高维数据空间的性质及度量选择

何进荣 丁立新 胡庆辉 李照奎
(武汉大学计算机学院软件工程国家重点实验室 武汉 430072)

摘 要 高维数据分析是机器学习和数据挖掘研究中的主要内容,降维算法通过寻找数据表示的最优子空间来约减维数,在降低计算代价的同时,也提高了后续分类或者聚类算法的性能,从而成为高维数据分析的有效手段。然而,目前缺乏高维数据分析的理论指导。对高维数据空间的统计和几何性质进行了综述,从不同的角度给出了高维数据空间中"度量集中"现象的直观解释,并讨论了通过度量选择的方式来提高经典的基于距离度量的机器学习算法在分析高维数据时的性能。实验表明,分数距离度量方式可以显著提高K近邻和Kmeans算法的性能。

关键词 高维数据,维数灾难,度量集中
中图法分类号 TP181 **文献标识码** A

Properties of High-dimensional Data Space and Metric Choice

HE Jin-rong DING Li-xin HU Qing-hui LI Zhao-kui
(State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan 430072, China)

Abstract High-dimensional data analysis is the core task of machine learning and data mining. By finding optimal subspace for data representation, dimensionality reduction algorithms can reduce computational cost and improve the performance of subsequent classification or clustering algorithms, leading to effective techniques for high-dimensional data analysis. However, there is very little guidance for theoretical analysis on high-dimensional data. This paper reviewed some statistical and geometrical properties of high-dimensional data space, and gave some intuitive explanations on "concentration of measure" phenomenon from different perspectives. In order to improve performances of classical machine learning algorithms based on distance metric, this paper discussed the effects of metric choice on high-dimensional data analysis. Empirical results show that fractional distance metric can improve performances of K Nearest Neighbor and Kmeans significantly.

Keywords High-dimensional data, Curse of dimensionality, Concentration of measure

1 引言

随着大数据时代的来临,高维数据分析已经成为应用驱动的热点研究问题[1,2]。机器学习算法在直接处理高维数据时,无可避免会遇到"维数灾难"问题[3],即要达到同样的精度,学习模型所需要的样本数随着样本维数的增加呈指数增长,在算法应用研究中表现为"小样本"问题[4],在数学分析上则表现为"度量集中"现象[5]。

由于在计算机处理中,数据通常是作为向量进行运算,因此高维数据空间本质上就是向量空间。目前关于高维数据空间性质的讨论,主要集中于数据库技术中相似性检索方法的研究[6-9],当数据库中每条记录的属性较多时,欧氏度量下的最近邻就失去意义。文献[10]在分析基因和蛋白质表达数据时,以此为例解释了高维数据数据空间中的一些性质。文献[11,12]以超光谱数据分析为例,对高维数据空间的特性进行了研究,并讨论了降维算法在高维数据分类中的必要性。文献[13]在化学信息学中分析分子之间的相似性度量时,讨论高维分子描述空间的一些现象,比如空空间现象、距离度量集中现象等等。

然而,国内外研究文献中关于高维数据空间的性质方面的专题研究非常少见,从而导致对高维数据的理论分析和算法设计缺乏指导。基于此,本文从几何和统计的角度,总结归纳了高维数据空间的性质,当数据空间的样本维数无限增长时,本文从各个侧面给出了"度量集中"现象的直观解释。最后通过实验分析,讨论了不同的距离度量对机器学习算法性能的影响。

2 高维数据空间的统计性质

大数定律和集中不等式是分析高维数据空间统计性质的基本工具。下面给出分析高维数据空间统计性质的相关定义和结论。

到稿日期:2013-05-19 返修日期:2013-09-16 本文受中央高校基本科研业务费专项资金(2012211020209),广东省省部产学研结合专项(2011B090400477),珠海市产学研合作专项资金(2011A050101005,2012D0501990016),珠海市重点实验室科技攻关项目(2012D0501990026)资助。
何进荣(1984—),男,博士生,主要研究方向为特征提取、高维数据分析,E-mail:hejinrong@whu.edu.cn;丁立新(1967—),男,教授,博士生导师,主要研究方向为智能信息处理、云计算;胡庆辉(1976—),男,博士生,主要研究方向为机器学习;李照奎(1976—),男,博士生,主要研究方向为机器学习。

2.1 高维数据空间中的度量

定义 1 d 维向量空间中的某点 $x=(x^{(1)},x^{(2)},\cdots,x^{(d)})\in R^d$ 的 ℓ^p 范数定义为：

$$\|x\|_p=(\sum_{i=1}^{d}(x^{(i)})^p)^{\frac{1}{p}}$$

当 $p<1$ 时，该范数又称为分数范数[14]。

定理 1[15] 任给 $k>0$，如果 $\lim\limits_{d\to\infty}\frac{\mathrm{var}(\|x\|_p^k)}{E(\|x\|_p^k)^2}=0$，则 $\lim\limits_{d\to\infty}P(\frac{\max(\|x\|_p)-\min(\|x\|_p)}{\min(\|x\|_p)}<\varepsilon)=1$。

