

子空间搜索广义主成分分析

曹 扬 罗予频 杨士元

(清华大学自动化系 北京 100084)

摘 要 GPCA(Generalized Principal Component Analysis)是近几年提出的一种数据聚类 and 降维方法,它通过将样本聚类为不同的子空间得到样本的低维表达. GPCA 方法已经被应用于图像分割、图像聚类等问题. 原有的 GPCA 算法具有指数计算复杂度,很难应用于高维数据的实际处理. 文中针对此问题,提出了基于子空间搜索的 SGPCA 算法,将聚类问题分解为单个平面的单个垂直向量的搜索问题,对不同子空间分别搜索,从而实现多项式复杂度算法. 实验表明,新方法不仅计算复杂度低,而且对噪声的鲁棒性也更强.

关键词 主成分分析;子空间分割;数据降维;最小化;局部极小解

中图法分类号 TP391

Subspace Searching Based Generalized Principal Component Analysis

CAO Yang LUO Yu-Pin YANG Shi-Yuan

(Department of Automation, Tsinghua University, Beijing 100084)

Abstract GPCA(Generalized Principal Component Analysis) is a new clustering and dimensionality reduction algorithm. It classifies and represents data in some subspaces. GPCA is used to solve some computer vision problems such as image segmentation and face clustering. Original GPCA algorithm has an exponential computational complexity so that it cannot be applied to high-dimension data. A new SGPCA (Subspace Searching Based Generalized Principal Component Analysis) algorithm is proposed in this paper. Clustering problem is reduced to searching of orthogonal vectors of subspaces. It has a polynomial computational complexity because it searches every subspace one at a time. Experiments show that new method runs faster and more robust to noise than the original algorithm.

Keywords principal component analysis; subspace segmentation; dimensionality reduction; minimization; local minimum

1 引 言

GPCA 是由 Vidal 等人提出的数据聚类和降维算法^[1]. GPCA 根据样本所在子空间对其进行聚类,每类数据可在其子空间内得到低维表示. 它主要应用于基于运动的图像分割^[2-3]、基于纹理的图像分割^[4]、混合线性时不变系统的辨识^[5]和图像压缩^[6]等问题.

假设样本集 X 由 N 个 K 维样本组成,需要聚类为 n 个不同的子空间,GPCA 用一个高维曲面统一表示所有 n 个子空间,并通过数据拟合得到曲面方程,再将曲面分解为不同子空间得到低维子空间表达. 由于 GPCA 算法在第一步中采用曲面拟合的方法,所以要求对一个 $C_{n+K-1}^{K-1} \times N$ 维的矩阵进行 SVD 分解,具有指数计算复杂度,随着计算空间的增长问题将很快无法计算. 文献[6-7]的作者提出先对样本集 X 降维,再用 GPCA 聚类,从而减小计算

量的方法.但这种方法在降维过程中可能造成不同子空间的混叠,使混叠的子空间不可区分.

本文针对现有方法计算复杂度高的问题,提出了基于子空间搜索的 SGPCA 算法.新算法对所有子空间单独计算,将计算量降低到多项式复杂度.本文第 2 节将对原有 GPCA 算法进行简单介绍;第 3 节提出新的 SGPCA 算法,文中将由特殊到一般地逐步解决子空间聚类问题,3.1 节求解单个 $K-1$ 维子空间的问题,3.2 节求解多个 $K-1$ 维子空间,3.3 节求解多个低维子空间的问题;第 4 节将给出实验结果.

2 GPCA

GPCA 算法考虑如下问题:假设 K 维空间 R^K 中的样本集 $X = \{x_j \in R^K\}_{j=1}^N$ 分别处于 n 个未知的线性子空间 $S = \{S_i \subset R^K\}_{i=1}^n$,要求 S_i 与 X_i ,其中 $X_i = \{x_j \in X; x_j \in S_i\}$.

