浅述降维方法及非线性降维方法举例

吉林大学珠海学院公共基础课教学与研究中心数学教研室　宋　靓

[摘　要]随着新工具,新技术的产生,人们使用越来越多的变量数据来描述某一现象,即为高维数据。但随着数据变量的增多,产生了前所未有的困难。文章介绍了降维方法来解决这一问题,并通过举例介绍了非线性降维方法在解决降维问题过程中的优势和重要意义。

[关键词]降维　非线性　LLE 方法

物质世界和人类社会中存在着大量的复杂事物及现象，人们总是希望揭示藏在这些纷繁复杂现象下的事物的客观规律。长久以来,人们不断地研制新的观察工具,发展新的观察技术。如对天气状况的研究,随着气象学的不断发展,可用来描述气象特征的指标越来越多,例如温度、湿度、气压、风力、降雨量、辐射强度等,可以获得更多的关于每时每刻天气状况的更加完善的信息。从而对天气状况，这一抽象自然界现象,可通过上述变量组成的数据来进行细致的综合描述。这些描述"某一现象"的多变量数据,即为高维数据。显然随着数据维数的不断提高,数据能提供有关客观现象更加丰富、细致的信息。但同时数据维数的大幅度提高又会给随后的数据处理工作带来前所未有的困难，针对这个问题提出降维理论。

降维方法是用来克服"维数灾难"和模型化高维数据的一种典型数据处理技术,是用来解决这一问题的有效手段之一。它可通过对离散数据集合的分析来探求嵌入在高维数据空间中本征低维流形的不同样式,寻求事物的本质规律性。

一、降维问题的分类

降维问题可初步地分类为:

硬降维问题:数据维数从几千到几万甚至几十万的变化,此时需要对数据集进行"严厉"地降维,以至于达到便于处理的大小,如图像识别、分类问题以及语音识别问题等。

软降维问题:此时数据集合的维数不是太高,降维的需求不是非常的迫切。如社会科学、心理学以及多元统计分析领域皆属于此类。

可视化问题:此时数据集合的绝对维数不是很高,但为了便于利用人们的直观洞察力,即为了可视化,我们将其降到 2 或 3 维。虽然我们可以可视化更高维数的数据,但是它们通常难于理解,不能产生数据空间的合理形态。

若我们还考虑时间变量的话可以对降维问题进行更加进一步的分类,静态降维问题和动态降维问题。后者对于向量时间序列来讲是有用的,如视频序列、连续语音信号等的处理。

二、降维的定义

若$\{x_n\}_{n=1}^N \subset R^D$为 D 维空间中的一个容量为 N 的数据集合,假设其来自于(或近似来自于)维数为 d(《D)的某一数据流形的采样。我们处理的目标是探求数据流形合适的低维坐标描述，将原数据集合投影到低维空间,获得原数据集合的低维简洁表示。此时,降维问题通常由下列部分组成:

高维数据空间,Ω^D,通常为 R^D 的某一子集;

降维空间(或称低维表示空间),Ω^d,通常为 R^d 的某一子集;

降维映射 F,

$F:\Omega^D \to \Omega^d$, $x \to y=F(x)$

称 y 为 x 的降维表示;

重构映射(不是必须的)F^{-1},

$F^{-1}:\Omega^d \to \Omega^D$, $y \to x=F^{-1}(y)$

满足下面的一些限制:d<D,并且在不损失原数据信息的基础之上,尽可能的小;流形 $S \triangleq F^{-1}(\Omega^d)$近似含有全部的采样点,即$\{x_n\} \subset S$(1)或从另一角度重新描述为重构误差尽可能的小。其中重构误差定义为:

$$Error(\{x_n\}) \triangleq \sum_{n=1}^{N} \rho(x_n, x_n^*), x_n^* \triangleq F^{-1}(F(x_n))$$

