SVD 及 PCA 降维方法研究

靳宗明，网络空间安全学院，南开大学，天津，中国，zongming_jin@mail.nankai.edu.cn

本文受 Web 大数据挖掘课程推荐系统相关内容启发，数据挖掘中的推荐系统会利用 SVD 奇异值分解进行数据降维，所以我会以 SVD 为出发点讲解降维的相关内容，然后具体讲解目前最广泛利用的降维方法 PCA，最终介绍 PCA 和 SVD 之间的区别和联系。

ABSTRACT 降维是机器学习中很重要的一种思想。在机器学习中经常会碰到一些高维的数据集，而在高维数据情形下会出现数据样本稀疏，距离计算等困难，这类问题是所有机器学习方法共同面临的严重问题，称之为“ 维度灾难 ”。另外在高维特征中容易出现特征之间的线性相关，这也就意味着有的特征是冗余存在的。所以降维就成了数据预处理必须的一个步骤，本文就以 SVD 奇异值分解为出发点，讲解降维的相关内容，然后具体详细讲解目前最广泛使用的降维方法主成分分析 PCA，最后简述 SVD 和 PCA 之间的区别和关系。

INDEX TERMS 降维，奇异值分解，主成分分析。

I. 引言

降维是将数据映射到较低的维度空间，从而丢弃数据的无用信息变量，或者检测数据所在的子空间[1]。降维作为一种数据可视化和提取核心低维特征（例如，从其高维图像表示中的对象的二维特征）的方法由来已久。在某些情况下，我们只需要某些低维特征，然而我们拿到的数据是极高维的数据。降维不仅可以让我们获得低维度特征信息，还可以使我们获得更好的推理模型。

降维可以通过两种不同的方式进行：通过仅保留原始数据集中最相关的变量（此技术称为特征选择），或通过利用输入数据的冗余性并通过查找较小的新变量集来实现。输出的变量组合，基本上包含与输入变量相同的信息（此技术称为降维）。

降维在数学统计分析中并不是一种新技术。实际上，最广泛使用的降维技术之一是主成分分析（PCA），Karl Pearson[2]在1901就提出了相关概念。 关键思想是找到一个新的坐标系，在该坐标系中输入数据可以用更少的变量表示而没有明显的误差。这个新的坐标可以是全局的，也可以是局部的，并且可以实现非常不同的属性。近年来，随着数据的爆炸式增长以及越来越强大的计算资源，统计、计算机科学和应用数学领域的许多研究人员都开始关注这一问题，他们开发了各种处理降维问题的计算技术[3][4][5]。

本文将从最简单的降维方法SVD出发，然后讲解目前最广泛使用的降维方法PCA，然后简单讲解PCA核函数扩展，最后总结SVD和PCA的区别和联系。

II. 奇异值分解SVD

A. 特征值和特征向量

特征值和特征向量的定义如下：

$$Ax = \lambda x \tag{2.1}$$

其中 A 是一个 $n \times n$ 的实对称矩阵， x 是一个 n 维向量，那么 λ 是矩阵 A 的特征值， x 是矩阵 A 的特征值 λ 所对应的特征向量。通过计算矩阵的特征值和特征向量，就是对矩阵做了特征分解， n 个特征值 $\lambda_1 \le \lambda_2 \le ... \le \lambda_n$ 对应的特征向量 $\{w_1, w_2, ...w_n\}$ 是线性无关的，那么矩阵 A 就可以用下式特征分解表示：

$$A = W \Sigma W^{-1} \tag{2.2}$$

一般我们会把 W 的 n 个特征向量标准化，即满足 $\|w_i\|_2 = 1$ ，此时 W 的 n 个特征向量就是一组标准正交基，满足 $W^T W = I$ ，即 $W^T = W^{-1}$ ，也就是酉矩阵，然后就可以将特征分解表达式写成：

$$A = W \Sigma W^T \tag{2.3}$$

这里需要注意 A 是一个 $n \times n$ 的方阵，将上述分解扩展到非奇异矩阵范围，那就是奇异值分解SVD。

B. 奇异值分解

SVD也是对矩阵进行分解，但是和特征分解不同，SVD并不要求要分解的矩阵为方阵。假设 A 是一个 $m \times n$ 的矩阵，那么矩阵SVD分解定义为[6]：

$$A = U \Sigma V^T \tag{2.4}$$

其中 U 是一个 $m \times m$ 的矩阵， Σ 是一个 $m \times n$ 的对角矩阵，对角线上的元素称为奇异值， V 是一个

$n\times n$ 的矩阵。U 和 V 都是酉矩阵，即满足 $U^TU=I, V^TV=I$，如图1所示。

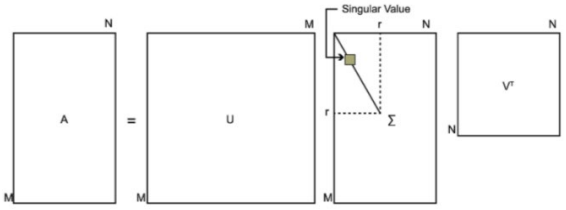

图 1. SVD特征值分解

中间的特征值对角阵 $\sum$ 是有序的，为了实现降维压缩，$\sum$ 可以取前k行k列，U 可以取前k列，V 可以取前k行，然后计算SVD分解的近似值，这样矩阵的奇异值分解空间会大大降低，同时也能达到降噪的目的，如图2所示。