2.1 GPCA 算法^[7]

设子空间 S_i 的维数为 $K - k_i$,则 S_i 可以用 k_i 个线性方程来表示:

$$S_i = \{x \in R^K; B_i^T x = 0\} = \{x \in R^K; \bigwedge_{j=1}^{k_i} (b_{ij}^T x = 0)\}$$
 (1)

其中 $B_i \doteq [b_{i1}, \dots, b_{ik_i}] \in R^{K \times k_i}$ 是 S_i 正交补的一组基.任意点 x 必属于一个子空间,即

$$\begin{aligned} \bigvee_{i=1}^n \bigwedge_{j=1}^{k_i} (b_{ij}^T x = 0) &\Leftrightarrow \bigwedge_{\sigma} \bigvee_{i=1}^n (b_{i\sigma(i)}^T x = 0) \\ &\Leftrightarrow \bigwedge_{\sigma} \prod_{i=1}^n (b_{i\sigma(i)}^T x) = 0 \\ &\Leftrightarrow \bigwedge_{\sigma} p_{n\sigma}(x) = 0 \end{aligned}$$
 (2)

其中, σ 表示一种选取 $b_{i\sigma(i)}$ 的组合方式.设 $\nu_n: R^K \rightarrow R^{M_n(K)}$ 是 n 维 Veronese 映射^[1],其中 $M_n(K) = C_{n+K-1}^{K-1}$,则 K 元 n 次多项式可表示为 $p_{n\sigma}(x) = c_{n\sigma}^T \nu_n(x)$,其中 $c_{n\sigma}$ 为各单项式系数.

综上所述, $\forall x \in X, c_{n\sigma}^T \nu_n(x) = 0$,即

$$c_{n\sigma}^T V_n = 0$$
 (3)

其中, $V_n = [\nu_n(x_1), \nu_n(x_2), \dots, \nu_n(x_N)]$.通过求解方程(3)可以得到 $c_{n\sigma}$.

设 $x_j \in S_i$,则有 $Dp_{n\sigma}(x_j) = \sum_{i=1}^n (b_{i\sigma(i)}^T) \prod_{l \neq i} (b_{l\sigma(l)}^T x_j)$,

即 $b_{i\sigma(i)} = \frac{Dp_{n\sigma}(x_j)}{\|Dp_{n\sigma}(x_j)\|}$.只要为每个子空间 S_i 选定一个样本 $x_{j(i)}$,对所有多项式 $c_{n\sigma}^T \nu_n(x)$ 在点 $x_{j(i)}$ 处求导,即可得到 S_i 的所有垂直向量 b_{ij} .对样本点的分

类结果为 $\hat{X}_i = \{x \in X; d(x, S_i) = \min_j d(x, S_j)\}$,其中 $d(x, S_i)$ 表示点 x 到子空间 S_i 的距离.

2.2 GPCA 算法的不足

已有的 GPCA 算法能够很好地解决数据聚类 and 降维问题,但计算复杂度高,当 n, K 较大时不可计算.

GPCA 算法中运算复杂度最高的部分是求解方程(3)和选择样本 $x_{j(i)}$.前者对 $M_n(K) \times N$ 维的矩阵 V_n 进行 SVD 分解,时间复杂度为 $O(n^{2K}N)$,空间复杂度为 $O(n^K N)$,后者对每个样本求 m_n 维矩阵广义逆,计算复杂度为 $O(Nm_n^3)$,其中 m_n 为方程(3)的解个数.当 K 增大时,计算量呈指数增长,而且随着计算空间的增长将导致问题无法计算.文献[7]的作者建议先对 X 降维,再在低维空间中聚类.文中证明了能够保证子空间不混叠的映射方向集是稠密开集,但也指出不能判断一个特定映射是否能够保留样本特性,所以必须取各种不同映射方向进行计算,从中选取最好的解作为最终结果.

3 SGPCA

针对上面提出的问题,本文提出一种新的具有多项式计算复杂度的 SGPCA 算法.新算法将空间聚类问题转换为优化问题,通过求解问题的各个极小解得到各个子空间的垂直向量.下面将由特殊到一般地解决子空间分解问题:由单个子空间到多个子空间的问题,由 $K-1$ 维子空间到低维子空间问题.

3.1 单个子空间 $k_i=1$ 的情况

设 $S_1 = \{x \in R^K; (b_1^T x = 0)\}$,且样本被加性噪声 $x^n = [x^n(1), x^n(2), \dots, x^n(K)]^T \in R^K$ 污染.只要噪声分布满足如下条件: $\exists r_s$,使得 $\forall b, c \neq 0$ 满足 $P(|b^T x^n| < r_s) > P(|b^T x^n - c| < r_s)$.则对 $\forall b$ 有 $P(|b_1^T x| > r_s) < P(|b^T x| > r_s)$,所以子空间垂直向量是如下问题的解:

$$b_1 = \arg \min_{|b|=1} \sum_{x \in X} h(|b^T x| - r_s)$$
 (4)

其中, $h(s) = \begin{cases} 0, & s \leq 0 \\ 1, & s > 0 \end{cases}$.可以证明,常用的零均值广义高斯分布满足上述条件,所以上述条件并不影响算法适用范围.

