基于信息熵的高维稀疏大数据降维算法研究

何兴高，李蝉娟，王瑞锦，邓伏虎，刘 行

(电子科技大学信息与软件工程学院 成都 610054)

【摘要】数据降维是从高维数据中挖掘有效信息的必要步骤。传统的主成分分析（PCA）算法应用于超高维稀疏数据降维时，存在着无法将所有数据特征一次性读入内存以进行分析计算的问题，而之后提出的分块处理PCA算法由于耗时太长，并不能满足实际需求。本文引入信息熵的思想对PCA算法进行改进，提出E-PCA算法，先利用信息熵对数据进行特征筛选，剔除大部分无用特征，再使用PCA算法对处理后的超高维稀疏数据进行降维。通过实验结果表明，在保留相同比例原数据信息的情况下，本文提出的基于信息熵的E-PCA算法在内存占用、运行时间以及降维结果都优于分块处理PCA算法。

关 键 词 分块处理； 降维处理； 高维稀疏大数据； 信息熵； 主成分分析

中图分类号 TP309 文献标志码 A doi:10.3969/j.issn.1001-0548.2018.02.012

Research on Dimensional Reduction of Sparse Matrix Data Based on Information Entropy

HE Xing-gao, LI Chan-juan, WANG Rui-jin, DENG Fu-hu, and LIU Xing

(School of Information and Software Engineering, University of Electronic Science and Technology of China Chengdu 610054)

Abstract Data dimensionality reduction is a necessary step in mining effective information from high-dimensional data. When applying the traditional principal component analysis (PCA) algorithm to high-dimensional sparse data dimensionality reduction, there is a problem that unable to read all data features at once into memory for analysis and calculation, furthermore, the improved block processing PCA algorithm also can not meet the actual requirements because of the time consuming. In this paper, we propose the E-PCA algorithm by introducing the concept of information entropy to improve the PCA algorithm. First, the useless features are eliminated through feature selection based on information entropy, and then PCA algorithm is used to reduce the dimensionality of large, high-dimensional sparse data. The experimental results show that in the case of keeping the same proportion of raw data, the information entropy-based E-PCA algorithm proposed in this paper is superior to block processing PCA algorithm in terms of memory usage, run time and the results of dimension reduction.

Key words block processing; dimensionality reduction; high-dimensional sparse data; information entropy; principal component analysis

随着大数据产业的快速发展，人们关注的数据对象日渐复杂，业界对数据分析、处理技术的需求更为迫切，特别是对高维数据的分析与处理技术。直接处理高维数据会面临以下困难[1-6]：维数灾难、空空间、不适定及算法失效等。为解决以上问题，一种有效的方法就是对高维数据进行降维，分为特征选择和特征变换两种方式[2]。按不同划分标准，算法可分为线性与非线性、监督与非监督、全局与局部等，如PCA、ICA、LDA、LLE、ISOMAP、LTSA、KPCA等。PCA适用于数值型数据，先将数据转换为矩阵形式，再进行相关计算，算法无参数限制，但在某些情况下运行效率不佳。如在处理用户访问网站记录数据时，网站数目庞大，用户能访问的网站数目甚少。这类数据特征维高，有用信息少，即高维稀疏大数据。本文就PCA在处理高维稀疏数据时存在的受内存限制、处理时间长的问题，给出了改进的解决方法。实验结果显示，改进算法能够保留相同比例原数据信息的情况下降低时间成本。

1 相关研究

1901年，统计学领域首先提出主成分分析(principal component analysis, PCA)[7]概念。1923年，

收稿日期：2017－01－04；修回日期：2017－06－15

基金项目：国家自然科学基金(61472064, 61602096)；四川省科技计划项目(2016FZ0002, 2015JY0178, 2016ZC2575)；四川省教育厅重点项目(17ZA0322)；中央高校基本科研基金(ZYGX2014J051, ZYGX2014J066)；网络与数据安全四川省重点实验室开放课题(NDSMS201606)

作者简介：何兴高(1963－)，男，高级工程师，主要从事移动数据管理及其应用、网络安全方面的研究.

