基于核稀疏表示的特征选择算法*

邓战涛，胡谷雨，潘志松，张艳艳
（解放军理工大学 指挥自动化学院，南京 210007）

摘　要：为了解决高维数据在分类时导致的维数灾难，降维是数据预处理阶段的主要步骤。基于稀疏学习进行特征选择是目前的研究热点。针对现实中大量非线性可分问题，借助核技巧，将非线性可分的数据样本映射到核空间，以解决特征的非线性相似问题。进一步对核空间的数据样本进行稀疏重构，得到原数据在核空间的一种简洁的稀疏表达方式，然后构建相应的评分机制选择最优子集。受益于稀疏学习的自然判别能力，该算法能够选择出保持原始数据结构特性的“好”特征，从而降低学习模型的计算复杂度并提升分类精度。在标准 UCI 数据集上的实验结果表明，其性能上与同类算法相比平均可提高约 5%。

关键词：特征选择；稀疏表示；核技巧
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1001-3695(2012)04-1282-03
doi:10.3969/j.issn.1001-3695.2012.04.022

Feature selection based on kernel sparse representation

DENG Zhan-tao, HU Gu-yu, PAN Zhi-song, ZHANG Yan-yan
(*Institute of Command Automation, PLA University of Science & Technology, Nanjing* 210007, *China*)

Abstract: In order to solve the problem of dimension illness in classification of high-dimensional data, dimensionality reduction is a key approach in pretreatment. Feature selection based on sparse representation is one of the hottest research topics recently. In the face of the nonlinear problem, this paper motivated by kernel trick, nonlinear data was mapped into kernel space in which the nonlinear similarity of the features could be captured and reconstructed by sparse representation to get concision expression of original data in kernel space. Then it designed evaluate mechanism to select excellent feature subsets. As the natural discriminative power of sparse representation, "good" feature which preserved the original structure would be selected so that feature selection could reduce computational complexity and improve precision. The results of experiment in standard UCI data sets show that the performance compare with similar algorithms improves about the average of 5%.

Key words: feature selection; sparse representation; kernel trick

0　引言

近年来，随着数据采集技术和存储技术的发展，大量高维数据（如生物数据、文本数据、遥感数据等）易于获取，却难以处理。由于维数过高，使用传统的分类方式时往往需要更多的训练样本，导致维数灾难，从而引起 Hughes 现象[1]。然而，降维技术中的特征选择保留了数据的原始特征，依据某种评估准则，选择最优特征子集对数据进行低维表示，保持了数据的内在信息，使得分类器在低维子空间上进行工作，从而有效提高了分类器的效率与预测精度。

特征选择不改变原始空间的性质，过程一般由四个部分组成：产生器、评估准则、停止条件和验证[2]。本文的研究主要着重于评估准则的构建。过滤式模型的评估准则不涉及学习算法，与后者独立，虽然在精度上弱于封装式模型，但有着较高的效率，是当前研究的热点。如 Variance Score[3] 是一种简单的无监督特征选择方法，其基本思想是特征构成的子空间的方差越大越重要；Laplacian Score[4] 也是无监督方法，通过构建无项加权图描述样本间的局部结构，特征子集不仅方差最大，还保持了局部性；Fisher Score[3] 是利用类标号的有监督方法，寻找类间方差尽可能大、类内方差尽可能小的特征子集；Constraint Score[5] 是利用成对约束信息的有监督方法，寻找最能保持样本间约束关系的特征子集；为了利用大量的无标号数据信息，Zhao 等人[6] 提出了基于谱图分析理论的半监督特征选择；刘峤等人[7] 提出了基于零范数特征选择的支持向量机模型，但受限于 NP 难问题的局限性；Sparsity Score[8] 是基于稀疏理论的特征选择算法，保持稀疏重构误差最小。Liang 等人[9] 通过稀疏接近将其运用在人脸识别中的特征选择。

本文借鉴了稀疏表示的思想，通过经验核映射将原始数据投影到线性可分的高维空间中，使得线性不可分的数据样本在高维空间中可能具有更高的区分性。然后对核空间中的高维数据稀疏表示，构建样本稀疏重构矩阵，依据评分机制寻找最优的特征子集。实验表明，基于核稀疏表示（kernel sparse representation，KSR）的特征选择算法在多个 UCI 数据集上取得了良好的分类性能。

收稿日期：2011-09-19；**修回日期**：2011-10-20　　**基金项目**：国家自然科学基金资助项目（60603029）；国家“863”计划资助项目（2010AAJ163）

作者简介：邓战涛（1988-），男，陕西汉中人，硕士研究生，主要研究方向为数据挖掘、人工智能（dengzhantao@sina.cn）；胡谷雨（1963-），男，教授，博士，主要研究方向为网络管理、人工智能；潘志松（1973-），男，副教授，博士，主要研究方向为数据挖掘、人工智能.