由于函数 $h(s)$ 不可导,很难求式(4)的最小解.为了使用基于梯度的优化算法,使用可导的 Sigmoid 函数 $h_s(s) = \frac{1}{1 + \exp(-\frac{s}{\lambda})}$ 代替阶跃函数,从而使

问题(4)可导,其中 λ 只影响目标函数的平滑性,只要目标函数可解,其取值对结果影响不大,在实验中,当 $\lambda \in [10, 1000]$ 时,计算结果基本相同,本文实验中取 $\lambda = 20$. 对于引入 $h_s(s)$ 函数导致的误差,我们使用迭代算法使计算结果只受置信区间内的样本影响. 每次迭代中抛弃距离上次搜索的子空间较远的样本,逐步得到精确解,具体算法见算法 1,其中 α 为样本中单个子空间样本所占比例, β 为每次迭代保留的样本比例,实验中取 $\beta = \alpha^{0.2}$.

算法 1. 搜索算法.

1. $w=1, Y_w=X$;
2. 求解 $b = \arg \min_{|b|=1} \sum_{x \in Y_w} h_s(|b^T x| - r_s)$;
3. 若 $|Y_w| < \alpha \cdot |X|$, 算法结束, 否则继续;
4. 计算样本到子空间的距离 $l_j = |b^T x_j|$;
5. 设 $l_{(j)}$ 为 l_j 的顺序统计量, $w=w+1, Y_w = \{x_j: x_j \in Y_{w-1}, l_j < l_{(\beta \cdot |Y_{w-1}|)}\}$, 转步 2.

3.2 多个子空间 $k_i=1$ 的情况

当问题涉及多个子空间时,每个子空间的垂直向量都是问题(4)的一个极小解. 设共有 n 个子空间, $S_i = \{x \in R^K: (b_i^T x = 0)\}$, 各子空间样本点数 $N_i = |X_i|$, 则

$$P(|b^T x| > r_s) = \sum_{i=1}^n \frac{N_i}{N} P(|b^T x| > r_s | x \in S_i).$$

对某一个子空间的垂直向量 b_i , $P(|b_i^T x| < r_s, x \in S_i) \gg P(|b_i^T x| < r_s, x \notin S_i)$, 在 b_i 的小邻域内忽略其它子空间的影响有

$$P(|b^T x| < r_s) \approx \frac{N_i}{N} P(|b^T x| < r_s | x \in S_i),$$

可知 b_i 是问题(4)的一个极小解. 因此每个子空间的垂直向量对应问题(4)的一个极小解, 求各个子空间的垂直向量即为求问题(4)的各个极小解.

在求出一个极小解后,为了计算其它极小解,这里需要设计新的不包含已有解的最小化问题,重复计算直至得到所有解,最后再去掉由噪声造成的虚假解. 新的最小化问题可以通过减少样本点来构造,算法 1 中的参数 α, β 根据剩余样本中的子空间数量确定,极小解是否已完全找到以及极小解是否对应真实子空间均可根据剩余点数进行判定.

设噪声 x^n 满足: $\exists r_t, p_t$ 对 $\forall b, c \neq 0$ 满足 $P(|b^T x^n| < r_t) > P(|b^T x^n - c| < r_t)$ 且 $P(|b^T x^n| < r_t) \geq p_t$. 当已有 m 个子空间时,排除已有子空间的 r_t 半径内的点,则剩余样本中属于已有子空间 S_i 的样本数小于 $(1-p_t)N_i$, 在剩余样本中重新解问题(4)就可以得到其它极小解. 如果已知全部子空间数量的上界 n , 则剩余样本中包含子空间数上界为 $n-m$, 单个子

空间样本数比例平均为 $\frac{1}{n-m}$, 由于各子空间样本数可能不同, 实验中取搜索算法参数 $\alpha = \frac{1}{3(n-m+1)}$.