文献[8]认为它是比方差分析更适合于相应数据的模型分析。1933年，文献[9]将其推广到随机变量，成为数据挖掘界熟知的一种无监督、线性学习方法。它关注事物的主要性质，将原始变量通过线性变换进行线性组合，从n维特征映射到k维上($k<n$)，这k维数据是重新构造出来的正交特征，被称为主成分。PCA算法简单，具有无线性误差、无参数限制等优点[10-12]。但存储空间大，计算复杂度高，采用的线性映射方法也会影响最后的效果，同时协方差矩阵的大小与样本点的维数成正比，导致计算高维数据的特征向量困难。

针对PCA的局限性，如无明确准则来确定主成分，且存在着诸如高斯假设、线性假设及未考虑数据序列相关性等局限，学者给出了多种改进算法，如动态PCA、非线性PCA、多尺度PCA等。文献[13]探讨对分子数据的降维，为解决传统PCA易受噪声影响的问题，提出了NFPCA(noise free PCA)，在PCs的计算步骤基础上增加一个惩罚项来控制噪声。文献[14]针对基因组单核苷酸多态性数据特征急剧增长，经典PCA处理非常耗时的问题，提出了基于随机算法的高性能PCA的实现方法flash PCA。文献[15]针对大型数据集不能存到随机存储器的问题，采用分块Lanczos方法的随机版本进行处理，迭代次数很少，结果几乎最优，参数l越大，计算复杂度越高，但l的选择没有确定的方法。文献[16]针对人脸识别中存在的图像特征维数高、样本小、耗时长及内存消耗大等问题，基于人脸识别特征和图像特性的考虑，采用分块处理，提出分块PCA。在表情和光照变化的时候，可以捕捉人脸局部特征，并将小样本问题大样本化，在识别性能和识别率上明显优于PCA。

本文针对PCA算法内存消耗大、耗时长，数据特征维高时，处理时间不能满足应用需求的问题，提出基于信息熵的高维稀疏大数据降维算法(E-PCA)。该算法引入信息熵，首先进行特征筛选，降低特征数量，将大型稀疏矩阵稠密化后再做降维处理。

2 基于信息熵的E-PCA算法

2.1 特征值与特征向量

在信息处理上[17]，针对对称矩阵$\boldsymbol{A}$，可以求得正交特征向量，把$\boldsymbol{A}$所代表的空间进行正交分解，使得$\boldsymbol{A}$中的向量可表示为它在各个特征向量上的投影长度，这个投影长度即为对应的特征值。据此，可求出投影前特征值为k的那些分量，丢弃剩下分量，尽最大可能保存矩阵包含的信息，同时能够降低矩阵维度，即降维。在下文的算法中会用此思想来确定具体的k值。

2.2 信息熵

对于信源要考虑所有可能发生情况的平均不确定性。如果信源发出的信号有n种取值，$\boldsymbol{X}=\{a_1,a_2,\cdots,a_n\}$，对应的概率为$p_1,p_2,\cdots,p_n$，并且满足$\sum_{i=1}^{n}p_i=1$。此时，信源的平均不确定性应为单个符号的不确定性$-\log p_i$的统计平均值，称为信息熵，为：

$$\boldsymbol{H}=\mathrm{E}[-\log p_i]=-\sum_{i=1}^{n}p_i\log p_i \tag{1}$$

式中，对数以2为底，单位为比特(bit)；以e为底，单位为奈特(Nat)；以10为底，单位为迪特(Det)。

从系统有序性上考虑[18]，一个系统越有序，信息熵就越低；反之，一个系统越混乱，信息熵就越高。所以，信息熵也可以说是系统有序化程度的一个度量。信息熵应用于特征时[19-21]，在信息量表达上也可以这样解释，如果一个特征的信息熵越大，说明其包含的数据信息量越大，能提供更多的信息，在降维时，应该保留该特征；反之，信息熵越小，说明其包含的数据信息量越小，能提供的信息有限，在做特征提取或者降维时，该特征不列入备选项，也就是应该被剔除或者不保留的特征。以信息熵阈值δ来界定特征是被剔除或者保留，δ可由以下两种方案确定：1) 根据特征对应用分析的重要性判断，适用于有用特征和无用特征信息熵值有明显的分界点；2) 利用表达式计算选择的特征在原始数据中的比重：

$$\frac{\sum_{i=1}^{k}\boldsymbol{H}(i)}{\sum_{j=1}^{n}\boldsymbol{H}(j)}\geqslant \text{threshold} \tag{2}$$