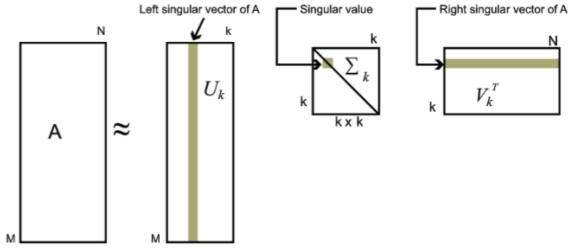

图 2. SVD近似分解

C. SVD在图像压缩中的应用

从图3所示SVD图片压缩结果可以看出来，奇异值可以被看作成一个矩阵的代表值，或者说奇异值能代表这个矩阵的信息，当奇异值越大时，它代表的信息越多。因此，我们取前面若干个最大的奇异值，就可以基本上还原数据本身。

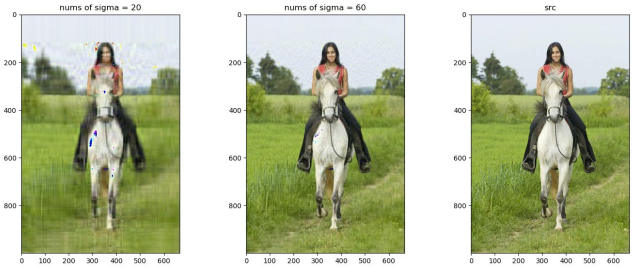

图 3. SVD图像压缩

III. 主成分分析PCA

Principal Component Analysis(PCA)是最常用的线性降维方法，它的目标是通过某种线性投影，将高维的数据映射到低维的空间中表示，并期望在所投影的维度上数据的方差最大，以此使用较少的数据维度，同时保留住较多的原数据点的特性。

下面是一个简单的例子，将原始二维数据降到一维上，如图4所示。我们希望找到某一个维度方向，它可以代表这两个维度的数据。图中列了两个向量方向，u_1 和 u_2，从直观上看 u_1 可以更好的代表原始数据集。

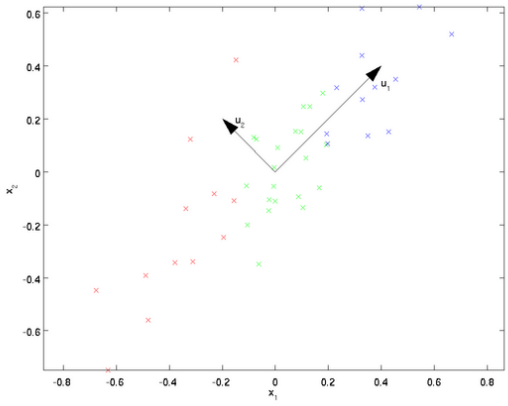

图 4. 主成分直观选择

为什么 u_1 比 u_2 好呢？可以有两种解释，第一种解释是样本点到这个直线的距离足够近，第二种解释是样本点在这个直线上的投影能尽可能的分开。假如我们把n'从1维推广到任意维，则我们的希望降维的标准为：样本点到这个超平面的距离足够近,或者说样本点在这个超平面上的投影能尽可能的分开。基于上面的两种标准，我们可以得到PCA的两种等价推导：基于最小投影距离、基于最大投影方差。

A. 基于最小投影距离推导

假设m个n维数据 $(x^{(1)},x^{(2)},...,x^{(m)})$ 都已经进行了中心化，即 $\sum_{i=1}^{m}x^{(i)}=0$。经过投影变化后得到的新坐标系为 $\{w_1,w_2,...,w_n\}$，其中 w 是标准正交基，即 $||w||_2=1, w_i^Tw_j=0$。

如果我们将数据从n维降到n'维，即丢弃新坐标系中的部分坐标，则新的坐标系为 $\{w_1,w_2,...,w_{n'}\}$，样本点 $x^{(i)}$ 在n'维坐标系中的投影为：$z^{(i)}=(z_1^{(i)},z_2^{(i)},...,z_{n'}^{(i)})^T$。其中 $z_j^{(i)}=w_j^Tx^{(i)}$ 是 $x^{(i)}$ 在低维坐标系里第j维的坐标。如果我们用 $z^{(i)}$ 来恢复原始数据 $x^{(i)}$，则得到的恢复数据 $\overline{x}^{(i)}=\sum_{j=1}^{n'}z_j^{(i)}w_j=Wz^{(i)}$，其中 W 为标准正交基组成的矩阵。

