文章编号: 1001-0920 (2013) 06-0915-05

基于互信息的主成分分析特征选择算法

范雪莉, 冯海泓, 原 猛

(中国科学院声学研究所 东海研究站，上海 200032)

摘 要: 主成分分析是一种常用的特征选择算法, 经典方法是计算各个特征之间的相关, 但是相关无法评估变量间的非线性关系. 互信息可用于衡量两个变量间相互依赖的强弱程度, 且不局限于线性相关, 鉴于此, 提出一种基于互信息的主成分分析特征选择算法. 该算法计算特征间的互信息, 以互信息矩阵的特征值作为评价准则确定主成分的个数, 并衡量主成分分析特征选择的效果. 通过实例对所提出方法和传统主成分分析方法进行比较, 并以神经网络为分类器分析分类效果.

关键词: 互信息；主成分分析；特征选择

中图分类号: TP391.4 **文献标志码:** A

PCA based on mutual information for feature selection

FAN Xue-li, *FENG Hai-hong*, *YUAN Meng*

(Shanghai Acoustics Laboratory，Chinese Academy of Sciences Bionic Ear and Sound Technology Laboratory, Shanghai 200032，China．Correspondent：FAN Xue-li，E-mail：fanxueli@mail.ioa.ac.cn)

Abstract: Principal component analysis(PCA) is a common method for feature selection. The classical procedure to obtain principal components is calculating the correlation matrix between features. However, the correlation cannot reflect the nonlinear relationship. Mutual information measures the interdependence strength between variables which are not limited to the linear correlation. PCA based on mutual information(MIPCA) for feature selection is presented. The algorithm calculates the mutual information matrix and extracts the eigenvalues as the criteria to determine the number of principal components and assess the effect of feature selection. Finally, the proposed algorithm is compared with PCA by cases, and the efficiency of classification is tested by neuron network.

Key words: mutual information；principal component analysis；feature selection

0 引 言

特征选择[1]的基本任务是从众多特征中找出能反映不同类别之间差异的那些有效特征. 作为分类问题的预处理, 特征选择广泛应用于数据挖掘[2]、模式识别[3]、机器学习和人工智能[4]等领域. 随着信息科学与计算机技术的发展、应用领域的拓宽和相关研究成果的不断涌现, 特征选择越来越受到人们的重视. 目前, 信息获取水平逐步提高, 存储能力不断增强, 可获得的数据量越来越大, 数据的维数也越来越高, 这些新情况一方面对后端分类器的设计与训练带来挑战, 如随着特征维数的增高带来的“维数灾难”[5], 使得分类效果恶化, 计算复杂度大幅度提升, 所需样本数量急剧增加; 另一方面, 特征之间的冗余造成虚假信息, 或者特征集包含与分类无关的特征进而造成计算量的提高. 因此, 如何更有效地进行特征选择是目前研究的一个热点.

特征选择算法主要包括两类: 第 1 类是组合优化方法, 如搜索树[6]、遗传算法[7]等, 这些算法在每一次搜索迭代过程中, 通过可分性判据评价当前特征选择的优劣; 第 2 类是转换法, 如主成分分析[8]、独立成分分析[9]等, 这些算法大都是将高维的特征空间进行线性转化, 压缩至低维特征空间.

主成分分析是目前应用较为广泛的特征选择方法, 最早由 Pearson 等提出, 将数据集从原始空间转换到主成分空间, 主成分空间内每一个主成分表示转换后有效的新特征. 计算数据集的协方差矩阵或相关矩阵, 利用矩阵的特征值确定主成分的维数, 利用矩阵的特征向量确定主成分方向. 这种方法意义明确, 易

收稿日期: 2012-02-09；**修回日期:** 2012-06-01.

基金项目: 国家自然科学基金青年科学基金项目(11104316)；上海市自然科学基金项目(11ZR1446000).

作者简介: 范雪莉(1984−), 女, 博士, 从事信号与信息处理的研究；冯海泓(1966−), 男, 研究员, 博士生导师, 从事声学与信息处理、听觉康复等研究.