图1所示为Arcene数据集(来自UCI机器学习库)前50个属性的信息熵值曲线图，如应用需要保留至少45个属性，δ不能大于0。

2.3 E-PCA降维算法

文献[16]中的方法，可能由于分块不随机，原始数据分布不均匀等情况，导致整体数据的某些主成分在分块中相对占比少而被丢弃；而某些成分虽算不上主成分，但在分块中占比很高，却被算作主成分而被保留，导致最后的结果不够完美。基于以上局限性，本文从全局的角度去考虑主成分占比，提出一种基于信息熵的E-PCA算法，用于超高维稀疏

数据的降维，处理流程如图2所示，其中 δ 代表信息熵阈值。

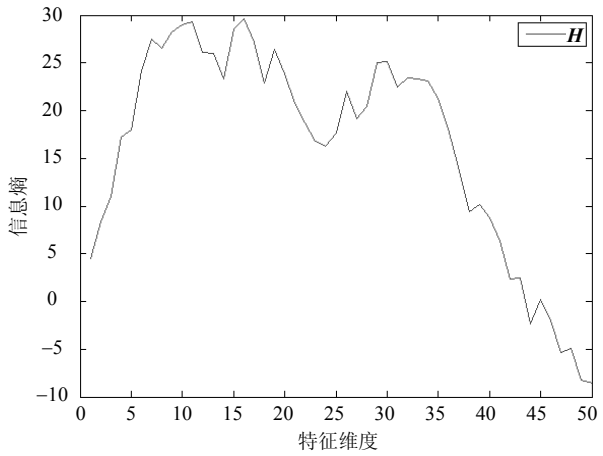

图1　Arcene数据集前50个属性的信息熵值

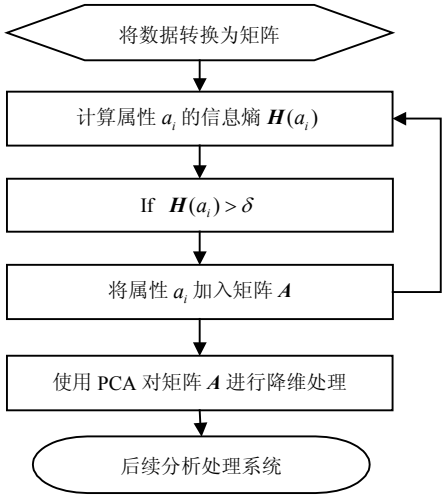

图2　基于信息熵的降维处理流程

基于信息熵的PCA降维算法(E-PCA)原理和PCA一样，区别仅在于E-PCA在应用PCA算法对数据做降维以前做了一次特征筛选：设置信息熵阈值 δ，过滤掉那些几乎无用的原始数据信息的属性(特征)，E-PCA算法的具体步骤如下所示。

输入：数据矩阵 $\boldsymbol{U}_{n\times m}$，其中 m 代表样本个数，n 代表属性(特征)个数；信息熵阈值 δ；贡献率 f。

输出：降维结果 $\boldsymbol{Y}_{k\times m}$。

1) 计算每个属性的信息熵值，与阈值 δ 比较，进行特征筛选，对 $\boldsymbol{U}$ 做如下操作：

For $i=1:n$

计算属性 a_i 的信息熵 $\boldsymbol{H}(a_i)$

If $\boldsymbol{H}(a_i)>\delta$

将属性 a_i 放入集合 $\boldsymbol{A}$ 中

End if

End for

2) 样本矩阵中心化，得矩阵 $\boldsymbol{X}_{n\times m}$：

$$\boldsymbol{X}=\boldsymbol{A}-\text{repmat}(\text{mean}(\boldsymbol{A},2),1,m)$$

3) 计算不同属性维度之间的协方差，构成协方差矩阵 **Cov**：

$$\textbf{Cov}=(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})/(\text{size}(\boldsymbol{X},2)-1)$$