定理 2[16] 假设样本数目 n 足够大，使得 $E(\|x\|_p^k)^2\in[\min\limits_{1\leqslant i\leqslant n}\|x_i\|_p^k,\max\limits_{1\leqslant i\leqslant n}\|x_i\|_p^k]$ 成立。如果 $\lim\limits_{d\to\infty}P(\frac{\max(\|x\|_p)-\min(\|x\|_p)}{\min(\|x\|_p)}<\varepsilon)=1$，则 $\lim\limits_{d\to\infty}\frac{\mathrm{var}(\|x\|_p^k)}{E(\|x\|_p^k)^2}=0$。

此处 $\frac{\max(\|x\|_p)-\min(\|x\|_p)}{\min(\|x\|_p)}$ 称为范数的相对差异，$RV_d=\frac{\mathrm{var}(\|x\|_p^k)}{E(\|x\|_p^k)^2}$ 称为相对方差。

该定理表明，随着数据空间维数的增加，样本点范数的相对差异和相对方差都趋于 0。在高维数据空间中，某个样本点到其最近邻居点和最远邻居点之间的距离趋于相等，从而导致一些基于距离度量的机器学习算法性能降低。这种现象通常称为“度量集中”，最早由 Milman 在描述高维概率分布时引入[5]。随着维数的增长，欧氏空间中任意两点间距离度量的差异性变得越来越弱，从而导致数据趋向于均匀分布。

已有相关研究表明[17-20]，高维空间中数据点之间的相似性度量对 ℓ^p 范数中 p 值的选取比较敏感。图 1 显示了 $\|x\|_p=1$ 在二维情形下的图形，p 值越小，其边界越靠近坐标轴，在机器学习算法中越容易导致稀疏解。

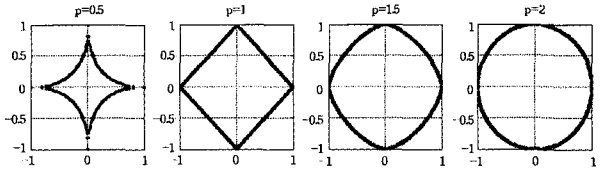


图 1 不同 ℓ^p 范数下的单位圆

定理 3[20] 给定 n 个 d 维样本点，其每个分量相互独立且来自于均匀分布，则存在常数 C，使得

$$C\cdot\sqrt{\frac{1}{2p+1}}\leqslant\lim_{d\to\infty}E(\frac{dist_{\max}^p-dist_{\min}^p}{dist_{\min}^p})\leqslant C\cdot(n-1)\cdot\sqrt{\frac{1}{2p+1}}$$

这里 $dist_{\max}^p$ 和 $dist_{\min}^p$ 分别表示 n 个样本点之间的最大 ℓ^p 距离度量和最小 ℓ^p 距离度量。

此定理表明，由分数距离度量所计算的样本点之间的相对差异性更大。

2.2 高斯空间

定义 2 d 维标准高斯空间由各个分量相互独立且来源于标准正态分布的 d 维随机向量构成，即

$$X=\{(x^{(1)},x^{(2)},\cdots,x^{(d)}):x^{(i)}\sim N(0,1),i=1,\cdots,d\}$$

其概率密度函数为

$$p(x)=\frac{1}{(2\pi)^{\frac{d}{2}}}\exp(-\frac{\|x\|^2}{2})$$

定理 4 $\forall x\in X,E(\|x\|_2^2)=d\cdot E((x^{(i)})^2)=d$。

定理 5[21] $\forall x\in X,\sum\limits_{i=1}^{d}(x^{(i)})^2\sim\chi_d^2$，且 $\lim\limits_{d\to+\infty}\sum\limits_{i=1}^{d}(x^{(i)})^2\sim N(0,1)$。

定理 6[22] $\forall x\in X,\|x\|_2\dot{\sim}N(\sqrt{d-\frac{1}{2}},\frac{1}{2})\approx N(\sqrt{d},\frac{1}{2})$。即当维数 d 趋向于无穷大时，$\|x\|_2$ 近似服从正态分布。

该定理表明，d 维标准高斯空间中的绝大部分概率集中于一个超球壳上，即

$$\sqrt{d}-\varepsilon<\|x\|_2<\sqrt{d}+\varepsilon$$

例如：当 $d=1000,\varepsilon=3.46$ 时：

$$P(28.16<\|x\|<35.08)\geqslant 1-10^{-6}$$

3 高维数据空间的几何性质

投影法和截面法是分析高维数据空间几何性质的基本工具。类比于平面几何和立体几何中的概念，我们可以定义如下的高维几何体，并推导出高维几何体的一些奇特性质，这些性质可以看作是“度量集中”现象的几何直观解释。