现在我们考虑整个样本集，我们希望所有的样本到这个超平面的距离足够近，即最小化 $\sum_{i=1}^{m}||\overline{x}^{(i)}-x^{(i)}||_2^2$，将这个式子进行整理，可以得到：

$$\begin{aligned}\sum_{i=1}^{m}||\bar{x}^{(i)}-x^{(i)}||_2^2 &= \sum_{i=1}^{m}||Wz^{(i)}-x^{(i)}||_2^2 \\ &= \sum_{i=1}^{m}(Wz^{(i)})^T(Wz^{(i)})-2\sum_{i=1}^{m}(Wz^{(i)})^Tx^{(i)}+\sum_{i=1}^{m}x^{(i)T}x^{(i)} \\ &= \sum_{i=1}^{m}z^{(i)T}z^{(i)}-2\sum_{i=1}^{m}z^{(i)T}W^Tx^{(i)}+\sum_{i=1}^{m}x^{(i)T}x^{(i)} \\ &= \sum_{i=1}^{m}z^{(i)T}z^{(i)}-2\sum_{i=1}^{m}z^{(i)T}z^{(i)}+\sum_{i=1}^{m}x^{(i)T}x^{(i)} \\ &= -\sum_{i=1}^{m}z^{(i)T}z^{(i)}+\sum_{i=1}^{m}x^{(i)T}x^{(i)} \\ &= -tr(W^T(\sum_{i=1}^{m}x^{(i)}x^{(i)T})W)+\sum_{i=1}^{m}x^{(i)T}x^{(i)} \\ &= -tr(W^TXX^TW)+\sum_{i=1}^{m}x^{(i)T}x^{(i)}\end{aligned} \tag{3.1}$$

注意 $\sum_{i=1}^{m}x^{(i)}x^{(i)T}$ 是数据集的协方差矩阵，W 的每一个向量 w_j 是标准正交基，而 $\sum_{i=1}^{m}x^{(i)T}x^{(i)}$ 是一个常量。最小化上式等价于:

$$\underbrace{arg\ min}_{W} -tr(W^TXX^TW)\ \ s.t.W^TW=I \tag{3.2}$$

利用拉格朗日函数可以得到

$$J(W)=-tr(W^TXX^TW+\lambda(W^TW-I)) \tag{3.3}$$

对 W 求导有 $-XX^TW+\lambda W=0$，整理得:

$$XX^TW=\lambda W \tag{3.4}$$

W 是 XX^T 的n'个特征向量组成的矩阵，而 λ 是 XX^T 的若干个特征值组成的矩阵，特征值在主对角线上，其余值为0。当我们将数据从n维降到n'维时，需要选择最大的n'个特征值对应的特征向量。这n'个特征向量组成的矩阵 W 即我们需要的矩阵。对于原始数据，我们只需要用 $z^{(i)}=W^Tx^{(i)}$ 就可以把原始数据降维到最小投影距离的n'维数据上。

B. 基于最大投影方差推导

假设m个n维数据 $(x^{(1)},x^{(2)},...,x^{(m)})$ 都已经进行了中心化，即 $\sum_{i=1}^{m}x^{(i)}=0$。经过投影变化后得到的新坐标系为 $\{w_1,w_2,...,w_n\}$，其中 w 是标准正交基，即 $||w||_2=1, w_i^Tw_j=0$。

如果我们将数据从n维降到n'维，即丢弃新坐标系中的部分坐标，则新的坐标系为 $\{w_1,w_2,...,w_{n'}\}$，样本点 $x^{(i)}$ 在n'维坐标系中的投影为：$z^{(i)}=(z_1^{(i)},z_2^{(i)},...,z_{n'}^{(i)})^T$。其中 $z_j^{(i)}=w_j^Tx^{(i)}$ 是 $x^{(i)}$ 在低维坐标系里第j维的坐标。对于任意一个样本 $x^{(i)}$，在新的坐标系中的投影为 $W^Tx^{(i)}$，在新坐标系中的投影方差为 $W^Tx^{(i)}x^{(i)T}W$，要使所有的样本的投影方差和最大，也就是最大化 $W^Tx^{(i)}x^{(i)T}W$ 的迹，即:

$$\underbrace{arg\ max}_{W} tr(W^TXX^TW)\ \ s.t.W^TW=I \tag{3.5}$$

利用拉格朗日函数可以得到

$$J(W)=tr(W^TXX^TW+\lambda(W^TW-I)) \tag{3.6}$$

对 W 求导有 $XX^TW+\lambda W=0$，整理得:

$$XX^TW=(-\lambda)W \tag{3.7}$$

同A基于最小投影距离推导，W 是 XX^T 的n'个特征向量组成的矩阵，而 λ 是 XX^T 的若干个特征值组成的矩阵，特征值在主对角线上，其余值为0。当我们将数据从n维降到n'维时，需要选择最大的n'个特征值对应的特征向量。这n'个特征向量组成的矩阵 W 即我们需要的矩阵。对于原始数据，我们只需要用 $z^{(i)}=W^Tx^{(i)}$ 就可以把原始数据降维到最小投影距离的n'维数据上。

