

基于 GHA 的核主成分分析及其应用

潘石柱, 姜伟群, 王令群

(同济大学 控制理论与控制工程学院, 上海 200092)

摘要:文中提出了一种将 GHA(Generalized Hebbian Algorithm)学习规则应用到核主成分分析的新方法,它结合了核主成分分析和 GHA 学习规则的优点,既能利用核主成分分析的方法方便地提取数据的非线性特征,又能避免在大样本数据的情况下运算复杂和存储空间大的问题。实验证明了该方法的可行性和高效性。

关键词:GHA;核主成分分析;特征提取

中图分类号:TP18

文献标识码:A

文章编号:1673-629X(2006)10-0023-03

Research and Application of Kernel Principal Component Analysis Based on Generalized Hebbian Algorithm

PAN Shi-zhu, SHU Wei-qun, WANG Ling-qun

(Department of Control Theory and Control Engineering, Tongji University, Shanghai 200092, China)

Abstract: Presents a new method that combines the algorithm of GHA with kernel principal component analysis which can make good use of respective advantage of two algorithms. First, this method uses the algorithm of kernel principal component analysis to extract the non-linear feature of data. Second, it can also avoid the computational complexity and high dimensionality of space. The experiments had proved this method is feasible and efficient.

Key words: GHA; kernel principal component analysis; feature extraction

在图像处理和模式识别领域,主成分分析被广泛地应用于特征提取,大大地简化了问题处理的难度和提高了识别的性能。但是,主成分分析是一种线性映射算法,在处理非线性问题时,往往不能表现出好的性能。Scholkopf 将主成分分析推广到非线性领域,通过核函数使非线性问题转化为普通的特征值问题^[1,2],这样可以利用已有的线性算法求解原本复杂的非线性特征空间。但是在样本数目很大时,计算求解的核矩阵维数太高,增添了计算的复杂度和保存整个矩阵的难度。

文中提出了一种基于 GHA 的核主成分分析的新方法,该方法结合了 GHA 和核主成分分析的优点,既能获取数据的非线性特征,又解决了运算量和存储空间大的问题,实验证明了该方法的可行性和高效性。

1 基于 GHA 的主成分分析

对数据进行主成分分析时,高维矩阵的特征值分解是一个难题,Oja 提出了一种稳定的学习规则^[3],学习的结

果是网络的权矢量自适应地学习输入矢量的最大主特征向量。由于在实际应用中往往需要提取多个主成分,利用 Sanger 的 GHA 算法^[4]将这个单线性神经元模型扩展到单层线性神经元的前馈网络,可以对输入进行任意大小的主成分分析,其结构如图 1 所示。

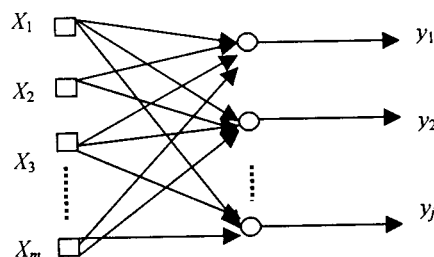


图1 主成分分析的网络结构图

设训练样本集为 $X = \{x_1, x_2, \dots, x_M\}$, 其中 $x_k \in R^N$ 是一个列矢量, M 为训练样本总数, w_{ji} 为连接输入层和输出层的权值。

第 j 个神经元的输出为:

$$y_j(n) = \sum_{i=1}^m w_{ji}(n) x_i(n), j = 1, 2, \dots, l \quad (1)$$

根据 Hebb 学习的广义形式,权值修改为:

$$\Delta w_{ji}(n) = \eta [y_j(n) x_i(n) - y_j(n) \sum_{k=1}^l w_{ki}(n) y_k(n)] \quad (2)$$