其中 $\rho(*,*)$为 Ω^d 中的距离测度,如 R^d 中的欧氏距离等。

根据 Ω^D 和 Ω^d 的差别,可以看出将会存在一个维数为 D-d 的子流形被映射为一点。即:

$$F^{-1}(y) \triangleq \{x \in \Omega^D / F(x)=y\}$$

是一个 D-d 维的数据流形。

三、非线性降维方法及举例

对具有线性结构的数据而言,进行分析和描述相对来讲是简单的,在实际中是便于执行、解释的。这使得线性降维方法在应用上比非线性降维方法更加广泛、更加深入。但就现实世界中所获得的真实数据集合而言,线性往往只是人们的一种理想情形,数据集合更多的是呈现出非线性的结构,甚至是高度非线性的结构,如图像数据、金融数据、多源空间数据等。而线性假设的内在不足,使得线性降维方法无法揭示出数据集合中所含有的非线性结构故为了弥补线性降维方法的不足，有效地探求数据集的内在非线性结构，人们发展了许多有效的非线性降维手段,如多维尺度法(MDS)、ISOMAP 方法、局部线性嵌入方法、Laplacian Eigenmap 方法等。

对数据集合内蕴结构而言,有下列特性:

(一)由泰勒定理,任何可微函数在一点的充分小的邻域之内满足线性。形象地来讲,相当于认为曲面流形可由大小不一的局部线性块拼接而成;数据流形经常是由许多可分割的子流形所组成。

(二)数据流形的本征维数沿着流形不断地发生变化,只有局部性才能抓住其根本特性。

接下来,将介绍一种有效的非线性降维方法:局部线性嵌入方法。

局部线性嵌入方法(LLE)是由 Roweis 和 Saul 于 2000 年提出的一种崭新的非线性降维方式。根据泰勒定理我们知道,可微函数具有良好的局部线性,即每点的微小邻域总可以用线性模型比较好的近似。同样地对任意光滑流形而言，它的微小局部一定程度上也应该具有线性的特征,只要我们可以清晰地描述这一局部线性特征,那么我们就在一定程度上抓住了数据流形的根本所在。所以如果数据集合可以认为是来自于某一个连续可微流形的离散采样，那么只要我们能够抓住数据集合的局部线性,也就等于抓住了数据集合的根本特征。LLE 降维便是基于这种考虑,即数据流形的局部线性的一种非线性降维方法。

LLE 方法的具体做法如下:

(一)寻求数据集的拓扑结构

$f_D(X)=W \in \Gamma_\pi \subset \prod^{N*N}$

其中 Γ_π 为由数据点的邻域确定的稀疏矩阵的全体。对选定的 k 及每个 $x_i \in X$ 考虑 x_i 的邻域 U(i),使得 $x_j^i \in U(i) \subset X$, j=1,2,…,k,是距 x_i 第 j 近的点。则,

$$\Gamma_\pi=\{W \in \prod^{N*N} / \text{若 } j \notin U(i) w_{ij}=0, \text{且} \sum_{j=1}^{N} w_{ij}=1\}$$

(二)拓扑结构的 δ 度量取为

$$\mu_\delta(W)=\sum_{i=1}^{N}(x_i-\sum_{j=1}^{N} w_{ij}x_j)^2=tr(XMX^T) \quad (1)$$

其中 $M=(I-W^T)^T(I-W^T)$,I 为 N*N 的单位矩阵。则 δ 最优结构映射为

$$f_D^\delta(X)=\arg\min_{W \in \Gamma\pi} \{\mu_\delta(W)\}$$

(三)降维准则为

$$\mu^\delta(Y)=\sum_{i=1}^{N}(y_i-Yw_i^T)^2=tr(YMY^T)$$

则,

$$Y^*=\arg\min_{Y \in \Omega_N^d} \mu^\delta(Y)=\arg\min_{Y \in \Omega_N^d} \{tr(YMY^T)\}$$

其中

$\Omega_N^d=\{Y/Y=(y_1,y_2,\cdots,y_N)$,满足 $\sum_{i=1}^{N} y_i=0, YY^T=I_{d*d}, I_{d*d}$ 为单位矩阵$\}$。

可以看出 LLE 方法的关键是设计相应的 δ 度量(1),探求最优结构 $f_D^\delta(X)$,然后采用与 μ_δ 一致的降维准则 μ^δ,寻求相应的降维空间中保持结构的最优表示向量 Y^*。