3.1 超立方体

3.1.1 相关定义

定义 3 中心在坐标原点、边长为 $2r$ 的 d 维超立方体 $C^d(r)$ 为

$$C^d(r)=\{(x^{(1)},\cdots,x^{(d)})\mid -r\leqslant x^{(i)}\leqslant r\text{ for all }i\}\triangleq[-r,r]^d$$

特别地，0 维立方体就是一个点，1 维立方体是一条线段，二维立方体是一个正方形。显然，超立方体 $C^d(r)$ 共有 2^d 个顶点，$2d$ 个 $d-1$ 维侧面，$2^k A_d^k$ 个 $d-k$ 维的侧面（A_d^k 表示 d 中取出 k 个的排列数），且每个侧面可看作是超立方体。超立方体的顶点为 $v=(\pm r,\cdots,\pm r)$，到坐标原点的距离为 $r\sqrt{d}$。单位超立方体可以表示为 $C^d(\frac{1}{2})$，其直径（超立方体上任意两点之间距离的最大值）为 $\sqrt{d}$。

定义 4 超立方体 $C^d(r)$ 的赤道面为

$$H_0=\{x:\sum_{i=1}^{d}x^{(i)}=0\}$$

则 $H_c=\{x:\sum\limits_{i=1}^{d}x^{(i)}=c\}$ 就表示与 H_0 平行的超平面，点 $x=(x^{(1)},\cdots,x^{(d)})$ 到 H_0 的垂直距离为

$$dist(x,H_0)=\frac{1}{\sqrt{d}}|\sum_{i=1}^{d}x^{(i)}|$$

定义 5 d 维超立方体 $C^d(r)$ 的体积为：

$$V(C^d(r))=\underbrace{(2r)\times(2r)\times\cdots\times(2r)}_{d\ times}=(2r)^d$$

注意到，超立方体的体积随着维数呈指数增长。

定义 6 d 维超立方体 $C^d(r)$ 的表面积为其所有侧面的体积之和，即

$$S(C^d(r))=(2d)\times V(C^{d-1}(r))$$

3.1.2 重要性质

定理 7 $\lim\limits_{d\to\infty}\cos\langle\frac{v}{\|v\|},e_i\rangle=0$，其中 e_i 表示坐标轴上的单位向量。

该定理说明，随着维数的增长，超立方体的对角线逐渐正交于所有的坐标轴。

定理 8 $\forall x=(x^{(1)},\cdots,x^{(d)})\in R^d$，且 $x^{(i)}\overset{i.i.d}{\sim}U(-0.5,0.5)$，则

$$E(dist^2(x,H_0))=\frac{1}{12}$$

证明：根据定义 4，可知

$$E(\frac{1}{\sqrt{d}}|\sum_{i=1}^{d}x^{(i)}|)^2=\frac{1}{d}Var(\sum_{i=1}^{d}x^{(i)})=\frac{1}{d}\sum_{i=1}^{d}Var(x^{(i)})$$

$$=\frac{1}{d}\sum_{i=1}^{d}(\frac{1}{12})=\frac{1}{12}$$

该定理说明，单位超立方体内任一点到其赤道面 H_0 的平均平方距离为$\frac{1}{12}$。

定理 9 $\lim\limits_{d\to\infty}V(C^d(r))=\begin{cases}0, & r<\frac{1}{2}\\ 1, & r=\frac{1}{2}\\ \infty, & r>\frac{1}{2}\end{cases}$

定理 10 $\lim\limits_{d\to\infty}V(C^d(\frac{1}{2})-C^d(\frac{1}{2}-\frac{\varepsilon}{2}))=1,\forall\varepsilon\in(0,1)$。

该定理表明，单位超立方体的体积主要集中在其外壳上。这也启发我们，原始高维数据的某个度量实际上分布在某个维数较低的子空间，这也是降维算法实施的依据之一。

定理 11 $\lim\limits_{d\to\infty}V(C^d(\frac{1}{2})-C^d(\frac{1}{2}-\frac{\varepsilon}{2d}))=1-e^{-\varepsilon},\forall\varepsilon\in(0,d)$。

定理 12 $\lim\limits_{d\to\infty}\frac{V(C^d(r))}{S(C^d(r))}=0,\forall r\in(0,+\infty)$。

3.2 超球体

3.2.1 相关定义

定义 7 圆心在坐标原点、半径为 r 的 d 维超球体定义为

$$B^d(r)=\{(x^{(1)},\cdots,x^{(d)})|\sum_{i=1}^{d}(x^{(i)})^2\leqslant r^2\}$$

其体积[23]为

$$V(B^d(r))=\frac{2r^d\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2})}$$

这里 $\Gamma(s)=\int_0^{\infty}e^{-t}t^{s-1}dt$ 是 Gamma 函数，且 $\Gamma(\frac{1}{2})=\sqrt{\pi},\Gamma(1)=1,\Gamma(x+1)=x\Gamma(x)$。