当所有子空间都搜索到时, 剩余野值点数量 N_r 为

$$N_r \leq \sum_{i=1}^n (1-p_t) N_i = (1-p_t) N \quad (5)$$

如果某极小解不对应真实子空间, 则置信范围内的样本点数 $N(S_i) = |\{x: x \in X, d(x, S_i) \leq r_t\}|$ 应小于 $(1-p_t)N$. 本文中将阈值放大 γ 倍, 以增强算法鲁棒性, 实验中取 $\gamma = 3$. 由于子空间数只作为真实子空间的判据, 所以算法不需要子空间的准确数量, 只要估计子空间数量的上界. 算法 2 结合了以上几点, 将一个空间分解为几个子空间.

算法 2. 分解算法.

1. $S = \emptyset$;
2. $Y = \{x: x \in X, \forall S_i \in S, d(x, S_i) > r_t\}$, 设 $\alpha = \frac{1}{3(n-|S|+1)}$, $\beta = \alpha^{0.2}$, 在 Y 内应用算法 1, 得到新子空间 S^{new} ;
3. 若 $|S| < n$ 或 $N(S^{\text{new}}) > \min_{S_i \in S} N(S_i)$, 则继续; 否则算法结束, 搜索结果集合为 $\{S_i \in S: N(S_i) > \gamma(1-p_t)N\}$;
4. $S = S \cup \{S^{\text{new}}\}$, 若 $|S| > n$, 去掉 $N(S_i)$ 最小的 $|S| - n$ 个子空间, 转步 2.

3.3 多个子空间 $k_i \geq 1$ 的情况

对于低维子空间, 同样符合算法 2 的使用条件, 算法将得到多个 $K-1$ 维子空间, 分别包含所有真实子空间. 与文献[4-5, 7]中提出的方法类似, 对任意 $K-1$ 维子空间, 只要将其中样本降至 $K-1$ 维, 就可以继续使用算法 2 得到子空间的其它垂直向量, 直到算法 2 不能满足终止条件(5), 则当前子空间不能继续分解. 由于算法 2 不要求已知子空间准确数量, 这里也只需要子空间数量的上界 n_{\max} . SGPCA 在计算过程中并不预先假设噪声的具体分布, 算法中只使用噪声的两个置信区间 r_s, r_t 和对应的 p_t .

算法 3. SGPCA 算法.

1. 设 $A = \{R^K\}$ 和 $B = \emptyset$;
2. 取出 A 中任意子空间 S_i , 设 A 中剩余子空间为 $\{S_i^A\}_{i=1}^a, B$ 为 $\{S_i^B\}_{i=1}^b$;
3. 设 $Y = \{x \in X: d(x, S_i) \leq r_t, \forall i \leq a, d(x, S_i^A) > r_t, \forall i \leq b, d(x, S_i^B) > r_t\}$;
4. 将 Y 降维到 S_i 上, 取 $n = n_{\max} - a - b$, 对 Y 使用算法 2, 得到子空间集合 $\{S_i^{\text{new}}\}_{i=1}^m$;
5. $X_r = \{x \in Y: \forall i \leq m, d(x, S_i^{\text{new}}) > r_t\}$, 若 $|X_r| > \gamma(1-p_t)|Y|$, 则 $B = B \cup \{S_i\}$; 否则 $A = A \cup \{S_i^{\text{new}}\}_{i=1}^m$;
6. 若 A 为空, 算法结束, 否则转步 2.

3.4 算法复杂度

由于最小化问题可导,所以本文中使用 SQP (Sequential Quadratic Programming)方法求解式(4),如果限定迭代次数,则求解计算复杂度为 $O(K^2 + KN)$,一般有 $N \gg K$,则复杂度为 $O(KN)$.

SGPCA 中每次调用算法 1 可以将每个子空间降低一维,所以调用次数为 $O(\sum_{i=1}^n k_i)$,总计算量为 $O(KN \sum_{i=1}^n k_i)$. SGPCA 不需要将样本映射到高维空间中,只需要存储样本原始数据,所以算法空间复杂度为 $O(KN)$.

4 实验结果

这里使用人工数据进行实验,选择 $K=12, n=3, N_i=1000$,各子空间 $S_i = \{[x(1), x(2), \dots, x(K)] \in R^K : x(i) = 0\}$,其它各维服从均匀分布 $U[-100, 100]$. 设样本带有加性零均值高斯噪声 $\mathbf{x}^n \sim N(0, v^2 I)$,其中方差 $v=5$. 对高斯噪声,取两组置信区间 $r_s = v, r_t = 2v$,则 $p_t = 0.95$. 实验 1 从时间、空

间复杂度以及聚类正确率三方面比较新算法与原有算法^[8],表 1 中列出了实验结果. 可以看出,新算法在三个方面均表现出优势.