4) 计算 **Cov** 的特征值 eigenValue 和特征向量 **eigenVector**

5) 选定变换基：

选择最大的 k 个特征值对应的 k 个特征向量分别作为列向量组成特征向量矩阵 $\boldsymbol{V}_{n\times k}$。

6) 计算降维结果：

$$\boldsymbol{Y}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{X}$$

7) 算法结束。

E-PCA算法中k值并没有作为输入参数，而是在计算中根据特征值的贡献率选取的，贡献率是指选取的特征值的和与所有特征值的和的比值，用式(3)计算：

$$f=\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{k}\lambda_i}\geqslant\text{threshold} \tag{3}$$

3　算法实验与分析

3.1　实验环境与方法

实验环境：本文仿真实验借助仿真工具Matlab R2014a在服务器上实现，服务器参数如下：操作系统Ubuntu 16.04 LTS，内存64 GB，CPU E5-2609，8核，主频1.70 GHz。

实验方法：分块PCA和E-PCA算法。

3.2　实验数据

本文算法所研究的数据来自R公司用户一定时间段内浏览网站的记录数据，带有类标签1、–1，分别代表正类、负类，即属于某年龄段的用户类和不在相应年龄段内的用户类。某位用户的访问量记录如下：24 409, 38 115, 44 944, 57 604, 112 224, 115 110, 127 659, 131 203, 134 084, 137 383, 149 874, 175 643, 194 142, 194 506, 202 770, 203 189, 212 584, 217 724, 229 474, 244 441, 250 507, 264 338, 270 530，代表正类用户在一段时间内浏览网站的记录。将数据转换成矩阵形式，列(属性)为网站编码(最大为282 646)，一行为一位用户的浏览记录，被用户浏览过的网站列对应位置为1，没被浏览过的置为0，在一定时间段内被用户访问的网站很少，所以矩阵

里值为1的元素数量很少，值为0的元素对于研究用户访问特点几乎无意义。而数据处理过程中，特征维数太高，不能一次性读入内存进行分析计算，基于信息的含义，引入信息熵[22]，保留原始数据中包含信息丰富的特征维，剔除那些即使丢掉也对原始数据信息完整性影响很小的特征维以达到降维的目的。

本文使用的原始数据大小246 KB，包含5 000个样本，282 669个属性。转换为矩阵以后将样本按4:1的比例随机分为训练集和检验集，即4 000个样本做训练集，1 000个样本做检验集。本文研究数据中有用特征与无用特征信息熵值有明显分界点，用方案1确定信息熵阈值。对于某一个属性 a_i (此例中为网站编号)取值如果全为0，则 $p(a_i=0)=1$，$p(a_i=1)=0$，意味着没有访客访问过该网站，计算信息熵的时候，必然会出现无穷小与负无穷大的乘积，结果不是一个数(为 NaN)。实际收集的数据中，绝大部分网站是没被用户访问的，会有很多列没有取值为1的项，计算出的属性信息熵大多数都是 NaN 。由于 $p_i<1$，所以 $p_i\log p_i<0$，进而 $\boldsymbol{H}>0$。综上，E-PCA算法中信息熵阈值 δ 设置为0。分块PCA算法处理时，将28万维的实验数据分为5块，平均每块包含56 535维属性。实验从以下几方面进行结果验证：内存占用、运行时间、降维后结果维数以及分类准确率。

3.3 结果对比与分析

1) 内存占用

PCA算法用于高维数据降维时，主要计算属性协方差矩阵的特征值和特征向量，内存消耗大、耗时长，特别是计算属性协方差矩阵**Cov**时，时间、空间复杂度都会随着维数的增长而急剧增长。设数据X包含N个属性，属性的协方差矩阵包含 $N\times N$ 个数据，以布尔值存储时，每个数据占据一个字节，则此协方差矩阵需占用内存 $(N\times N\times 8)/1\,024^2$ MB，而实际情况下，内存占用比理论值大。表1列出了不同属性维度下的**Cov**理论内存占用量。