1　基于核稀疏表示的特征选择算法

1.1　核稀疏表示

稀疏表示[10]源于信号在过完备字典上进行分解的思想,将信号 $x \in R^M$ 分解为一系列基信号 $X=[x_1,x_2,\cdots,x_n] \in R^{M\times n}$ 的线性组合 $x \approx \sum_{i=1}^{n} s_i x_i$,并希望向量 $s=[s_1,s_2,\cdots,s_n]^{\mathrm{T}}$ 尽可能稀疏。由于 ℓ_0 范数的求解属于 NP 难问题,Donoho[11]研究表明:只要满足所求解足够的稀疏,ℓ_0 和 ℓ_1 范数等价。为了减小重构误差,引入最大错误容忍变量 ε,即有问题 P_1:

$$\min_{s} \| s \|_1$$
$$\text{s.t.} \quad \| x - sX \|^2 < \varepsilon \tag{1}$$

式(1)是一个凸优化问题,使用经典的 ℓ_1-magic 软件包(http://www.public.asu.edu/~jye02/Software/SLEP)求解最小化问题。

为了解决现实中的非线性可分问题,本文通过核方法将原始数据映射到核空间[12]。文献[13]提出了核稀疏表示,但在求解最优化稀疏矩阵时却相当困难。研究[14]表明,经验核映射(empirical kernel mapping, EKM)所对应的核矩阵与隐形核映射(implicit kernel mapping, IKM)所对应的核矩阵完全一致,由 EKM 及 IKM 所生成的映射样本在各自的特征空间中有着相同的几何结构,而且文献[15]已经证明了 EKM 相对于 IKM 更容易处理及分析核对输入空间的适宜度,所以本文选取经验核映射方法。

对样本集合 $\{x_i\}_{i=1}^{N}$,$K=[\ker_{ij}]_{N\times N}$ 表示 $N\times N$ 的核矩阵,其中 $\ker_{ij}=\Phi(x_i)\times\Phi(x_j)=\ker(x_i,x_j)$,核矩阵 K 是一对称半正定矩阵。假定显性核矩阵 K 的秩为 r,则核矩阵 K 可分解为

$$K_{N\times N}=Q_{N\times r}\Lambda_{r\times r}Q_{N\times r}^{t} \tag{2}$$

其中:Λ 是一对角矩阵,其对角线上每个元素对应于 K 矩阵的 r 个正特征值;矩阵 Q 则是由 K 矩阵的 r 个正特征值所对应的特征向量组成。

那么,显性核映射定义[15]如下:

$$\Phi^e: X \rightarrow F^r$$
$$x \rightarrow \Lambda^{-1/2}Q^t[\ker(x,x_1),\cdots,\ker(x,x_N)]^t \tag{3}$$

令 $B=KQ\Lambda^{-1/2}$,则将通过 EKM 生成的特征样本 $\{\Phi^e(x_i)\}_{i=1}^{N}$ 所对应的经验核矩阵定义为

$$BB^t=KQ\Lambda^{-1/2}\Lambda^{-1/2}Q^tK=K \tag{4}$$

将原始数据 $X=[x_1,x_2,\cdots,x_m] \in R^{n\times m}$ 通过核映射 Φ^e 生成新的特征样本 $\{\Phi^e(x_i)\}_{i=1}^{m}$,对核空间的新样本进行稀疏表示,称之为经验核稀疏表示(empirical kernel sparse representation,EKSR),即有问题 P_2:

$$\min_{s} \| S_i \|_1$$
$$\text{s.t.} \quad \| \Phi^e(x_i) - S_i \times \{\Phi^e(x_i)\}_{i=1}^{m} \|^2 < \varepsilon \tag{5}$$