表 1 对子空间过零点问题的计算性能比较			
方法	计算时间/s	使用内存/MB	错分点数
原有算法	39.6	>100	272
新方法	17.5	<2	242

实验 2 测试算法对噪声的鲁棒性,取 v 为 0, 1, 2, \dots , 25,以错分点数为指标比较新算法与原有算法的性能. 设 $\hat{S}(\mathbf{x})$ 表示样本 \mathbf{x} 的聚类结果,则使用真实子空间进行分类时分类错误概率如式(6),对方差的各个取值分别进行数值积分可得错分点数的理论值.

$$P(|\hat{S}(\mathbf{x}) \neq S_i | \mathbf{x} \in S_i) =$$
$$1 - P(|\mathbf{x}^n(1)| < |\mathbf{x}(2) + \mathbf{x}^n(2)|,$$
$$|\mathbf{x}^n(1)| < |\mathbf{x}(3) + \mathbf{x}^n(3)|) \tag{6}$$

图 1 显示了两种算法在不同噪声下错分点数以及理论最优值,原有算法在噪声较大时不稳定,聚类容易失败,而新算法在较大噪声条件下错分点数仍然接近理论值.

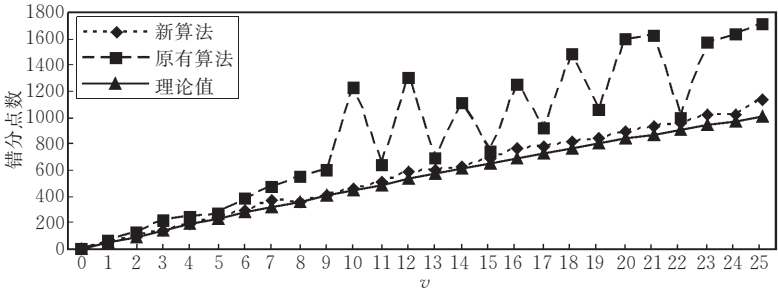


图 1 不同噪声下算法性能比较

图 2 所示为算法所用内存、时间及错分点数随 K 增加的变化趋势,实验中仍使用上述数据模型. 可以看出在 N 与 k_i 固定的情况下,算法计算量相对 K 线性增长,而错分点数与数据维数基本无关. 图 3

所示为算法所用内存、时间及错分点数随 n 增加的变化趋势,在实验中固定 K 与 $k_i, N_i=1000$,可以看出算法计算量相对 N 呈线性增长,错分点数与理论值较为接近.