表1 不同属性维度下的内存占用情况

属性维数	内存占用(理论)/MB	内存占用(实际)/MB
1 000	7.63	309
5 000	190.73	867
10 000	762.94	3 816
15 000	1 716.61	4 169
20 000	3 051.76	6 214
30 000	6 866.46	14 558.51
40 000	12 207.03	27 587.89
56 535	16 930.88	40 828.93
169 605	219 466.06	-
282 669	609 602.08	-

PCA算法运行时内存与CPU占用率如图3所示，其中实验机器有运行安全保护等后台程序。

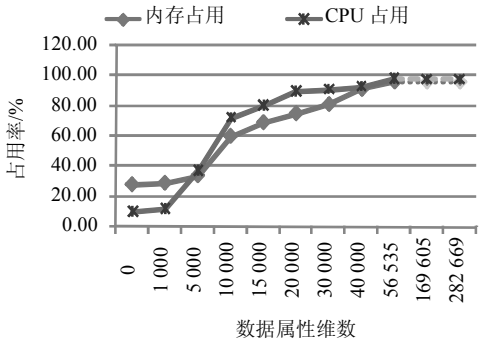

图3 PCA处理不同属性维数的内存与CPU占用率

当数据属性维数为0时的内存与CPU占用率为Matlab软件自身所耗资源。当属性维数为20 000时，CPU占用率达90%，当属性维数继续增长时，CPU占用率增长缓慢。当属性维数在56 535以下，内存占用率随维数增加呈上升趋势，属性维数等于56 535时，程序运行至计算协方差矩阵**Cov**处，终止退出并报错“内存不足”，属性维数再增加时，已经没有更多的内存可供使用，服务器可以提供的CPU占用率和内存占用率只在它力所能及范围内，在图3中维数大于56 535以后，内存占用以延长虚线表示，CPU占用以延长短线表示。

E-PCA用于高维稀疏大数据降维时，内存占用与数据维数、信息熵以及贡献率有关。E-PCA仍然需要计算**Cov**的特征值和特征向量，但这里**Cov**的规模受属性信息熵阈值 δ 的影响，δ 越大，**Cov**规模越小；反之，越大。分布均匀的数据，在贡献率 f 相同的条件下，数据维数越高，计算开销越大。实验给出了不同数据维度下，算法E-PCA和PCA运行时，内存占用情况的对比，如图4所示，其中 δ=0，f=0.95 。

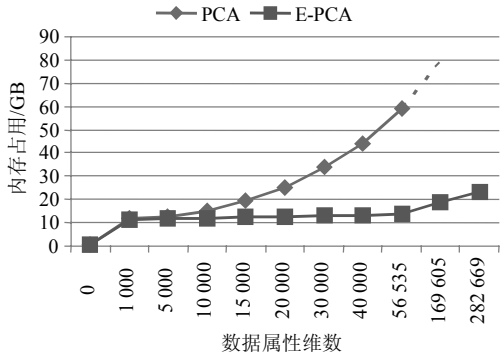

图4 E-PCA和PCA运行时内存占用情况对比

图4中数据显示，当属性维数小于5 000，E-PCA和PCA在运行时，内存占用率差别不大，E-PCA以

微弱的优势胜于PCA。但是随着属性维数的增大，E-PCA呈现了明显的优势。当将属性均匀分为6块，PCA处理其中一块时出现运行错误，程序终止退出并提示在计算**Cov**处“内存不足”，所以实验设定PCA能处理的最大块为将原始数据属性均分为5块，即每块属性维数为56 535，图4中数据属性维数大于56 535之后，PCA算法以延长虚线表示。当数据属性继续增加，不借助于分布式平台、云平台，也不采用分块等技术时，PCA显得无能为力。对于E-PCA算法而言，内存占用呈上升趋势，但一直不大于PCA，即使数据属性维数达到本文实验的最高值282 669，测试服务器仍然能给出足够的空间以供降维使用。就图4结果而言，E-PCA性能优于PCA。理论上，E-PCA首先利用信息熵做特征筛选，使得特征个数减少，后续协方差的相关计算内存开销一定会小于PCA。在图4中体现为E-PCA曲线永远在PCA之下。