特别地，单位超球体的体积为：

$$V(B^d(1))=\begin{cases}\frac{\pi^p}{p!}, & d=2p\\ \frac{2\cdot(2\pi)^p}{1\cdot3\cdot5\cdots(2p+1)}, & d=2p+1\end{cases}$$

显然，$V(B^d(r))=V(B^d(1))\cdot r^d$。于是，超立方体 $C^d(r)$ 的外接球为 $B^d(r\sqrt{d})$，内切球为 $B^d(r)$。

定义 8 超球体 $B^d(r)$ 的表面称为球面，记作 $\partial(B^d(r))$，即

$$\partial(B^d(r))=\{(x^{(1)},\cdots,x^{(d)})|\sum_{i=1}^{d}(x^{(i)})^2=r^2\}$$

注意，$\partial(B^d(r))$ 可以看作是 d 维欧氏空间中的 $d-1$ 维流形。从拓扑观点来看，d 维球面可以表示为 $\partial(B^{d+1}(a))=R^d\cup\{\infty\}$，其局部同构于 d 维欧氏空间 R^d。

定义 9 超球体 $B^d(r)$ 的表面积定义为

$$S(B^d(r))=S(B^d(1))r^{d-1}$$

等价地，d 维单位超球体可以看作是对 $d-1$ 维球壳的积分，即

$$V(B^d(1))=\int_0^1 S(B^d(1))r^{d-1}dr$$

于是

$$S(B^d(r))=\frac{d}{dr}(V(B^d(r)))=d\cdot V(B^d(1))\cdot r^{d-1}$$

定义 10 超球体的中心切片定义为

$$B_\varepsilon^d(r)=\{x:\|x\|\leqslant r,-\varepsilon\leqslant x^{(1)}\leqslant\varepsilon,\varepsilon\in(0,r)\}$$

3.2.2 重要性质

根据上面的定义，容易导出如下的相关定理。

定理 13 与 d 维超立方体的每个 $d-1$ 维侧面相交的 d 维超球体不一定包含超立方体的中心。如图 2 所示。

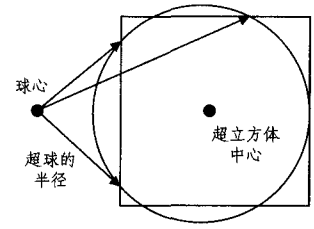


图 2 定理 13 的几何解释

考虑下面的反例。假设中心点在坐标原点的单位超立方体，当 $d=16$ 时，假设超球体的球心在 $(0.2,\cdots,0.2)$ 处，该点到坐标原点的欧氏距离为 $\sqrt{16\cdot0.2^2}=0.8$，此时定义该超球体的半径为 0.7，则该超球体与单位超立方体的所有 15 个侧面相交。显然，超立方体的中心点并不包含在超球体中。

定理 14 $V(B^d(1))=\frac{S(B^d(1))}{d},S(B^{d+1}(1))=2\pi V(B^{d-1}(1)),V(B^d(1))=\frac{2\pi}{d}V(B^{d-2}(1)),S(B^d(1))=\frac{2\pi}{d-2}S(B^{d-2}(1))$。

该定理容易由定义 7 和定义 9 得出，反映了单位超球体的体积与其表面积之间的递归关系。

定理 15 $\lim\limits_{d\to\infty}V(B^d(r))=0,\lim\limits_{d\to\infty}S(B^d(r))=0$。

定理 16[24] $\lim\limits_{d\to\infty}V(B^d(r\sqrt{d}))=\begin{cases}0, & r\leqslant\frac{1}{\sqrt{2\pi e}}\\ \infty, & r>\frac{1}{\sqrt{2\pi e}}\end{cases}$。

上面两个定理表明，任给超立方体 $C^d(r)(r>\frac{1}{\sqrt{2\pi e}})$，随着维数 d 的增加，其外接超球体的体积趋向于无穷大，而内切超球体的体积趋向于 0。

定理 17 $\lim\limits_{d\to\infty}\frac{V(B^d(r-\varepsilon))}{V(B^d(r))}=0$。

该定理表明，高维超球体的体积集中在球壳上。例如，当 $d\geqslant500$ 时，至少 99% 的体积包含在厚度为 1% 的球壳上。

定理 18 $\lim\limits_{d\to\infty}\frac{V(B^d(r-\frac{\varepsilon}{d}))}{V(B^d(r))}=e^{-\frac{\varepsilon}{r}}$。