式(5)是凸优化问题,可通过与式(1)相同的方法求解最小化问题。对于每一个训练样本 $x_i(i=1,2,\cdots,m)$,计算其最优的稀疏重构向量 S_i,即可得到稀疏重构权重矩阵 $S=[S_1,S_2,\cdots,S_m]^{\mathrm{T}}$。

1.2　Empirical kernel sparsity score(EKSS)算法

Qiao 等人[16]指出,稀疏重构权重向量虽然没有数据标号,但是具有一定的判别性能,能够反映数据的全局结构信息。对于样本集 $X=[x_1,x_2,\cdots,x_m] \in R^{n\times m}$,样本的个数为 m,维数为 n,x_{ij} 表示第 i 个样本的第 j 个特征,通过 EKSR 得到样本 x_i 的稀疏重构向量 $S_i=[S_{i1},\cdots,S_{i(i-1)},0,,S_{i(i+1)},\cdots,S_{im}]^{\mathrm{T}} \in R^m$,其中 S_{ij} 表示第 j 个样本 x_j 对重构 x_i 作出的贡献,则样本 x_i 重构如下:

$$x_i=S_{i1}x_1+\cdots+S_{i(i-1)}x_{i-1}+S_{i(i+1)}x_{i+1}+\cdots+S_{in}x_n \tag{6}$$

因此,“好”的特征将最能够保持原始数据的稀疏重构特性,重构误差将达到最小。推广到整个样本集,依据重构误差构建评分准则为

$$\text{KS\_Score}_j^1=\sum_{i=1}^{m}(x_{ij}-\sum_{r=1}^{m}S_{ij}x_{rj})^2 \tag{7}$$

考虑样本特征的全局方差后,评分准则为

$$\text{KS\_Score}_j^2=\frac{\sum_{i=1}^{m}(x_{ij}-\sum_{r=1}^{m}S_{ij}x_{rj})^2}{\sum_{i=1}^{m}(x_{ij}-\mu_j)^2} \tag{8}$$

KS_Score 的值越小即保持稀疏结构的能力越强,依据得分值进行特征选择,即得到基于核稀疏表示的特征选择算法。EKSS 算法流程如下所示:

输入:数据集 $X=[x_1,x_2,\cdots,x_m] \in R^{n\times m}$。

输出:特征排列。

a)原始数据 $\{x_i\}_{i=1}^{m}$ 通过经验核映射得到 $\{\Phi^e(x_i)\}_{i=1}^{m}$;

b)利用式(5)构造核空间中样本的稀疏重构矩阵 S;

c)利用式(7)或(8)计算特征的重构得分;

d)依据重构得分升序排列 n 个特征。

2　实验结果与分析

2.1　实验设置

在实验中,经验核的核函数采用三种不同类型:线性核($\ker(x_i,x_j)=x_i^t x_j$)、RBF 核($\ker(x_i,x_j)=\exp(-\| x_i-x_j \|_2^2/2\sigma^2)$)以及多项式核($\ker(x_i,x_j)=(x_i^t x_j+1)^d$)。其中核参数 σ 的选择如文献[17]的设置,为所有训练样本的平均 ℓ_2 范数距离。核稀疏表示中的错误容忍变量 $\varepsilon=0.001$。

实验中使用了 11 个 UCI 标准数据集(http://archive.ics.uci.edu/ml/)测试特征选择算法的性能,表 1 中列出了这些数据集的属性(C 表示类别数目,D 表示维数,N 表示样本数目)。对于上述数据集,从每类中抽取 2/3 作为训练样本集,剩余 1/3 作为测试样本集。由于基于核稀疏表示的特征选择算法是无监督的,所以对比算法主要选择无监督的 Variance Score、Laplacian Score 和 Sparsity Score。同时作为参考,还将未进行特征选择的原始数据的分类结果作为基准 Baseline。分类器选择 k 近邻分类器(KNN),分类的精度作为特征选择算法的评价指标。