四、结语

图像识别作为高维数据处理的一个领域，长久以来一直引起人们浓厚的兴趣。然而其出发点一般是基于线性降维的,所寻求的特征是基于数据集合整体的,所以对于大量的数据来讲,这种方法经常是非常繁琐、耗时的,需要大量的细节的探讨。相反通过　(下转第 120 页)

离散数学中的等价关系性质探讨

商丘师范学院计算机科学系　田素霞

[摘　要]等价关系是离散数学中非常重要的内容之一，本文介绍了等价关系的概念，给出了等价关系的若干性质。
[关键词]离散数学　等价关系　等价类

1.预备知识

"离散数学"是计算机专业的重要基础课程和核心课程，等价关系是离散数学中非常重要的内容之一，本文介绍了等价关系的概念，给出了等价关系的若干性质。

定义1 设R为非空集合A上的二元关系，如果R是自反的、对称的和可传递的，则称R为A上的等价关系。

定义2 设R为非空集合A上的等价关系，$\forall x\in A$，令$[x]R=\{y\mid y\in A\wedge xRy\}$，则称$[x]R$为x关于R的等价类，简记为$[x]$。

定义3 设R为非空集合A上的等价关系，以R的所有等价类作元素的集合称为A关于R的商集，记为A/R，即$A/R=\{[x]R\mid x\in A\}$。

2.主要结果

引理1 设R_1,R_2是A上的等价关系，则$R_1\cap R_2$也是A上的等价关系。

引理2 设R是A上的等价关系，C是A关于R的商集，$C=\{C_1,C_2,C_n\}$，则$R=\bigcup_{i=1}^{m}C_i\times C_i$。

证明 对任意的$x,y\in A$，若$\langle x,y\rangle\in\bigcup_{i=1}^{m}C_i\times C_i$，则必存在某个$k(1\leqslant k\leqslant m)$使得$\langle x,y\rangle\in\bigcup_{i=1}^{m}C_k\times C_k$，故有$x\in C_k,y\in C_k$。

令$C_k=[a]_R$，则$x\in[a]_R\wedge y\in[a]_R$，所以$\langle a,x\rangle\in R\wedge\langle a,y\rangle\in R$，所以$\langle x,a\rangle\in R\wedge\langle a,y\rangle\in R$，由于R具有传递性，所以$\langle x,y\rangle\in R$，故$\bigcup_{i=1}^{m}C_i\times C_i\subseteq R$。

反之，对任意的$x,y\in A$，若$\langle x,y\rangle\in R$，则$[x]_R=[y]_R$。令$[x]_R=C_k$，因为$x\in[x]_R,y\in[y]_R$，故$x\in C_k,y\in C_k$，所以$\langle x,y\rangle\in\bigcup_{i=1}^{m}C_i\times C_i$，即$R\subseteq\bigcup_{i=1}^{m}C_i\times C_i$。

综上可知，$R=\bigcup_{i=1}^{m}C_i\times C_i$。

定理1 设R_1,R_2是A上的等价关系，C_1,C_2是A关于R_1,R_2的商集，即$C_1=A/R_1,C_2=A/R_2$，则$R_1\subseteq R_2\Leftrightarrow C_1$的每个等价类包含在$C_2$的一些等价类中。

证明 设$C_1=A/R_1=\{C_{11},C_{12},\cdots,C_{1m}\},C_2=A/R_2=\{C_{21},C_{22},\cdots,C_{2n}\}$，由引理2知$R_1=\bigcup_{i=1}^{m}C_i\times C_i,R_2=\bigcup_{k=1}^{n}C_{2k}\times C_{2k}$。