证明：根据定义 7，可得

$$\lim_{d\to\infty}\frac{V(B^d(r-\frac{\varepsilon}{d}))}{V(B^d(r))}=\lim_{d\to\infty}\frac{2(r-\frac{\varepsilon}{d})^d\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2})}\frac{d\Gamma(\frac{d}{2})}{2r^d\pi^{\frac{d}{2}}}$$

$$=\lim_{d\to\infty}(1-\frac{\varepsilon}{rd})^d=e^{-\frac{\varepsilon}{r}}$$

定理 19 $\lim\limits_{d\to\infty}\dfrac{V(B^d(r))}{V(C^d(r))}=0$。

该定理表明，随着维数的增加，超立方体的体积主要集中于超立方体的边角上，即其内切超球体的体积所占比重越来越小。因此，在高维数据空间的结构分析中，其内部中心往往是“空”的，这种现象被称为“空空间”现象。

定理 20 $\lim\limits_{d\to\infty}\dfrac{V(B_\varepsilon^d(r))}{V(B^d(r))}=1,\lim\limits_{d\to\infty}\dfrac{S(B_\varepsilon^d(r))}{S(B^d(r))}=1$。

该定理表明，高维超球体的体积、表面积主要集中于中心切片上。

3.3 其他高维几何体

下面再介绍几类具有解析形式的体积计算公式的高维几何体。

3.3.1 超长方体

d 维的超长方体 $R^d(a)$ 定义如下：

$$R^d(a)=\{(x^{(1)},\cdots,x^{(d)})\mid -a^{(i)}\leqslant x^{(i)}\leqslant a^{(i)},a^{(i)}\in\mathbb{R}_+\}$$

其体积为：

$$V(R^d)=2^d\prod_{i=1}^{d}a^{(i)}$$

3.3.2 超平行几何体

超平行几何体是平行四边形和平行六面体概念在高维空间中的推广，可定义为：

$$P^d(a)=\left\{\begin{matrix}x:-a\leqslant Ax\leqslant a,a\in\mathbb{R}_+^d\\A\in R^{d\times d},\det(A)\neq 0\end{matrix}\right\}$$

超平行几何体 $P^d(a)$ 可以看作是由超长方体 $R^d(a)$ 经过可逆线性变换得到的，因此

$$V(P^d(a))=|\det(A^{-1})|\cdot V(R^d(a))$$

3.3.3 超单纯形

d 维超单纯形是三角形概念在高维的推广，可以定义为

$$S^d=\{x:0\leqslant x^{(1)}\leqslant x^{(2)}\leqslant\cdots\leqslant x^{(d)}\leqslant 1\}$$

显然，超单纯形 S^d 具有如下形式的 $d+1$ 个顶点：

$\{(0,0,\cdots,0),(0,1,0,\cdots,0),(0,1,1,\cdots,0),\cdots,(1,1,\cdots,1)\}$

下面推导其体积计算公式。注意到在超立方体 $\{x:0\leqslant x^{(i)}\leqslant 1\}$ 内共有 $d!$ 种不同的顶点坐标排列。下面考虑另一个超单纯形，其 $d+1$ 个顶点如下：

$$S_1^d=\{(0,0,\cdots,0),(1,0,0,\cdots,0),(0,1,0,\cdots,0),\cdots,(0,0,\cdots,0,1)\}$$

根据顶点 $x\in S^d$ 和 $y\in S_1^d$ 的关系：

$$x^{(1)}=y^{(d)},x^{(2)}=\sum_{i=1}^{d}y^{(i)},x^{(3)}=\sum_{i=2}^{d}y^{(i)},\cdots,x^{(d)}=\sum_{i=d-1}^{d}y^{(i)}$$