收稿日期:2006-01-10

作者简介:潘石柱(1976-),男,河南信阳人,博士研究生,研究方向为图像处理、模式识别、多媒体信息处理、数字视频处理等;姜伟群,博士,教授,博士生导师,主要的研究方向为测量理论、信号处理、流量测量、多媒体技术等。

$$i = 1, 2, \dots, m; j = 1, 2, \dots, m$$

其中 $\Delta w_{ji}(n)$ 是在时刻 n 对 $w_{ji}(n)$ 的修改, η 是学习率。

这样便求得了最大特征值对应的特征向量(即权值)。

为了求得多个分量,因此修改规则为:

$$\Delta w_{ji}(n) = \eta y_j(n) [x'_i(n) - w_{ji}(n) y_j(n)] \quad (3)$$

其中 $x'_i(n)$ 为输入向量 $x(n)$ 的第 i 个分量的修改形式:

$$x'_i(n) = x_i(n) - \sum_{k=1}^{j-1} w_{ki}(n) y_k(n) \quad (4)$$

分析式(4)可知,当 $j = 1$ 时, $x'_i(n) = x_i(n)$ 相当于单个神经元模型,即求取其中最大的主成分。当 $j > 1$ 时,相当于将 $x(n)$ 的前 $j - 1$ 个主分量去除,所以 w_j 将会趋于余下的最大特征对应的特征向量。因此,该网络中所有神经元的权向量都将按照特征值由大到小的顺序收敛于对应的特征向量,这就使该网络具有提取多个主分量的能力。

2 基于 GHA 的核主成分分析

2.1 核主成分分析(KPCA)

KPCA 是一种非线性信号处理方法,在分类前对输入数据进行预处理,可以有效地提取输入数据集的非线性信息,L.J. Cao^[5]实验结果表明:经过特征提取后的 SVM 算法具有更好的分类精度和分类速度,并且利用 KPCA 进行特征提取后的效果明显优于 ICA(Independent Component Analysis)和 PCA(Principal Component Analysis)。

设训练样本集为 $X = \{x_1, x_2, \dots, x_M\}$, 其中 $x_k \in R^N$ 是一个列向量, M 为训练样本总数,设 ϕ 是一个非线性映射,且满足 $\sum_{k=1}^M \phi(x_k) = 0$, 对应的空间记为 F , 则对应的协方差矩阵为:

$$C = 1/M \sum_{j=1}^M \phi(x_j) \phi(x_j)^T \quad (5)$$

对 C 进行特征分解得:

$$\lambda v = C v \quad (6)$$

对所有特征值 $\lambda \geq 0$ 。特征向量 v 是由 $\Phi(x_1), \Phi(x_2), \dots, \Phi(x_M)$ 张成的空间。由此可得式(6)等价于 $\lambda < \Phi(x_k), v^r > = < \Phi(x_k), C v^T >, k = 1, 2, \dots, M$

$$(7)$$

考虑到 v^r 是 $\Phi(x)$ 的线性组合,即

$$v^r = \sum_{i=1}^M c'_i \Phi(x_i) \quad (8)$$

将式(5),(8)代入式(7),并令 $K_{ij} = < \Phi(x_i), \Phi(x_j) >, i, j = 1, 2, \dots, M$, 得

$$M \lambda c^r = K c^r \quad (9)$$

同理可知 $M \lambda^r$ 和 c^r 是对应于 K 的特征值和特征向量,假设对应于特征值大于 0 的特征向量分别为 c^0, c^{0+1}, \dots, c^M , 将 c^r 归一化可得 $M \lambda < c^r, c^r > = 1$, 此时样本 $\Phi(x)$ 在 c^r 上的投影为:

$$g_r(x) = < v^r, \Phi(x) > =$$

$$\sum_{i=1}^M c^r < \Phi(x_i), \Phi(x_j) > \quad (10)$$

$$r = 0, 1, \dots, M$$

称 $g(x)$ 为对应于 $\Phi(x)$ 非线性主元分量,将所有投影值形成一个矢量 $g(x) = [g_1(x), g_2(x), \dots, g_l(x)]^T$ 作为样本的新特征。