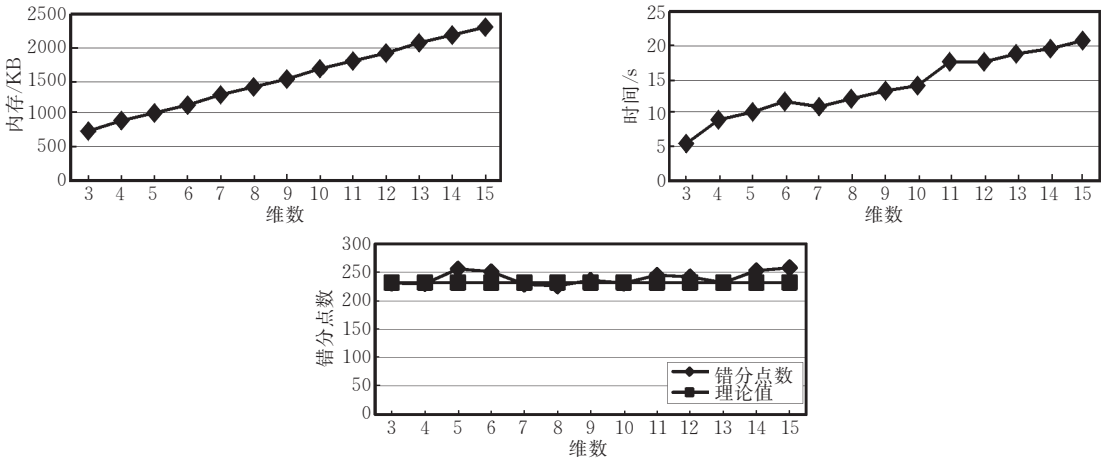


图 2 算法所用内存、时间和错分点数随空间维数增加的变化

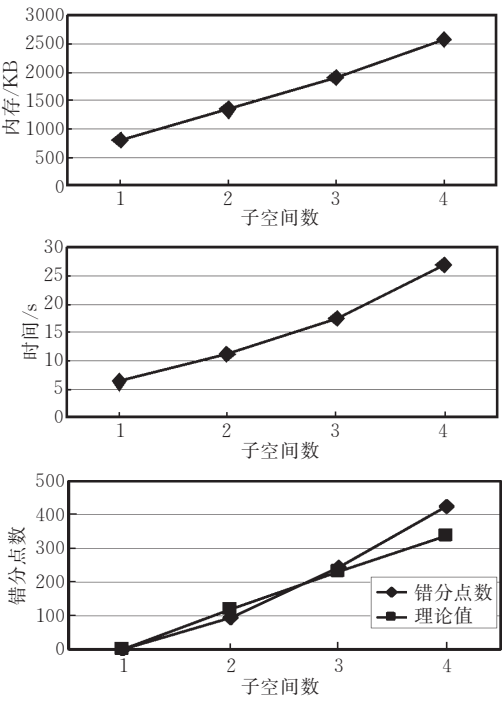


图 3 算法所用内存、时间和错分点数随子空间数增加的变化

5 结 论

本文针对原有 GPCA 算法计算复杂度高的问题,提出了新的 SGPCA 算法. 通过将问题分解为单个子空间的单个垂直向量的搜索,降低了算法的时间空间复杂度. 实验结果证明,新算法计算时间短,

且对噪声的鲁棒性更高.

参 考 文 献

[1] Vidal R, Ma Y, Sastry S. Generalized principal component analysis(GPCA)//Proceedings of the IEEE CVPR. Madison, WI, 2003; 621-628

[2] Vidal R, Hartley R. Motion segmentation with missing data using powerfactorization and GPCA//Proceedings of the IEEE CVPR. Washington DC, 2004; 310-316

[3] Hartley R, Vidal R. The multibody trifocal tensor: Motion segmentation from 3 perspective views//Proceedings of the IEEE CVPR. Washington DC, 2004; 769-775

[4] Huang K, Ma Y, Vidal R. Minimum effective dimension for mixtures of subspaces: A robust GPCA algorithm and its applications//Proceedings of the IEEE CVPR. Washington DC, 2004; 631-638

[5] Huang K, Wagner A, Ma Y. Identification of hybrid linear time-invariant systems via Subspace Embedding and Segmentation(SES)//Proceedings of the IEEE Conference on Decision and Control. Nassau, Bahamas, 2004; 3227-3234

[6] Huang K, Yang A Y, Ma Y. Sparse representation of images with hybrid linear models//Proceedings of the IEEE ICIP. Singapore, 2004; 1281-1284

[7] Vidal R, Ma Y, Sastry S. Generalized principal component analysis(GPCA). IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27; 1945-1959

[8] Vidal R, Ma Y, Piazzi J. A new GPCA algorithm for clustering subspaces by fitting, differentiating and dividing polynomials//Proceedings of the IEEE CVPR. Washington DC, 2004; 510-517



CAO Yang, born in 1979, Ph. D. candidate. His research interests include image denoising and image representation.

LUO Yu-Pin, born in 1959, professor, Ph. D. supervisor. His research interests include image processing, computer vision and computer aided design.

YANG Shi-Yuan, born in 1945, professor, Ph. D. supervisor. His research interests include fault test and diagnosis of electric system, digital community and intelligent home network.

Background

Generalized Principal Component Analysis is a new clustering algorithm. It supposes that data points lie in some subspaces. GPCA is used to solve a lot of computer vision problems because image is often self-similarity. The processed image is often divided into small regions such as 5×5 rectangles, which means that data has 25 dimensions. Original GPCA algorithm that has exponential computational complexity cannot deal with such high-dimension data. Previous

researches use some dimensionality reduction algorithm such as PCA before using GPCA to reduce computational cost.

In this work, the authors address the issue of how to reduce computational cost. A new Subspace Searching Based Generalized Principal Component Analysis algorithm is proposed to solve the same problem. New algorithm searches subspaces in the data one by one so that it can solve the problem with polynomial computational complexity.