必要性 若$R_1\subseteq R_2$，当C_{1i}是单个元素集合时，设$C_{1i}=\{x\}$，因为$x\in A$，所以必存在$j(1\leqslant j\leqslant n)$使得$x\in C_{2j}$，即$C_{1i}\subseteq C_{2j}$。设$C_{1i}$不是单个元素集合，则至少有两个元素$x,y\in C_{1i}$，因此$\langle x,y\rangle\in C_{1i}\times C_{1i}$，所以$\langle x,y\rangle\in R_1$。但$R_1\subseteq R_2$，所以$\langle x,y\rangle\in R_2$。由$\langle x,y\rangle\in\bigcup_{k=1}^{n}C_{2k}\times C_{2k}$，必存在某个$j(1\leqslant j\leqslant n)$使得$\langle x,y\rangle\in C_{2j}\times C_{2j}$，即$x\in C_{2j}\wedge y\in C_{2j}$。若除$x,y\in C_{1i}$外还有元素$z\in C_{1i}$，则$\langle x,z\rangle\in C_{1i}\times C_{1i}$，所以$\langle x,z\rangle\in R_1$，由于$R_1\subseteq R_2$，所以$\langle x,z\rangle\in R_2$，故必存在$t(1\leqslant t\leqslant n)$使得$\langle x,z\rangle\in C_{2t}\times C_{2t}$，所以$x\in C_{2t}\wedge z\in C_{2t}$。因为$C_{2j}\cap C_{2t}=\Phi$，故$C_{2j}=C_{2t}$。于是x,y,z在同一等价类$C_{1i}$中，必可得到x,y,z在$C_{2j}$中，依次类推可知$C_{1i}\subseteq C_{2j}$。因此，当$R_1\subseteq R_2$时，$C_1$的等价类必包含在$C_2$的某等价类中。

充分性 若$C_{1i}\subseteq C_{2k}$对每个$i=1,2,\cdots,m$有$C_{1i}\times C_{1i}\subseteq C_{2k}\times C_{2k}$，于是$R_1=\bigcup_{i=1}^{m}C_{1i}\times C_{1i}\subseteq\bigcup_{k=1}^{n}C_{2k}\times C_{2k}=R_2$。

定理2 设R_1,R_2是A上的等价关系，$|A/R_1|=r_1$，$|A/R_2|=r_2$，则$|A/(R_1\cap R_2)|\leqslant r_1\cdot r_2$。

证明 因为R_1,R_2是A上的等价关系，由引理1知$R_1\cap R_2$也是A上的等价关系，设$R_1,R_2,R_1\cap R_2$分别诱导的划分为$C_1=\{C_{11},C_{12},\cdots,C_{1m}\},C_2=\{C_{21},C_{22},\cdots,C_{2n}\},C_3=\{C_{31},C_{32},\cdots,C_{3p}\}$，任取$C_{3k}\in C_3(1\leqslant k\leqslant p)$，因为$C_{3k}\neq\Phi$，故必存在某$x\in A\wedge x\in C_{3k}$。因为$C_1、C_2$均为A的划分，所以必存在$C_{1i}\in C_1$使$x\in C_{1i}$，且存在$C_{2j}\in C_2$使$x\in C_{2j}$，即$x\in C_{1i}\cap C_{2j}$。

若$C_{3k}=\{x\}$，则$C_{3k}\in C_{1i}\cap C_{2j}$。若另有$y\in C_{3k}$，因$C_{3k}$是A关于$R_1\cap R_2$的等价类，故有$x\in C_{3k}\wedge y\in C_{3k}$，所以$\langle x,y\rangle\in R_1\cap R_2$，即$\langle x,y\rangle\in R_1\wedge\langle x,y\rangle\in R_2$。

由于$x\in C_{1i}\wedge x\in C_{2j}$，故有$y\in C_{1i}\wedge y\in C_{2j}$，即$y\in C_{1i}\cap C_{2j}$。

由y的任意性得$C_{3k}\subseteq C_{1i}\cap C_{2j}$且$C_{3k}\subseteq C_{1i}\cap C_{2j}$是唯一的。

若不然，如果还有$C_{3k}\subseteq C_{1s}\cap C_{2t}$，则对任意$x\in C_{3k}$必有$x\in C_{1i}\wedge x\in C_{2j}\wedge x\in C_{1s}\wedge x\in C_{2t}$，所以$x\in(C_{1i}\cap C_{1s})\wedge x\in(C_{2j}\cap C_{2t})$。因$C_{1i}、C_{1s}$是$C_1$中的等价类，$C_{2j}、C_{2t}$是$C_2$中的等价类，在每个划分中如果两个等价类有公共元素，这两个等价类相等，故有$C_{1i}=C_{1s},C_{2j}=C_{2t}$，即$C_{3k}$只能属于一个$C_{1i}\cap C_{2j}$中。