于操作. 一般进行主成分分析之前先对数据进行标准化预处理, 消除变量之间由于量纲不同造成的差异, 因此计算变量之间的协方差即计算变量之间的相关. 但是相关只能反映变量之间的线性关系, 对于非线性相关无法进行度量. 而互信息衡量两个变量间相互依赖的强弱程度, 表示两个变量间共同拥有信息的含量, 且这种度量不局限于线性关系, 对于变量之间的非线性关系也能进行评估. 互信息属于信息度量的范畴, 主要利用信息熵[10]等量化特征相对于分类类别的不确定性程度来判定其包含的类别信息. 它是一种无参的、非线性的标准, 目前在特征选择领域引起了广泛的关注[11-14].

本文针对传统的主成分分析特征选择算法中相关矩阵只能衡量变量之间线性关系的局限性, 提出一种基于互信息的主成分分析法用于特征选择. 首先根据样本数据, 得到特征的概率分布; 然后结合信息论理论, 计算特征的自信息和特征之间的互信息, 得到互信息矩阵后计算其特征值与特征向量, 其中特征向量表征新的特征空间——基于互信息的主成分空间内每一个主成分的方向, 特征值与总体特征值之和的百分比表征对应主成分占总体主成分信息量的比重, 以此作为选择主成分的维度的标准. 该算法从互信息的角度计算主成分, 可以提供更多特征之间的关系信息, 转换后的主成分较传统的主成分所需维数降低, 且保留了大部分原始信息. 通过实例分析, 验证了这一方法的有效性.

1 主成分分析特征选择

主成分分析 (PCA) 是一种以 K-L 变换[15]为基础的统计分析方法, 其基本思想是对一个维数较高、各维变量之间相互关联的数据集进行降维处理, 且降维后数据集尽量保留原始信息. PCA 通过线性转换, 将原始空间转换到低维的主成分空间, 转换后的新特征称为主成分, 满足各个主成分之间不相关, 并按照其对应方向的方差贡献率降序排列.

假设特征空间 $\boldsymbol{R}^n$ 上的样本数据集 X, 每一个数据 x 由 n 维特征变量组成, 即 $(x_1, x_2, \cdots, x_n)$. PCA 对 x 进行线性变换, 转换到主成分空间 $\boldsymbol{R}^n$ 内的数据 y, 其中每一个数据 y 由 m 维主成分组成, 即 $(y_1, y_2, \cdots, y_m)$, 有

$$y = A'x. \tag{1}$$

y 中第 k 个元素 y_k 表示第 k 个主成分, $k=1, 2, \cdots, m$, $m \leqslant n$. 除非特殊说明, 一般要求第 1 个主成分 y_1 表示在此主成分方向上数据的方差最大, 第 2 个主成分 y_2 表示在此主成分方向上数据的方差次大, 且第 2 主成分 y_2 与第 1 主成分 y_1 无关. 以此类推.

由式 (1) 中矩阵 A 的第 k 列向量 α_k 可得到

$$y_k = \alpha_k' x. \tag{2}$$

根据主成分的要求, 考虑第 1 主成分 $\alpha_1' x$ 满足此方向上的方差最大化要求 $\mathrm{var}[\alpha_1' x]$, 则有

$$\mathrm{var}[\alpha_1' x] = \alpha_1' \Sigma \alpha_1, \tag{3}$$

其中 Σ 表示 x 的协方差, 有时也用互相关代替. 由于满足此要求的向量 α_1 不唯一, 需要添加一定限制, 一般要求 $\alpha_1'\alpha_1 = 1$. 注意到, 其他限制如 $\max|\alpha_1| = 1$ 得出的结果是不同于主成分的.

利用拉格朗日因子法求解 α_1, 最大化 $\alpha_1'\Sigma\alpha_1 - \lambda(\alpha_1'\alpha_1 - 1)$, 可以推导得出

$$(\Sigma - \lambda I_n)\lambda_1 = 0,$$

其中 I_n 是 $n \times n$ 的单位阵. 因此, λ 是此主成分方向 α_1 上的方差, 且 λ 是 Σ 最大的特征值, α_1 是对应的特征向量. α_2 的求解与 α_1 类似, 再加上一个限制, $\alpha_1' x$ 与 $\alpha_2' x$ 无关, 即

$$\begin{aligned}&\mathrm{cov}[\alpha_1' x, \alpha_2' x] = \\ &\alpha_1'\Sigma\alpha_2 = \alpha_2'\Sigma\alpha_1 = \alpha_2'\lambda_1\alpha_1 = \lambda_1\alpha_2'\alpha_1 = 0,\end{aligned} \tag{4}$$