在计算 $g(x)$ 时,涉及到求解非线性映射的点积运算问题,若空间的维数很高甚至无限维时,直接计算点积困难,为解决这一问题,同时也为避免寻找映射 Φ , 根据 Mercer^[1]定理,用核函数 $K_1(x_i, x) = < \Phi(x_i), \Phi(x) >$ 代替空间的点积运算,式(10)可写成:

$$g(x) = < v^r, \Phi(x) > = \sum_{i=1}^M c^r K_1(x_i, x)$$

当 $\Phi(x)$ 的均值不等于 0 时,空间样本首先要变为:

$$\bar{\Phi}(x_i) = \Phi(x_i) - 1/M \sum_{i=1}^M \Phi(x_i)$$

其它的算法相似,此时式(9)中的 K 用下式代替:

$$\bar{K} = K - 1_M K - K 1_M + 1_M K 1_M$$

式中 1_M 是系数为 $1/M$ 的 $M \times M$ 阶单位阵。

实验证明^[5], KPCA 方法在空间内具有与线性 PCA 相同的数学和统计特性,如各主元分量互不相关;主元分量能够表示样本数据的最大方差;用主元分量进行样本数据重构其均方误差最小;此外,它比线性 PCA 能提取更多的样本信息,在达到同样分类性能的前提下, KPCA 所需的主元个数要少于 PCA,同时与其它非线性特征提取方法相比,它不需要解决非线性优化问题而只涉及矩阵的特征值分解计算。

2.2 核主成分分析的 GHA 求解

由上述分析可得,利用 KPCA 分析需要先计算核矩阵 K , 然后求解矩阵得到特征值和特征向量。然而训练数据样本较大时,计算过程中不可能一次性存储整个矩阵,并且特征值和特征向量的求解也非常难,根据 GHA 学习规则无需直接计算和存储协方差矩阵的特点,文中将 GHA 引入到 KPCA 的求解中。

首先将核引入到式(3)中可得:

$$W(t+1) = W(t) + \eta(t)(y(t)\Phi(d(t))^T - LT[y(t)y(t)^T]W(t)) \quad (11)$$

其中 $W(t)$ 为网络权值的矩阵形式, $y(t) = W(t)\Phi(d(t))$ 。 $LT[\]$ 将矩阵转化为下三角矩阵。 $\Phi(d(t))$ 为 t 时刻从映射数据点 $\{\Phi(d_1), \Phi(d_2), \dots, \Phi(d_n)\}$ 随机选择的样例。

因为 $W(t)$ 相当于核分量分析中的特征微量,即根据式(8),它可表示为映射数据点的展开式:

$$W(t) = A(t)\Phi(x(t)) \quad (12)$$

其中 $A(t) = (a_1(t)^T, a_2(t)^T, \dots, a_r(t)^T)$

由此新的规则为:

$$A(t+1)\Phi = A(t)\Phi + \eta(t)(y(t)\Phi(x(t))^T - LT[y(t)y(t)^T]A(t)\Phi) \quad (13)$$

引入 n 维列向量 $b(t) = (0, \dots, 1, \dots, 0)^T$,

则 $\Phi(d(t))$ 可表示为:

$$\Phi(d(t)) = \Phi d(t) \quad (14)$$

将式(14)代入式(13),得:

$$\mathbf{A}^T(t+1) = \mathbf{A}^T(t) + \eta(t)(\mathbf{y}(t)\mathbf{b}(t)^T - LT[\mathbf{y}(t)\mathbf{y}(t)^T]\mathbf{A}^T(t)) \quad (15)$$

由式(14)经过分解可得:

$$a_{ij}(t+1) = \begin{cases} a_{ij}(t) + \eta(t)y_i(t) - \eta(t)y_i(t) \cdot \sum_{k=1}^i a_{kj}(t)y_k(t), & b_j = 1 \\ a_{ij}(t) - \eta(t)y_i(t) \sum_{k=1}^i a_{kj}(t) \cdot y_k(t), & b_j \neq 1 \end{cases} \quad (16)$$

$$\text{其中 } y_i(t) = \sum_{r=1}^l a_{ir}(t)\Phi(d_r)\Phi(d(t)) = \sum_{r=1}^l a_{ir}(t)k(d_r, d(t))$$