表 1　无监督特征选择算法在 UCI 标准数据集上的最高分类性能分析　%

数据集	Baseline	Variance	LS	SS	EKSS-1	EKSS-2
Wine(3C,13D,178N)	94.03(13)	97.01(4)	**98.51(3)**	95.52(13)	95.52(12)	97.01(9)
Deramatology(6C,34D,366N)	92.42(34)	95.45(21)	96.97(22)	93.18(29)	**97.73(26)**	93.34(31)
Ionosphere(2C,34D,351N)	86.32(34)	88.89(22)	89.74(17)	89.74(30)	94.87(9)	**95.73(5)**
Diabetes(2C,9D,768N)	58.59(9)	63.67(5)	63.67(5)	63.28(3)	**63.89(8)**	62.89(8)
Sonar(2C,60D,208N)	85.33(60)	89.33(51)	**90.67(18)**	86.67(48)	86.67(47)	86.67(52)
Soybean(4C,36D,47N)	100(36)	100(9)	100(6)	100(25)	100(16)	**100(5)**
Breast Tissue(6C,9D,106N)	62.86(9)	68.57(4)	68.57(7)	71.43(6)	71.43(6)	**71.43(5)**
SPECT Heart(2C,22D,267N)	57.22(44)	75.4(1)	75.4(1)	87.6(1)	**89.84(1)**	**89.84(1)**
Water(2C,38D,116N)	100(38)	100(8)	**100(1)**	100(26)	100(24)	100(27)
Glass(6C,9D,214N)	44.29(9)	44.29(1)	48.57(4)	**57.14(5)**	**57.14(5)**	**57.14(5)**
Wdbc(2C,31D,569N)	54.5(31)	57.67(11)	**62.96(1)**	61.9(26)	59.79(21)	60.32(22)

实验前对数据进行归一化处理,使用处理后的数据进行特征选择和分类。

2.2 核函数对性能的影响

通过不同的核函数映射后,分类性能上也会有差异。图 1 和 2 分别给出了 Deramatology 和 Ionosphere 数据集在线性核、RBF 核和多项式核下 EKSS-1 及 EKSS-2 的分类精度。

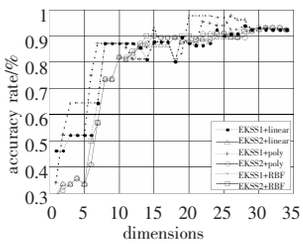

图 1 Deramatology 数据集上不同核函数的 KNN 分类精度

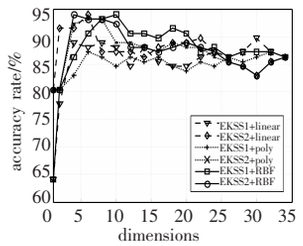

图 2 Ionosphere 数据集上不同核函数的 KNN 分类精度

从图中可以看出,线性核的效果偏弱,RBF 核和多项式核均可达到好的效果,而从复杂性上考虑本文选择了多项式核。

2.3 分类性能分析

表 1 记录了各个无监督算法在 11 个 UCI 数据集上的最高分类性能,括号内的数字代表性能最佳时所选特征的数目,加粗的数值为所有算法中的最高值。根据上文分析结果,核函数选择多项式核。分类器的近邻数选择 $k=5$。图 3 和 4 是 Deramatology 和 Ionosphere 数据集在不同特征选择算法下,特征维数与分类精度的曲线图。

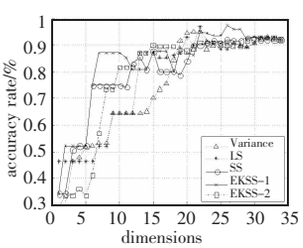

图 3 特征选择算法在 Deramatology 数据集上的 KNN 分类精度

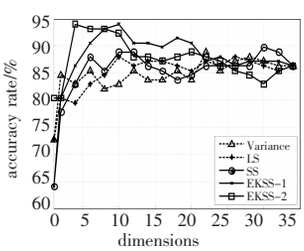

图 4 特征选择算法在 Ionosphere 数据集上的 KNN 分类精度

从表 1 可以看出,EKSS-1 和 EKSS-2 在多个数据集上得到优于其他选择算法的性能。在 Ionosphere 数据集上,EKSS-2 只需要 1 个特征就可以达到 100% 的分类精度;而在 Soybean 数据集上,虽然各个选择算法均能达到 100% 的分类精度,但是 EKSS-2 只需要选择 36 个特征中的 5 个即可达到。Laplacian Score 算法选择了更能保持数据局部特性的特征,在部分数据集上性能优于本文提出的算法;基于方差的 Variance 算法性能弱于其他选择算法,但在一般情况下均优于 Baseline。实验结果说明,特征选择算法对于高维数据有着良好的性能,能够寻找出最优子集,达到降维效果,而本文提出的 EKSS 算法在大多数数据集上的性能优于其他算法。由于 EKSS-2 考虑了全局方差信息,在大多数情况下的分类精度优于 EKSS-1。