即 $\alpha_2'\alpha_1 = 0$. 利用拉格朗日因子法得到 α_2 为 Σ 次大的特征值对应的特征向量. 以此类推, 各个主成分为降序排列的 Σ 的特征值对应的特征向量, 因此有

$$A'\Sigma A = \Lambda. \tag{5}$$

其中: A 是由 Σ 的特征向量为列向量组成的矩阵, Λ 是由 Σ 的特征值组成的对角阵.

在主成分分析前, 一般先对数据进行标准化预处理, 消除不同量纲之间的影响. 经过处理后的数据进行协方差运算等效于进行相关运算. 主成分维数由前 m 个主成分对应特征值之和所占总体特征值之和的比重确定, 一般选择 85%~95%.

2 基于互信息的主成分分析特征选择算法

2.1 互信息

信息是认识主体所感受和所表达的事物运动的状态和运动状态变化的方式[16]. 信息的度量方式有两种, 一种是对消息或消息集合本身所含信息量多少的度量, 一种是对消息之间或消息集合之间相互提供信息量多少的度量. 前者用自信息和消息熵 (自信息的平均值) 来描述, 后者用互信息和平均互信息来描述.

互信息衡量两个变量间相互依赖的程度, 表示两个变量间共同拥有信息的含量. 给定两个随机变量 X 和 Y, 若它们各自的边缘概率分布和联合概率分布分别为 $p(x), p(y)$ 和 $p(x, y)$, 则它们之间的互信息 $I(X;Y)$ 定义为

$$I(X;Y)=\sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \tag{6}$$

当变量 X 和 Y 完全无关或相互独立时, 互信息最小, 结果为 0, 这意味着两个变量之间不存在重叠的信息; 反之, 两者的相互依赖程度越高, 互信息的值越大, 所包含的相同信息也越多.

2.2　基于互信息的主成分分析特征选择算法

相关或相关系数只能反映两个变量之间的线性相关, 无法衡量两个变量之间的非线性关系. 而互信息从信息论的角度出发, 可以评估变量间共有信息量的多少, 并且不局限于线性关系, 与相关比较有很大优势. 因此, 考虑在主成分分析中用互信息来替代协方差或相关, 提出一种基于互信息的主成分分析 (MIPCA) 特征选择算法. 将式 (7) 改写为

$$B'\Sigma_{I_{XY}}B=\Lambda. \tag{7}$$

其中: $\Sigma_{I_{XY}}$ 为数据集的互信息矩阵, B 为特征向量 $(\beta_1,\beta_2,\cdots,\beta_n)$ 组成的矩阵, Λ 为特征值 $(\mu_1,\mu_2,\cdots,\mu_n)$ 组成的对角阵.

下面判断式 (9) 是否有解和解的特点. $\Sigma_{I_{XY}}$ 中对角线元素表示变量的自信息, 即变量的信息熵, 非对角线元素表示两个变量之间的互信息. 无论互信息或信息熵均为实数, 当两个变量之间不相关时, 互信息为 0, 否则为正数, 因此 $\Sigma_{I_{XY}}$ 为非负实数阵. 互信息满足

$$I(X,Y)=I(Y,X),$$

可得到 $\Sigma_{I_{XY}}$ 为非负实数对称阵, $\Sigma_{I_{XY}}$ 的特征值为实数, Λ 为实数对角阵, 对应的特征向量两两正交, 矩阵 B 为正交阵.

基于互信息的主成分为

$$z=B'x. \tag{8}$$

主成分 $z_k=\beta_k'x(k=1,2,\cdots,n)$, 且两两正交, 满足主成分要求, β_k 是主成分 z_k 的转换系数, 简称主成分 k 的系数.