3 实验结果及分析

在实验室里,利用了车牌牌照识别系统的前期工作得到大量的像素的字符图像,为了使算法具有较强的泛化性和实用性,图像都是在现场不同环境中获取的。文中只对车牌中的字母进行训练和识别。分别在各类中选取 30 个样本作为训练样本,再分别选取 10 个样本作为测试样本。

实验的目的是为了论证基于 GHA 的 KPCA 特征提取在 SVM 字符识别中的作用,以及此算法与 KPCA 特征提取在性能上的不同。分别采用无特征提取、KPCA 特征提取以及基于 GHA 的 KPCA 特征提取进行实验,获得各自的识别率和训练速度,其中软件的编写使用 VC++ , SVM 程序部分改写了 LIBSVM 工具箱中的函数,表 1 给出了统一提取前 50 个特征的实验结果。

从实验结果可以看出,KHA 虽然消耗更长的时间来进行学习,但它利用自适应迭代的学习方法无需直接计算核矩阵、无需寻求矩阵的特征值和特征向量,需要更少的

存储空间并可在线学习,尤其在使用大样本的应用实例时,更能显示此方法的优势。

表 1 实验数据

	SVM	
	识别率(%)	训练速度(m)
无特征	89.9	2.52
KPCA	94.3	5.31
KHA	93.8	6.35

4 总 结

文中提出了一种基于 GHA 的核成分分析的方法,它是利用核函数的方式求取数据空间中的非线性特征,在计算过程中,利用 GHA 求解 KPCA 问题,使 KPCA 特征提取的方式在大样本分类问题中成为可能,实验结果表明了此算法的可行性。但是在实际应用中,还存在着训练速度慢和选取特征数固定的问题,下一步的研究中将着力解决这些问题。

参考文献:

- [1] Scholkopf B, Smola A, Muller K. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation, 1999, 10(5):1299-1319.
- [2] Kim K I, Franz M O, Scholkopf B. Kernel hebbian algorithm for iterative kernel principal component analysis[R]. [s. l.]: Planck Institute for Biological Cybernetics, 2003.
- [3] Oja E. A simplified neuron model as a principal component analyzer[J]. Journal of Mathematical Biology, 1982, 15: 267-273.
- [4] Sanger T D. Optimal unsupervised learning in a single-layer linear feedforward neural network[J]. Neural networks, 1989 (2): 459-473.
- [5] Cao L J, Chua K S, Chong W K, et al. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine[J]. Neurocomputing, 2003, 55: 321-336.

(上接第 22 页)

的类。

由于每个方法都有其优缺点和不同的适用领域,在数据挖掘中,用户应该根据实际需要选择恰当的聚类算法。

参考文献:

- [1] Han Jiawei, Kamber M. Data Mining Concepts and Techniques[M]. 北京:高等教育出版社, 2001. 223-262.
- [2] Zhang T. BIRCH: An efficient data clustering method for very large databases[A]. Proc. of the ACM SIGMOD Int'l Conf on Management of Data[C]. Montreal: ACM press, 1996. 73-84.
- [3] Enter M. A density-based algorithm for discovering clusters in large spatial databases with noise[A]. In Proc of 2nd Int'l Conf on Knowledge Discovering in Databases and Data Mining KDD-96[C]. Portland: AAAI Press, 1996.
- [4] Guha S, Rastogi R, Shim K. CURE: An efficient clustering algorithm for large databases[A]. In Proc of the ACM SIGMOD Int'l Conf on Management of Data[C]. Seattle: ACM Press, 1998. 73-84.
- [5] GEHRKE J, AGRAWAL R, GUNOPUL O. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications[J]. ACM SIGMOD, 1998, 72(2): 94-105.
- [6] BETUR V, DASARAEH Y. Data Mining and knowledge Discovery: Theory Tool, and Technology II[A]. Orlando, florida 2000 SPIE-The International Society for Optical Engineering [C][s. l.]: [s. n.], 2000. 259-264.
- [7] 王 实, 高 文. 数据挖掘中的聚类方法[J]. 计算机科学, 2000(4): 42-45.