综上，在内存占用上，E-PCA性能优于PCA。

2) 运行时间

R公司原始数据转换成矩阵后，计算每个属性的信息熵，并与阈值比较，大于阈值的对应属性被留下加入矩阵$\boldsymbol{A}$。根据图2的流程，接下来进入降维处理步骤，表2列出了Arecene(10 000维)和R公司数据(2 826 669维)两种数据集的实验结果，R公司数据用了分块处理，最终时间为分块处理的和；降维后结果维数k都是在贡献率$f=0.95$条件下的结果。从表中数据显示，数据维数越高，降维时间开销越大。

表2　PCA算法运行时间记录

数据集	属性维数	运行时间/s
R 公司	282 669	153 487.65
Arecene	10 000	132

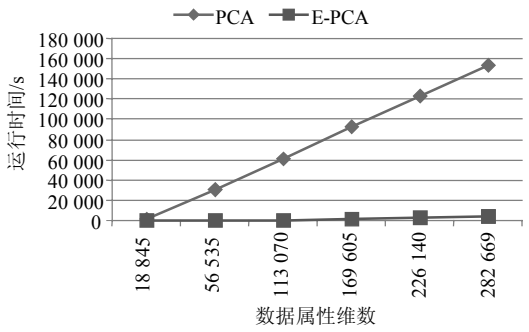

图5　PCA和E-PCA运行耗时对比

实际上，PCA算法用于降维时，时间开销会随着维数的增加急剧增长，图5展示了PCA和E-PCA两种算法处理时间随着维数的增长变化的趋势。

从表2和图5可以看出，当数据维数达到28万维的时候，PCA降维的时间要花上数小时，完全不能满足应用需求，而采用E-PCA降维算法可以极大地降低数据降维处理时间。

3) 降维后结果维数

PCA和E-PCA两种算法处理过程中都会预先设定贡献率f的值以确定k值，图6显示由PCA、E-PCA处理时，k随f的增大而增大，其中横轴表示式(3)中f的值，纵轴表示降维处理后的维数k。PCA处理的为原始数据中的属性数为18 845的一块数据。当$f=1$时，k(PCA)=1 191，k(E-PCA)=3 113，也就是说，经PCA投影后，18 845维的数据，保留最大的926个主成分就可以保留原18 845个属性所包含的信息；信息分布均匀的条件下，对282 669维的数据，需要保留15×1 191=17 865个主成分才能完全保持原始数据所包含的信息，而经E-PCA处理后，保留最大的3 113个主成分就可以保留282 669维属性包含的信息，节省了不少存储空间。表3列出了在$f=0.95$的情况下，E-PCA和PCA降维处理的结果对比，无论从结果维k还是运行时间来说，E-PCA都明显优于PCA。

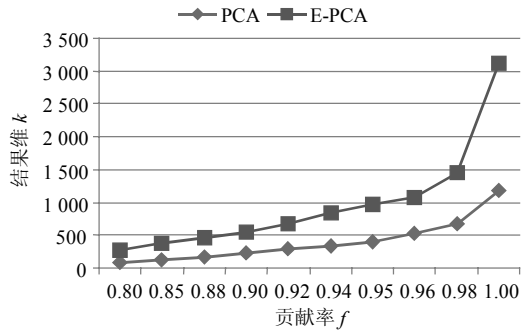

图6　PCA和E-PCA降维后结果对比

表3　PCA、E-PCA处理R公司高维数据结果

方法	时间开销/s	贡献率f	结果维k
E-PCA	3 365.83	0.95	961
PCA	15 487.65	0.95	6 323

4) 分类准确率

为评价降维后数据经KNN、SVM分类算法的准确率，本文还比较了降维前、后数据的分类准确率，表4记录了来自R公司高维稀疏大数据由PCA与E-PCA降维处理后数据的分类准确率。

由表4中数据可以看出，原始数据由KNN和SVM分类的准确率分别为53.1%、53.6%，PCA降维后的KNN和SVM分类准确率分别为52.5%、50.5%，

E-PCA降维以后数据由KNN和SVM分类的准确率分别为53.1%、53.9%。PCA降维后的分类准确率略低于原始数据分类准确率，而E-PCA稍高，因此，就分类准确率来说，依然是E-PCA优于PCA。