然后判断主成分的维数 m. 定义 MIPCA 的主成分贡献率 δ_k 为单一主成分占总体主成分信息量的比率, 即

$$\sigma_k=\mu_k\Big/\sum_{k=1}^{n}\mu_k, \tag{9}$$

其中 μ_k 为式 (9) 中互信息矩阵 $\Sigma_{I_{XY}}$ 第 k 大的特征值, 表示主成分 k 的信息量. 定义 MIPCA 的主成分累积贡献率 δ_k 为前 k 个主成分的贡献率之和, 有

$$\delta_k=\sum_{i=1}^{k}\sigma_i. \tag{10}$$

选择贡献率之和为 85%∼95% 的前 m 个主成分作为新的特征.

3　实验分析

3.1　数据库

选择 Iris 鸢尾花数据集, 此数据集是数据挖掘和模式识别中常用的数据集, 每一个鸢尾花数据包含 4 个特征: 花瓣长度、花瓣宽度、花萼长度和花萼宽度, 数据集共有 150 个鸢尾花数据, 分为 setosa, versicolor 和 virginica 3 类, 每类 50 个数据.

3.2　实验分析

实验 1　比较两维数据情况下, PCA 和 MIPCA 计算得到的主成分方向的异同.

只选择两维数据进行比较是为了清楚地看出两种方法得到的主成分方向的差异, 利用两维数据进行计算, 主成分方向可以展示在二维图形上, 可视化的结果可以进行较为直观的比较.

实验 1 采用鸢尾花的花瓣长度和花瓣宽度 2 维特征, 数据个数不变, 进行 PCA 和 MIPCA 特征选择. 表 1 为 PCA 和 MIPCA 各自得到的各自主成分系数 β (为了表示方便, PCA 的主成分系数在本实验和后续实验中均用 β 表示, 而非 α) 和累积贡献率 δ. 将表 1 中的主成分方向和两维鸢尾花数据绘制在二维图中, 如图 1 所示. 为了清楚地看出主成分方向, 将主成分方向长度增加, 方向保持不变. 图 1 显示的 3 类鸢尾花分别由不同标识符表示, 根据这些数据, 利用 PCA 和 MIPCA 两种特征选择方法计算得出各自的主成分方向, 同时也将这些主成分方向绘制在图 1 中, 并标出各自主成分方向 1.

表 1　PCA 和MIPCA 得到的主成分系数β和累积贡献率δ

算法	主成分 1		主成分 2	
	β_1	δ_1/%	β_2	δ_2/%
PCA	[0.707 1, 0.707 1]	98.1	[−0.707 1, 0.707 1]	100
MIPCA	[0.695 1, 0.718 9]	94.6	[−0.718 9, 0.695 1]	100

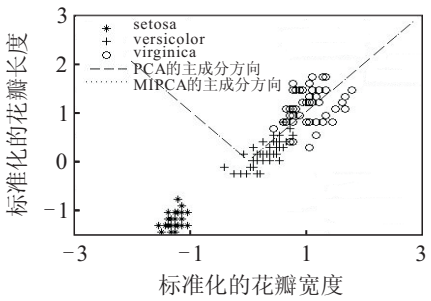

图 1　鸢尾花数据, MIPCA 和 PCA 得到的主成分方向

从图 1 可以看出, 相对于原来的两维特征, 利用 PCA 得到的主成分方向 1 或 MIPC A得到的主成分方向 1 能够更好地区分 3 类数据, 实现了数据的降维处理. 经过计算, MIPCA 得到的主成分 1 与 PCA 得到的

主成分1之间的夹角为0.97°, 约等于方向相同. 并且由表1得到PCA和MIPCA的主成分1的累积贡献率分别为98.1%和94.6%, 相差不大. 因此可以说, 数据集X转换到基于互信息的主成分空间后的数据集Z与X转换到主成分空间后的数据集Y之间近似相等. 因此在两维情况下, 基于互信息的主成分分析特征选择算法与传统的主成分分析特征选择算法结果相似.

实验2 比较多维数据情况下, PCA和MIPCA计算得到的主成分方向的异同.

选择Iris鸢尾花数据集的全部特征进行PCA和MIPCA特征选择, 得到两种方法各自的主成分系数β、主成分方向的信息μ、主成分贡献率σ和累积贡献率δ.