从图 3 和 4 中可以看出,通过特征选择算法对特征进行评分后,依据评分选取部分特征即可达到最好的分类效果,达到了降维的目的。

3 结束语

本文通过核映射将稀疏表示扩展为经验核稀疏表示(EKSR),并引入到特征选择算法理论中,提出了基于核稀疏表示的特征选择算法(EKSS)。EKSS 通过稀疏评分准则选取最优子集,达到降维的目的。本文在多个标准 UCI 数据集上进行了实验,结果表明 EKSS 是一种有效的特征选择算法。但是 EKSS 算法没有利用监督信息,仅仅考虑了稀疏结构信息,所以可以考虑选取以类标号为单位的变量集合,使得同类之间的样本不是稀疏的,而异类之间将保持稀疏性,从而有效利用监督信息。对于核学习还可以考虑多核的思想[18]。因此还需进一步深入研究。

参考文献:

[1] DUDA R O, HART P E, STORK D G. Pattern classification[M]. 2nd ed. New York: Wiley, 2000.

[2] DASH M, LIU H. Feature selection for classification[J]. Intelligent Data Analysis, 1997, 1(3): 131-156.

[3] BISHOP C M. Neural networks for pattern recognition[M]. New York: Oxford University Press, 1995.

[4] HE Xiao-fei, CAI Deng, NIYOGI P. Laplacian score for feature selection[C]//Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2005: 507-514.

[5] ZHANG Dao-qiang, CHEN Song-can, ZHOU Zhi-hua. Constraint score: a new filter method for feature selection with pairwise constraints[J]. Pattern Recognition, 2008, 41(5): 1440-1451.

[6] ZHAO Zheng, LIU Huan. Semi-supervised feature selection via spectral analysis[C]//Proc of the 7th SIAM International Conference on Data Mining. 2007.

[7] 刘峤, 秦志光, 陈伟, 等. 基于零范数特征选择的支持向量机模型[J]. 自动化学报, 2011, 37(2): 126-130.

[8] 孙丹. 基于成对约束和稀疏表示的特征选择算法研究[D]. 南京: 南京航空航天大学, 2010.

[9] LIANG Yi-xiong, WANG Lei, XIANG Yao, *et al*. Feature selection via sparse approximation for face recognition[C]//Proc of Computer Science and Pattern Recognition. 2011.

[10] MALLAT S G, ZHANG Zhi-feng. Matching pursuits with time-frequency dictionaries[J]. IEEE Trans on Signal Processing, 1993, 41(12): 3397-3415.

[11] DONOHO D L. Compressed sensing[J]. IEEE Trans on Information Theory, 2006, 52(4): 1289-1306.

[12] 张亮, 黄曙光, 郭浩. 快速核有监督局部保留投影算法[J]. 电子与信息学报, 2011, 33(5): 37-42.

[13] GAO Sheng-hua, TSANG I W H, CHIA L T. Kernel sparse representation for image classification and face recognition[C]//Proc of the 11th European Conference on Computer Vision. Berlin: Springer-Verlag, 2010.

[14] WANG Zhe, CHEN Song-can, SUN Ting-kai. MultiK-MHKS: a novel multiple kernel learning algorithm[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2008, 30(2): 348-353.

[15] XIONG Hui-lin, SWAMY M N S, AHMAD M O. Optimizing the kernel in the empirical feature space[J]. IEEE Trans on Neural Networks, 2005, 16(2): 460-474.

[16] QIAO Li-shan, CHEN Song-can, TAN Xiao-yang. Sparsity preserving projections with applications to face recognition[J]. Pattern Recognition, 2010, 43(1): 331-341.

[17] TSANG I N, KOCSOR A, KWOK J T. Efficient kernel feature extraction for massive data sets[C]//Proc of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2006: 724-729.

[18] DAS S, MATTHEWS B L, SRIVASTAVA A N, *et al*. Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study[C]//Proc of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010: 47-56.