表4 PCA和E-PCA算法降维前后数据的分类准确率对比

算法名	贡献率f	降维后结果k	降维后/%		降维前/%	
			KNN	SVM	KNN	SVM
E-PCA	0.95	961	53.1	53.9	53.1	53.6
PCA	0.95	3 323	52.5	50.5		

降维前后分类准确率都不高的原因在于：原始数据维数太高，信息繁杂，导致分类器辨识度不高。降维以后的数据仍然不能被分类器很好地识别的原因在于PCA降维的目标是使得信息的损失最小，并通过衡量在投影方向上的数据方差的大小来衡量该方向的重要性，往方差最大的方向投影使得投影后的数据尽最大可能保持原始数据信息。这期间，由于并没有对类间间距做过多的考虑，投影后对数据的区分作用并不大，反而可能使得数据点揉杂在一起无法区分。这也是PCA存在的最大一个问题，这导致使用PCA在很多情况下的分类效果并不好。在后续的研究中，会继续根据应用需求改进算法，选择合适的评价指标，得出更适合于应用的结果。

4 结 束 语

本文针对稀疏大数据特征维数过高，使用PCA降维时，矩阵计算内存消耗太大，使用文献[16]的分块处理技术，比较麻烦，运行时间远远不能满足应用需求，改进了降维算法PCA，给出基于信息熵的E-PCA降维算法。实验结果表明，E-PCA在保持原始数据尽可能多的信息的时候，运行耗时和内存消耗得到了极大的改善。接下来，将利用量子计算和通信[23-24]进一步提高算法的性能。

参 考 文 献

[1] JAIN A, CHANDRASEKARAN B. Dimensionality and sample size considerations in pattern recognition practice[J]. Handbook of Statistics, 1982(2): 835-855.

[2] HOU L, GAO J, CHEN R. An information entropy-based animal migration optimization algorithm for data clustering[J]. Entropy, 2016, 18(5): 185-200.

[3] WANG Rui-jin, LI Dong-fen, QIN Zhi-guang. An immune quantum communication model for dephasing noise using four-qubit cluster state[J]. International Journal of Theoretical Physics, 2016, 55(1): 609-616.

[4] 王珏, 杨剑, 李伏欣, 等. 机器学习的难题与分析[C]//第三届机器学习及应用研讨会. 南京: [s.n.], 2005.
WANG Yu, YANG Jian, LI Fu-xin, et al. Difficulties and analysis of machine learning[C]//The Third Machine Learning and Application Seminar. Nanjing: [s.n.], 2005.

[5] LI Dong-fen, WANG Rui-jin, ZHANG Feng-li, et al. Quantum information splitting of arbitrary two-qubit state by using four-qubit cluster state and Bell-state[J]. Quantum Information Processing, 2015, 14(3): 1103-1116.

[6] 尹芳黎, 杨雁莹, 王传栋, 等. 矩阵奇异值分解及其在高维数据处理中的应用[J]. 数学的实践与认识, 2011, 41(15): 171-177.
YIN Fang-li, YANG Yan-ying, WANG Chuan-dong, et al. Matrix singular value decomposition and its application in high dimensional data processing[J]. Mathematics in Practice and Theory, 2011, 41(15): 171-177.

[7] PEARSON K. On lines and planes of closest fit to systems of points in space[J]. Philosophical Magazine, 1901, 2(6): 559-572.

[8] FISHER R, KENZIE W M. Studies in crop variation II. The manorial response of different potato varieties[J]. Journal of Agricultural Science, 1923, 13(3): 311-320.

[9] HOTELLING H. Analysis of a complex of statistical variables into principal components[J]. British Journal of Educational Psychology, 1933, 24(6): 417-520.

[10] JOLLIFFE I T. Principal component analysis[J]. Journal of Marketing Research, 2002, 87(100): 513.

[11] GUEBEL D V, TORRES N V. Principal component analysis(PCA)[M]. New York: Springer, 2013.

[12] 张道强, 陈松灿. 高维数据降维方法[J]. 中国计算机学会通讯, 2009, 5(8): 15-22.
ZHANG Dao-qiang, CHEN Song-can. Research on dimension reduction methods of high dimensional data[J]. Communications of the CCF, 2009, 5(8): 15-22.