表2给出了经过PCA和MIPCA特征选择后的主成分系数β和累积贡献率δ. 由表2可见, MIPCA主成分1的贡献率明显大于PCA主成分1的贡献率, 前两维主成分累积贡献分别为95.81%和97.61%, 维数相同的情况下, MIPCA得出的主成分贡献更大, 若以85%的累积贡献率为限选择主成分维数, 则MIPCA特征选择方法仅需1维, 而PCA方法则需要2维.

表2 PCA和MIPCA得到的主成分的β和δ

	主成分	β	δ/%
PCA	1	[0.521 1, −0.269 3, 0.580 4, 0.564 9]	72.96
	2	[−0.377 4, −0.923 3 − 0.024 5, −0.066 9]	95.81
	3	[0.719 6, −0.244 4, −0.142 1, −0.634 3]	99.48
	4	[0.261 3, −0.123 5, −0.801 4, 0.523 6]	100
MIPCA	1	[−0.296 7, −0.243 8, −0.386 5, −0.838 5]	86.52
	2	[−0.525 2, −0.469 4, −0.464 7, 0.536 6]	97.61
	3	[0.790 9, −0.300 1, −0.530 7, 0.051 2]	99.65
	4	[0.102 7, −0.793 8, 0.594 1, −0.079 3]	100

实验3 以多层感知器(MLP)神经网络为分类器, 比较PCA和MIPCA得到的主成分的分类正确率.

MLP含有1个隐层, 隐层内含有4个神经元. 隐层和输出层的传递函数分别为对数-S形(logsid)函数和线性函数. 以不同维度的主成分作为输入, 输入层神经元个数与主成分维度一致. 输出层含有3个神经元, 与鸢尾花的类别数量一致. MLP的训练采用后向传播(BP)算法, 以最小均方误差为准则. 分类正确率如表3所示.

表3 分类正确率 %

算法	主成分维度			
	1	2	3	4
PCA	78	92	98	98
MIPCA	86	94	98	98

由表3可见, 维数相同的情况下, 采用MIPCA计算出的主成分作为神经网络输入得到的分类正确率大于等于PCA方法. 结合表2和表3, 以两维主成分为例, 表2表明MIPCA方法计算得到的2维主成分所含的累积贡献率大于PCA方法得到的结果; 由表3可知, 同时以神经网络作为分类器时, MIPCA得到的2维主成分为特征, 得到的分类正确率也大于PCA得到的两维主成分的分类正确率. 实验表明, 以式(11)和(12)为依据判断选择的主成分所包含的信息是合理的, 且采用MIPCA方法得到的累积贡献率更大, 分类正确率更高.

4 结 论

本文描述了基于互信息的主成分分析特征选择算法, MIPCA利用数据的互信息矩阵取代PCA中的协方差或相关矩阵计算主成分方向, 并确定主成分维度. 通过实验分析了基于互信息的主成分分析特征选择算法的性能. 实验表明: 1) 在两维情况下, 主成分分析与基于互信息的主成分分析得到的主成分方向相同; 2) 以主成分累积贡献率为准则判断主成分维数, 基于互信息的主成分特征选择算法较之主成分分析能够得到更少的主成分维度, 降低了特征维数, 以便减少后端分类或识别的计算量, 若选择主成分维数相同, 则基于互信息的主成分特征选择方法得到的主成分的累积贡献率更高; 3) 以神经网络为分类器, 同样维数的主成分, 采用基于互信息的主成分分析较之传统主成分分析特征选择算法能够得到更高的分类正确率.

参考文献(References)

[1] 边肇祺, 张学工. 模式识别[M]. 北京: 清华大学出版社, 2000: 176-177.
(Bian Z Q, Zhang X G. Pattern recognition[M]. Beijing: Tsinghua University Press, 2000: 176-177.)

[2] Witten I, Frank E. Data mining: Practical machine learning tools and techniques[M]. San Francisco: Morgan Kaufmann, 2005: 39-52.

[3] Bishop C. Pattern recognition and machine learning[M]. New York: Springer, 2006: 1-58.

[4] Russell S, Norvig P. Artificial intelligence: A modern approach[M]. New Jersey: Prentice Hall, 2010: 31-44.

[5] Bellman R. Adaptive control processes: A guided tour[M]. Princeton: Princeton University Press, 1966: 152-175.

[6] Donald E. The art of computer programming[J]. Sorting and Searching, 1999, 3: 426-458.

