计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
45(4): 612~618, 2008

面向机器学习的相对变换

文贵华
(华南理工大学计算机科学与工程学院 广州 510641)
(crghwen@scut.edu.cn)

Relative Transformation for Machine Learning

Wen Guihua
(*School of Computer Science and Engineering, South China University of Technology, Guangzhou* 510641)

Abstract Recently developed machine learning approaches such as manifold learning and the support vector machine learning work well on the clean data sets even if these data sets are highly folded, twisted, or curved. However, they are much sensitive to noises or outliers contained in the data set, as these noises or outliers easily distort the real topological structure of the underlying data manifold. To solve the problem, the relative transformation on the original data space is proposed by modeling the cognitive relative laws. It is proved that the relative transformation is a kind of nonlinear enlarging transformation so that it makes the transformed data more distinguishable. Meanwhile, the relative transformation can weaken the influence of noise on data and make data relative denser. To measure the similarity and distance between data points in relative space is more consistent with the intuition of people, which can be then applied to improve the machine learning approach. The relative transformation is simple, general and easy to implement. It also has clear physical meaning and does not add any parameter. The theoretical analysis and conducted experiments validate the proposed approach.

Key words machine learning; cognitive laws; relative transformation; noisy data; sparse data

摘 要 机器学习常常面临数据稀疏和数据噪音问题.根据认知的相对性规律提出了相对变换方法,证明了相对变换是非线性的放大变换,可提高数据之间的可区分性.同时在一定条件下相对变换还能抑制噪音,并使稀疏的数据变得相对密集.通过相对变换将数据的原始空间变换到相对空间后,在相对空间中度量数据的相似性或距离更加符合人们的直觉,从而提高机器学习的性能.理论分析和实践验证了所提方法的普适性和有效性.

关键词 机器学习;认知规律;相对变换;噪音数据;稀疏数据

中图法分类号 TP181

机器学习从数据中学习知识,通常要考虑两个问题:数据稀疏和数据噪音的影响.数据的稀疏性对现有机器学习方法有很大的影响,主要原因是它使得数据之间难以用现有的度量来区分,这使得数据分类、聚类等难以得到所期望的结果.数据噪音是普遍存在的,只要是来自实际的真实数据,噪音几乎是不可避免的.从数量上看,噪音必然是少数,但在很多情况下,它们常常导致现有机器学习算法产生剧烈的性能偏差,例如等距嵌入流形学习算法ISOMAP利用测地距离来刻画数据的全局几何结构,嵌入性能非常优越[1],但是它对噪音是拓扑不稳定的[2].有研究试图将噪音点删除[3],但在很多

收稿日期:2007-05-28;**修回日期**:2007-10-24
基金项目:广东省科技攻关基金项目(2007B030803006);教育部留学回国人员科研启动基金项目

情况下不合适,首先是噪音难以准确识别,其次有些看起来是噪音的数据对分析来说是不可少的,如孤立点挖掘. 最后噪音在很多情况并不是独立存在的,而是对正常数据点的污染,比如改变了数据点的坐标值,这样就更不能把被污染的正常数据点按噪音点而删除. 因此较合适的方法是消除噪音对正常点的影响而不是消除"噪音"本身.

考察目前机器学习方法中的基本度量发现,在计算数据点之间的度量时,绝大多数都没有考虑其他数据点的影响,如图1所示,这使得噪音与正常点的待遇相同,同时也没有考虑数据稀疏性以及数据的非均匀分布对度量的影响,这些是造成现有机器学习方法难以处理数据稀疏和数据噪音的重要原因之一. 例如对不均匀分布的样本集,基于近邻的机器学习算法难以选择合适的邻域参数,解决办法除了邻域参数的自适应确定外,还可以利用局部归一化方法重定义距离[4]. 不过本文提出不同的思路,不是发明新的距离度量公式来考虑所有数据点的影响,而是根据认知的相对性规律提出一种相对变换,将原始数据空间转换到相对空间,之后在相对空间中虽然仍然采用原来的距离公式,但计算出的值却考虑了所有数据点的影响. 相对变换能使噪音和孤立点远离正常点,稀疏的数据变得相对密集. 在相对空间中测量数据的相似性或距离能够更符合我们的直觉,从而提高数据分析的准确性.