[13] WANG Y. Semi-supervised dimensionality reduction[J]. Proceedings of the International Symposium on Computer Science, 2010, 41(9): 1993-1998.

[14] REZGHI M, OBULKASIM A. Noise-free principal component analysis: an efficient dimension reduction technique for high dimensional molecular data[J]. Expert Systems with Applications, 2014, 41(17): 7797-7804.

[15] ABRAHAM G, INOUYE M. Fast principal component analysis of large-scale genome-wide data[J]. Plos One, 2014, 9(4): e93766.

[16] HALKO N, MARTINSSON P G, SHKOLNISKY Y, et al. An algorithm for the principal component analysis of large data sets[J]. Siam Journal on Scientific Computing, 2010, 33(5): 2580-2594.

[17] 陈伏兵, 杨静宇. 分块PCA及其在人脸识别中的应用[J]. 计算机工程与设计, 2007, 28(8): 1889-1892.
CHEN Fu-bing, YANG Jing-yu. Realization of face recognition algorithm based on block PCA[J]. Computer Engineering and Design, 2007, 28(8): 1889-1892.

[18] CHEN Fu-bing, YANG Jing-yu. PCA face recognition algorithm based on local feature[J]. Mini-Micro Systems, 2006, 7(10): 1943-1947.
[19] 尹飞, 冯大政. 基于 PCA 算法的人脸识别[J]. 计算机技术与发展, 2008, 18(10): 31-33.
YIN Fei, FENG Da-zheng. Face recognition based on PCA algorithm[J]. Journal of Computer Technology and Development, 2008, 18(10): 31-33.
[20] LI Dong-fen, WANG Rui-jin, ZHANG Feng-li, et al. A noise immunity controlled quantum teleportation protocol[J]. Quantum Information Processing, 2016, 15(11): 4819-4837.
[21] AMPILOVA N, SOLOVIEV I. On application of entropy characteristics to texture analysis[J]. Wseas Transactions on Biology & Biomedicine, 2014, 11(1): 194-202.
[22] PHOENIX S J D. Elements of information theory[M]. [S.l.]: Wiley, 1992.
[23] LI Dong-fen, WANG Rui-jin, ZHANG Feng-li. Quantum information splitting of a two-qubit Bell state using a four-qubit entangled state[J]. Chinese Physical C, 2015, 39(4): 26-30.
[24] LI Dong-fen, WANG Rui-jin, ZHANG Feng-li, et al. Quantum information splitting of arbitrary three-qubit state by using seven-qubit entangled state[J]. International Journal of Theoretical Physics, 2015, 54(6): 2068-2075.

编 辑 蒋 晓

(上接第215页)

[5] BRUNO B, FRANCKY C, DENIS C. Energy efficiency of the IEEE 802.15.4 standard in dense wireless microsensor networks: Modeling and improvement perspectives[J]. Springer Netherlands, 2008, 1(S02): 196-201.
[6] WANG Qin, YANG W W. Energy consumption model for power management in wireless sensor networks[C]//4th Annual IEEE Communications Society, Conference on Sensor, Mesh and Ad Hoc Communications and Networks, 2007 (SECON'07). San Diego, USA: IEEE, 2007: 142-151.
[7] CIGDEM E, MERVE S V, CAGRI G. Lifetime analysis of wireless sensor nodes in different smart grid environments [J]. Wireless Networks, 2014(20): 2053-2062.
[8] LJILJANA S, STEVAN M B, KEVIN W S. Partner choice and power allocation for energy efficient cooperation in wireless sensor networks[J]. ICC, 2008, 14(3): 4255-4260.
[9] SALLABI F M, GAOUDA A M, EI-HAG A H. Evaluation of Zigbee wireless sensor networks under high power disturbances[J]. IEEE Transactions on Power Delivery, 2014, 29(1): 13-20.
[10] WANG Chu-fu, SHIH J, PAN Bo-han. A network lifetime enhancement method for sink relocation and its analysis in wireless sensor networks[J]. IEEE Sensors Journal, 2014, 14(6): 1932-1942.

编 辑 漆 蓉