另外，C_3的两个不同元素必是不同的$C_{1i}\cap C_{2j}$的子集。

设$C_{3k}\in C_3,C_{3p}\in C_3$，若$C_{3k}\subseteq C_{1i}\cap C_{2j}\wedge C_{3p}\subseteq C_{1i}\cap C_{2j}$，因$C_{3k}\cap C_{3p}=\Phi$，故必有$x\in C_{3k}\wedge y\in C_{3k}$但$x\notin C_{3p}\wedge y\notin C_{3p}$，因$x\in C_{3k}$，所以$x\in C_{1i}\wedge x\in C_{2j}$，$x\in C_{3p}$，所以$y\in C_{1i}\wedge y\in C_{2j}$，故$x\in C_{3k}\wedge y\in C_{3p}$，所以$x\in C_{1i}\wedge y\in C_{1i}\wedge x\in C_{2j}\wedge y\in C_{2j}$，从而$\langle x,y\rangle\in R_1\wedge\langle x,y\rangle\in R_2$，所以$\langle x,y\rangle\in R_1\cap R_2$，故$C_{3k}=C_{3p}$。

于是对任意的$C_{3k}\in C_3$必有唯一的$C_{1i}\cap C_{2j}$使得$C_{3k}\subseteq C_{1i}\cap C_{2j}$且对任一$C_{1i}\cap C_{2j}$至多只有一个$C_{3k}$满足$C_{3k}\subseteq C_{1i}\cap C_{2j}$。因为$|C_1|=r_1,|C_2|=r_2$，故$i=1,2,\cdots,r_1,j=1,2,\cdots,r_2$，$C_{1i}\cap C_{2j}$至多有$r_1\cdot r_2$个不同的交集，故$|C_3|$至多为$r_1\cdot r_2$，即$|A/(R_1\cap R_2)|\leqslant r_1\cdot r_2$。

参考文献

[1]耿素云，屈婉玲.离散数学[M].北京:高等教育出版社，2004.

[2]张文修，吴伟志，梁吉业等.粗糙集理论与方法[M].北京:科学出版社，2001.

[3]胡宝清.模糊理论基础[M].武汉大学出版社，2004.

[4]左孝凌，李为鉴，刘永才.离散数学.上海:上海科学技术文献出版社，1982.

[5]BEMARD K，ROBERT C，BUSBY.离散数学结构.北京:清华大学出版社，1997.

(上接第119页)　高维化数据，利用非线性降维手段，则可以探求数据的本征特征，在寻求特征的过程中，不需要考虑整个数据集合，从而可大大减少计算负担，改善常用方法的效能。另外，将非线性降维手段应用于图像识别问题也为图像识别问题的解决提供了一种崭新的思考方式，它的提出极大地拓展了关于降维的认识，引起了人们广泛的注意。一个直接的后果就是，人们开始更加关注数据集所蕴含的内蕴特征，通过内蕴特征的探讨来研究关于降维的问题，所以非线性降维方法的提出对于降维问题的发展来讲具有重大意义。

参考文献

[1]Zhenyue Zhang and Hongyuan Zha. Local Linear Smoothing for Nonlinear Manifold Learning [J].Technical Report, CSE, Penn State Univ, 2003.

[2]Sam T.Roweis and Lawrence K.Saul, Nonlinear Dimensionality Reduction byEmbedding[J].Science, 2000.

[3]David L.Donoho and Carne Grimes, Hessian Eigenmaps: new locally linear embeddingtechniques for high-dimensional data [J].Technical Report Department of Statistics, Stanford UnivErsity,2003.

[4]Yoshua Bengio,Jean-Francois Paiement etc. Out-of-Sample Extensions for LLE, Isomap, MDS,Eigenmaps, and Spectral Clustering, [J]. Technical Report 1238, Departement Informatique et Recherche Operationnelle, 2003.