[7] Goldberg D. Genetic algorithms in search, optimization and machine learning[M]. New York: Addison-wesley, 1989: 41.

[8] Jolliffe I. Principal component analysis[M]. New York: Springer-Verlag, 1986: 10-28.

[9] Kwon O, Lee T. Phoneme recognition using ICA-based feature extraction and transformation[J]. Signal Processing, 2004, 84(6): 1005-1019.

[10] Shannon C. A mathematical theory of communication[J]. ACM Sigmobile Mobile Computing and Communications Review, 2001, 5(1): 3-55.

[11] Battiti R. Using mutual information for selecting features in supervised neural net learning[J]. IEEE Trans on Neural Networks, 1994, 5(4): 537-550.

[12] Kwak N, Choi C. Input feature selection for classification problems[J]. IEEE Trans on Neural Networks, 2002, 13(1): 143-159.

[13] Yang H, Moody J. Feature selection based on joint mutual information[C]. Proc of Int ICSC Symposium on Advances in Intelligent Data Analysis. 1999: 22-25.

[14] 唐亮, 段建国, 许洪波, 等. 基于互信息最大化的特征选择算法及应用[J]. 计算机工程与应用, 2008, 44(13): 130-133.
(Tang L, Duan J G, Xu H B, et al. Mutual information maximization based feature selection algorithm in text classification[J]. Computer Engineering and Applications, 2008, 44(13): 130-133.)

[15] Jorgensen P, Song M. Entropy encoding, hilbert space and karhunen-loève transforms[J]. J of Mathematical Physics, 2007, 48(10): 103503.

[16] 田宝玉, 杨洁, 贺志强, 等. 信息论基础[M]. 北京: 人民邮电出版社, 2008: 1-8.
(Tian B Y, Yang J, He Z Q, et al. Foundations of information theory[M]. Beijing: People's Posts and Telecommunications Press, 2008: 1-8.)

(上接第914页)

[4] 魏洁, 李军. EPR 下的逆向物流回收模式选择研究[J]. 中国管理科学, 2005, 13(6): 18-22.
(Wei J, Li J. The choice of different take-back models in reverse logistics with the restriction of EPR[J]. Chinese J of Management Science, 2005, 13(6): 18-22.)

[5] 刑伟, 汪寿阳, 赵秋红, 等. 考虑渠道公平的双渠道供应链均衡策略[J]. 系统工程理论与实践, 2011, 31(7): 1249-1256.
(Xing W, Wang S Y, Zhao Q H, et al. Impact of fairness on strategies in dual-channel supply chain[J]. Systems Engineering－Theory & Practice, 2011, 31(7): 1249-1256.)

[6] 韩小花. 基于制造商竞争的闭环供应链回收渠道的决策分析[J]. 系统工程, 2010, 28(5): 36-41.
(Han X H. Decision analysis of recycling channel of closed-loop supply chain based on manufacturers competition[J]. Systems Engineering, 2010, 28(5): 36-41.)

[7] 李帮义. 作为阻止战略的再制造决策研究[J]. 控制与决策, 2010, 25(11): 1675-1678.
(Li B Y. Decision research on remanufacturing as prevention strategy[J]. Control and Decision, 2010, 25(11): 1675-1678.)

[8] Amaeshi Onyeka K, Osuji, Paul Nnodim. Corporate social responsibility in supply chains of global brands: A boundaryless responsibility Claricatios, exceptions and implications[J]. J of Business Ethics, 2008, 81(1): 223-234.

[9] Ni D, Kevin W L, Tang X W. Social responsibility allocation in two-echelon supply chains: Insights from wholesale price contracts[J]. European J of Operational Research, 2010, 207(3): 1269-1279.

[10] 汪翼, 孙林岩, 李刚, 等. 闭环供应链的回收责任分担决策[J]. 系统管理学报, 2009, 18(4): 378-384.
(Wang Y, Sun L Y, Li G, et al. Shareing take-back responsibility across closed-loop supply chain[J]. J of Systems & Management, 2009, 18(4): 378-384.)

[11] Fruchter G E, Kalish S. Closed-loop advertising strategies in a duopoly[J]. Management Science, 1997, 43(1): 54-63.