可知此处的线性变换矩阵的行列式为 1，即 $V(S^d)=V(S_1^d)$，因此超单纯形 S^d 的体积为

$$V(S^d)=\frac{1}{d!}$$

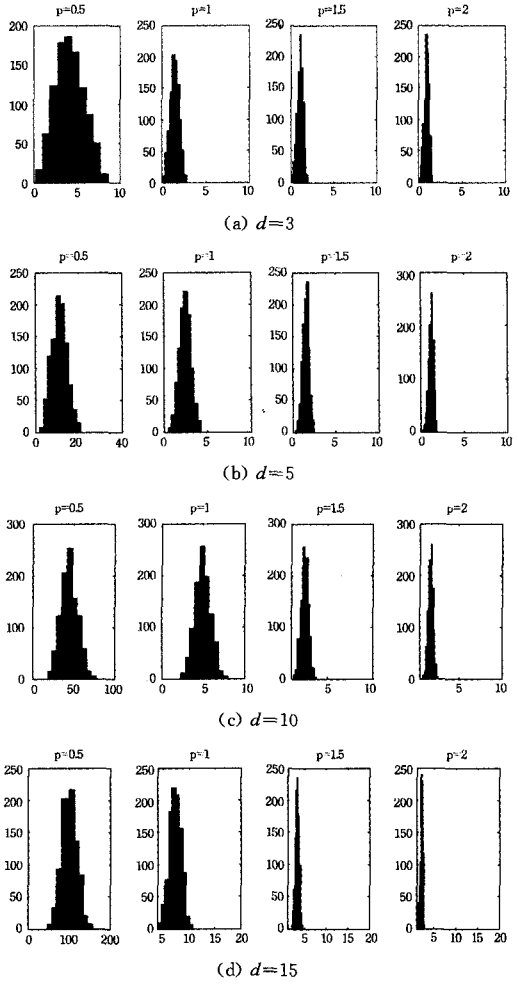
4 实验结果

根据以上对高维数据空间的奇特性质的综述可以看出，“度量集中”现象是高维数据分析的一大难点，传统的基于欧氏距离度量的算法不适合高维数据分析。维数约减是克服这一困难的常规方法，然而维数约减无可避免地会导致数据中蕴含的某些信息的损失，本文主要研究度量选择对机器学习算法性能的影响。下面通过实验来讨论高维数据空间中不同距离度量的分布，及其对经典的基于距离度量的机器学习算法（如 KNN 和 Kmeans）的影响。

4.1 人工数据集上的距离度量选择

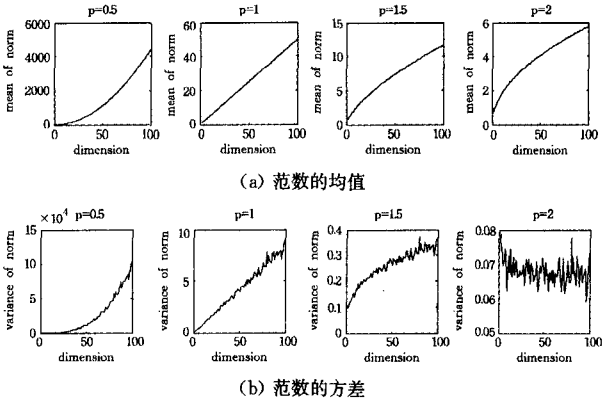
实验中，随机生成 1000 个 d 维样本点，每个维度分量相互独立且服从[0,1]区间上的均匀分布。随着维数 d 的增加，样本点的 ℓ^p 范数（实验中分别取 $p=0.5,1,1.5,2$ 四种情形）的分布呈现出较大的差异，如图 3 所示，维数 d 越高，范数分布的集中效应越明显，且 p 越大，这种集中趋势越突出。另外随着维数 d 的增加，其均值以不同的方式增大，其中 $p=0.5$ 时呈先慢后快增长趋势，$p=1$ 时呈线性增长趋势，$p>1$ 时呈先快后慢增长趋势（见图 4(a)）；当 $p<2$ 时，其方差随着维数增加而增加，$p=2$ 时，方差呈现减小趋势（见图 4(b)）；而相对方差和相对差异随着维数增加而快速减小（见图 4(c)和(d)）。

由图 3 和图 4 所示的实验结果可以看出，当采用分数范数（即 $p<1$ 时）的距离度量方式时，样本间的“度量集中”现象比 $p>1$ 时较弱。

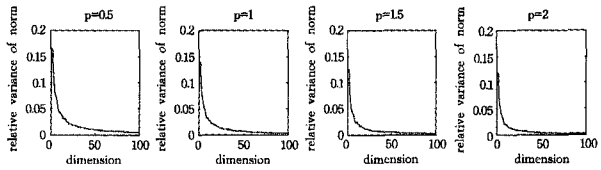


(a) $d=3$　(b) $d=5$　(c) $d=10$　(d) $d=15$

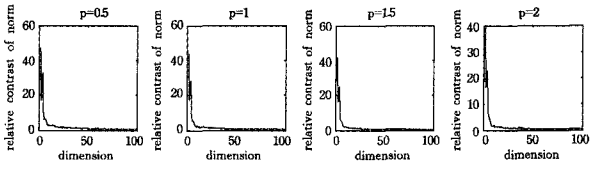
图 3　高维空间中不同维数下样本点范数的分布



(a) 范数的均值　(b) 范数的方差



(c) 范数的相对方差



(d) 范数的相对差异

图 4 高维空间中样本点范数的均值(a)、方差(b)、相对方差(c)和相对差异(d)

4.2 UCI 数据集上的距离度量选择

为了验证不同范数的选择对机器学习算法性能的影响，我们选用来自于真实世界的 UCI 测试数据集[1)]，数据集相关描述如表 1 所列。对于原始数据集中属性有缺失的情形，实验中直接对其赋值−1。同时，实验中首先对每个维度上的数据按照下面的公式规范化为 0 到 1 之间：

$$\frac{x^{(i)}-\min(x^{(i)})}{\max(x^{(i)})-\min(x^{(i)})}$$

表 1 数据集描述

名称	维数	样本数	类别数
Iris	4	150	3
wine	13	178	3
sonar	60	208	2
heart	13	297	2
glass	9	214	4
german	24	1000	2