C. PCA算法流程

从上面两节我们可以看出，求样本 $x^{(i)}$ 的n'维的主成分其实就是求样本集的协方差矩阵 XX^T 的前n'个特征值对应特征向量矩阵 W，然后对于每个样本 $x^{(i)}$，做 $z^{(i)}=W^Tx^{(i)}$ 变换，即达到PCA降维的目的。

下面是一个具体的算法流程:

（1）对所有的样本进行中心化：$x^{(i)}=x^{(i)}-\frac{1}{m}\sum_{j=1}^{m}x^{(j)}$

（2）计算样本的协方差矩阵 XX^T

（3）对矩阵 XX^T 进行特征值分解

（4）取出最大的n'个特征值对应的特征向量 $(w_1,w_2,...,w_{n'})$，将所有的特征向量标准化后，组成特征向量矩阵 W

（5）对样本集中的每一个样本 $x^{(i)}$，转化为新的样本 $z^{(i)}=W^Tx^{(i)}$

（6）输出降维后的样本集 $D'=(z^{(1)},z^{(2)},...,z^{(m)})$

在很多应用场景中，我们不指定降维后的n'值，而是指定一个降维到的主成分比重阈值 t。这个阈值 t 在在 $(0,1]$ 之间，假如n个特征值为 $\lambda_1\geq\lambda_2\geq...\geq\lambda_n$，可以通过下式求n'。

$$\frac{\sum_{i=1}^{n'}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \geq t \tag{3.8}$$

D. 核主成分分析KPCA

在上面的PCA算法中，我们假设存在一个线性的超平面，可以让我们对数据进行投影。但是有些时候，数据不是线性的，不能直接进行PCA降维。这里就需要用到和支持向量机一样的核函数的思想，先把数据集从n维映射到线性可分的高维 $N > n$，然后再从N维降维到一个低维度n’，满足 $n' < n < N$。

使用了核函数的主成分分析一般称之为核主成分分析(Kernelized PCA, 简称KPCA)。假设高维空间的数据是由n维空间的数据通过映射 ϕ 产生，对n维空间的特征分解：

$$\sum_{i=1}^{m} x^{(i)} x^{(i)T} W = \lambda W \tag{3.9}$$

映射为：

$$\sum_{i=1}^{m} \phi(x^{(i)})\phi(x^{(i)})^T W = \lambda W \tag{3.10}$$

通过在高维空间进行协方差矩阵的特征值分解，然后用和PCA一样的方法进行降维。一般来说，映射 ϕ 不用显式的计算，而是在需要计算的时候通过核函数完成。由于KPCA需要核函数的运算，因此它的计算量要比PCA大很多。

IV. 总结

SVD是一种矩阵分解的方法，相当于因式分解，他的目的纯粹就是将一个矩阵拆分成多个矩阵相乘的形式。PCA是一种广泛使用的降维方法，从多维数据中找出主成分来低误差的表示原始多维数据。

PCA在操作过程中要计算协方差矩阵，当样本数和特征数很多的时候，计算量是非常大的。SVD也可以得到协方差矩阵 XX^T 最大的k个特征向量张成的矩阵，但是SVD有一个好处，那就是有一些SVD的实现算法可以先不求协方差矩阵 XX^T，也能求出左奇异矩阵 U，也就是说PCA算法可以不用做特征分解，而是通过做SVD来完成，而且矩阵的奇异值分解迭代计算比协方差矩阵的特征值分解更快更准确。

REFERENCES

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, A learning algorithm for Boltzmann machines," Cognitive Science, vol. 9, 1985.
[2] Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, 1901, 2, 559-572
[3] Carreira-Perpiñán, M. A. A review of dimension reduction techniques Dept. Computer Science, Univ. Sheffield, 1997
[4] Fodor, I. K. A survey of dimension reduction techniques Lawrence Livermore Natl. Laboratory, 2002
[5] van der Mateen, L.; Postma, E. & van den Herik, J. Dimensionality Reduction: A Comparative Review Tilburg Centre for Creative Computing, Tilburg Univ., 2009
[6] Abdi, H. Salkind, N. (ed.) Encyclopedia of measurements and statistics Singular value decomposition (SVD) and Generalized Singular Value Decomposition (GSVD) Sage Publications, 2007, 907-912