia.ac.cn/china/3DFace%20Databases%20CH.asp>.