实验中，我们首先计算了各个数据集上不同范数度量下的相对差异。由表 2 的计算结果可知，$p=0.5$ 时，样本范数的相对差异最大，这也意味着在该范数度量下的机器学习算法的性能也应该是最好的。

表 2 UCI 数据集上的样本 ℓ^p 范数相对差异

名称	p=0.5	p=1	p=1.5	p=2
Iris	**22.4487**	10.2334	9.8902	9.567
wine	**3.7875**	2.4703	2.0935	1.9054
sonar	**2.5133**	1.9964	1.5703	1.2414
heart	**15.9696**	4.5836	2.8656	2.3328
glass	**3.9535**	0.79433	0.66121	0.6735
german	**6.8696**	2.3263	1.3565	0.94006

在分类任务的实验中，我们采用基于不同的范数度量的 KNN 算法，实验中 k 取 3，对样本集进行随机划分，50%作为训练集，50%作为测试集，其评价指标为误分率；在聚类任务的实验中，我们采用基于不同的范数度量的 K-means 算法，初始聚类中心随机选择，最大迭代次数为 1000，其评价指标为 Rand Index。每个数据集上的算法重复执行 50 次后取平均值，算法性能比较结果见表 3 和表 4。

表 3 ℓ^p 范数度量下的 KNN 分类结果

名称	p=0.5	p=1	p=1.5	p=2
Iris	0.0491	0.0421	**0.0379**	0.0387
wine	**0.1328**	0.2589	0.2964	0.3106
sonar	**0.2160**	0.2252	0.2437	0.2533
heart	**0.2588**	0.3254	0.3586	0.3731
glass	**0.2989**	0.3030	0.3161	0.3249
german	**0.2908**	0.3037	0.3188	0.3252

表 4 ℓ^p 范数度量下的 K-means 聚类结果

名称	p=0.5	p=1	p=1.5	p=2
Iris	**0.8679**	0.8519	0.8507	0.8492
wine	0.9313	0.9479	**0.9501**	0.9292
sonar	0.5005	0.5008	0.5020	**0.5023**
heart	**0.6913**	0.6710	0.6457	0.6406
glass	0.6112	0.6183	0.6163	**0.6245**
german	0.5232	**0.5249**	0.5187	0.5166

实验表明，分数范数度量可以显著提高经典的分类算法 KNN 和聚类算法 K-means 在高维数据集上的性能。

4.3 人脸数据集上的距离度量选择

人脸图像数据是典型的高维数据，在人脸识别实验中，首先将图像数据拉直变成向量形式，假设经过裁剪之后的人脸图像长为 32 个像素，宽为 32 个像素，则在算法处理中通常将其转换为 1024(32×32)维的向量。实验中选取了 Yale、Olivetti、UMIST 和 GeorgiaTech 等人脸数据集(见表 5)，分别在不同的距离度量设置下，采用 KNN 算法(k 取 1)进行人脸识别，其中每个人脸随机抽取 T 张图片作为训练集，其余的作为测试集。每次实验重复进行 50 次，平均识别准确率见表 6−表 9，当 $p=0.5$ 时，KNN 算法取得最高的识别准确率。

表 5 人脸数据集描述

名称	维数	样本数	类别数
Yale[2)]	1024	165	15
Olivetti[3)]	2576	400	40
UMIST[4)]	644	575	20
GeorgiaTech[5)]	1800	750	50

表 6 ℓ^p 范数度量下的 KNN 人脸识别结果(Yale)

T	p=0.5	p=1	p=1.5	p=2
4	**0.6623**	0.6392	0.6090	0.5768
5	**0.6780**	0.6564	0.6258	0.6002
6	**0.6917**	0.6712	0.6499	0.6200

表 7 ℓ^p 范数度量下的 KNN 人脸识别结果(Olivetti)

T	p=0.5	p=1	p=1.5	p=2
4	**0.9531**	0.9406	0.9304	0.9176
5	**0.9650**	0.9600	0.9600	0.9550
6	**0.9810**	0.9730	0.9639	0.9528

表 8 ℓ^p 范数度量下的 KNN 人脸识别结果(UMIST)

T	p=0.5	p=1	p=1.5	p=2
5	**0.8968**	0.8737	0.8653	0.8463
6	**0.9055**	0.8857	0.8681	0.8681
7	**0.9126**	0.9034	0.9011	0.8966

1) http://archive.ics.uci.edu/ml/datasets.html

2) cvc.yale.edu/projects/yalefaces/yalefaces.html

3) http://www.cs.nyu.edu/~roweis/data.html

4) http://images.ee.umist.ac.uk/danny/database.html

5) http://www.anefian.com/research/face_reco.htm

表 9 l^p 范数度量下的 KNN 人脸识别结果(GeorgiaTech)

T	p=0.5	p=1	p=1.5	p=2
6	**0.8244**	0.7978	0.7600	0.7356
7	**0.8550**	0.8100	0.7825	0.7600
8	**0.8743**	0.8400	0.8029	0.7686

结束语 本文对高维数据空间的统计性质和几何性质进行了系统的综述,这些性质都可以看作是"度量集中"现象的具体表现。当样本的维数增大时,数据集呈现出"度量集中"现象,即不同样本之间的距离度量的相对差异在逐渐减小,这使得基于样本间距离度量的机器学习算法的性能大大降低。因此,距离度量的选择对于机器学习算法至关重要,本文通过大量实验讨论了不同距离度量的选择对经典的机器学习算法(如 KNN 和 Kmeans)的影响,实验结果表明分数范数的距离度量可以显著提高算法性能。

参 考 文 献

[1] Skillicorn D B. Understanding High-Dimensional Spaces[M]. Springer-Verlag New York Incorporated, 2013

[2] Donoho D L. High-dimensional data analysis: The curses and blessings of dimensionality[J]. AMS Math Challenges Lecture, 2000:1-32

[3] Bellman R. Adaptive Control Process: A Guide Tour[M]. Princeton University Press, Princeton, New Jersey, 1961

[4] Fukunaga K. Introduction to Statistical Pattern Recognition(2nd ed)[M]. New York: Academic, 1990, 39-40(31-34): 220-221

[5] Mil'man V D. New proof of the theorem of A. Dvoretzky on intersections of convex bodies[J]. Functional Analysis and its Applications, 1971, 5(4): 288-295

[6] Weber R, Schek H-J, Blott S. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces[C]//Proceedings of the 24rd International Conference on Very Large Data Bases, ser. VLDB '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998: 194-205

[7] Gaede V, Günther O. Multidimensional access methods [J]. ACM Computing Surveys (CSUR), 1998, 30(2): 170-231

[8] Francois D, Wertz V, Verleysen M. Non-euclidean metrics for similarity search in noisy datasets[C]//Proc. of ESANN. 2005

[9] Kouiroukidis N, Evangelidis G. The Effects of Dimensionality Curse in High Dimensional kNN Search [C] // Informatics (PCI), 2011 15th Panhellenic Conference on. IEEE, 2011: 41-45

[10] Clarke R, Ressom H W, Wang A, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data[J]. Nature Reviews Cancer, 2008, 8(1): 37-49

[11] Jimenez L, Landgrebe D. Supervised Classification in High Dimensional Space: Geometrical, Statistical and Asymptotical Properties of Multivariate data[J]. IEEE Transactions on Geoscience and Remote Sensing, 1999, 37(6)

[12] Jimenez L, Landgrebe D. High dimensional feature reduction via projection pursuit[C]//Geoscience and Remote Sensing Symposium, 1994. IGARSS' 94. Surface and Atmospheric Remote Sensing: Technologies, Data Analysis and Interpretation. International. IEEE, 1994, 2: 1145-1147

[13] Rupp M, Schneider P, Schneider G. Distance phenomena in high-dimensional chemical descriptor spaces: Consequences for similarity-based approaches[J]. Journal of Computational Chemistry, 2009, 30(14): 2285-2296

[14] Francois D, Wertz V, Verleysen M. The concentration of fractional distances[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(7): 873-886

[15] Durrant R J, Kabán A. When is 'nearest neighbor' meaningful: A converse theorem and implications[J]. Journal of Complexity, 2009, 25(4): 385-397

[16] Beyer K, Goldstein J, Ramakrishnan R, et al. When is "nearest neighbor" meaningful? [M]//Database Theory—ICDT' 99. Springer Berlin Heidelberg, 1999: 217-235

[17] Hinneburg A, Aggarwal C C, Keim D A. What is the nearest neighbor in high dimensional spaces? [M]. Bibliothek der Universität Konstanz, 2000

[18] Francois D, Wertz V, Verleysen M. Non-euclidean metrics for similarity search in noisy datasets[C]//Proc. of ESANN. 2005

[19] Hsu C M, Chen M S. On the design and applicability of distance functions in high-dimensional data space[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(4): 523-536

[20] Aggarwal C C, Hinneburg A, Keim D A. On the surprising behavior of distance metrics in high dimensional spaces[C]//Proceedings of the 8th International Conference on Database Theory, ser. ICDT '01. London, UK: Springer-Verlag, 2001: 420-434

[21] Canal L. A normal approximation for the chi-square distribution [J]. Computational Statistics & Data Analysis, 2005, 48(4): 803-808

[22] Katafygiotis L S, Zuev K M. Geometric insight into the challenges of solving high-dimensional reliability problems[J]. Probabilistic Engineering Mechanics, 2008, 23(2): 208-218

[23] Wang Jian-zhong. Geometric Structure of High-Dimensional Data and Dimensionality Reduction[C]// Higher Education Press (China) and Springer. Beijing, 2011

[24] Hopcroft J, Kannan R. Computer Science Theory for the Information Age[M]. Spring, 2012: 7-27