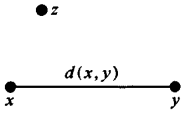

Fig. 1 Computation of the distance $d(x,y)$ without considering the interaction from point z.

图1 计算距离 $d(x,y)$ 时没有考虑数据点 z 的影响

1 相对变换

《Science》上发表的两种流形机器学习方法:局部线性嵌入算法和等距嵌入学习算法可认为是基于认知的方法[1,5]. 因为研究表明人类的感知是流形[6],而流形概念是流形机器学习方法的核心,正因如此这两种方法才具有原始的创新性,引起目前广泛研究. 但是正如前面提到的,它们所面临的困难需要模型化更多的认知规律,因此我们考察认知的相对性规律. 经验表明人类的感知具有相对性,例如在观察图2中的两个圆 x 和 y 时,通常都会认为圆 x 比圆 y 大,而实际上它们一样大[7]. 发生的原因是在观察圆 x 时,与其周围相比,其显得很大,而在观察圆 y 时,与其周围相比,其显得很小,因此这是一种相对性比较的结果.

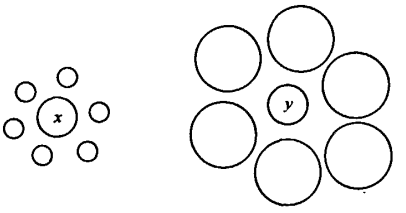

Fig. 2 Human perception on images is relative.

图2 视觉感知的相对性

为模型化这种认知规律,我们以原始数据空间中的每个数据点作为基向量来构造新的空间,这样任意点 x 到所有点的距离就构成该点在新空间中的坐标,这个过程称为相对变换:

$$\Gamma: X \to Y \subset R^{|X|},$$

$$\Gamma_X(x_i) = (d_{i_1}, d_{i_2}, \cdots, d_{i_{|X|}}) = y_i \in Y,$$

其中 $X=\{x_1, x_2, \cdots, x_{|X|}\}$,$|X|$ 为集合 X 的元素个数,y_i 也记为 x_i^r. 通过相对变换构造的空间称为相对空间.

定理1. $\forall_{x_i,x_j \in X} d(x_i,x_j) \leqslant d(x_i^r, x_j^r)$.

证明. 设 $x_i^r = (d(x_i,x_1), d(x_i,x_2), \cdots, d(x_i,x_{|X|}))$,则

$$d(x_i^r,x_j^r)^2 = \sum_{k=1}^{|X|}(d(x_i,x_k) - d(x_j,x_k))^2 =$$

$$\sum_{k=1,k\neq j}^{|X|}(d(x_i,x_k) - d(x_j,x_k))^2 +$$

$$(d(x_i,x_j) - d(x_j,x_j))^2 =$$

$$\sum_{k=1,k\neq j}^{|X|}(d(x_i,x_k) - d(x_j,x_k))^2 +$$

$$d(x_i,x_j)^2 \geqslant d(x_i,x_j)^2. \qquad \text{证毕.}$$

因此相对变换不是等距变换,而具有放大作用,这有利于我们观察数据之间的拓扑结构的细节.

定理2. $\exists_{X \wedge x_i,x_j,x_k \in X}(d(x_i,x_j) = d(x_i,x_k) \wedge d(x_i^r,x_j^r) \neq d(x_i^r,x_k^r))$.

我们举一个实例来说明. 从图3中可以看出,在原始数据空间中 $d(x_3,x_1) = d(x_3,x_4)$,此时,x_3 无法决定 x_1 和 x_4 谁离自己更近,这对基于最近邻选择的机器学习方法产生不利影响. 但是在转换后的相对空间中,$d(y_3,y_1) < d(y_3,y_4)$,很容易决定 y_1 与 y_3 更近,特别是这种情形也更符合人类的直觉,因此相对变换不是线性变换,它能够将原来在

原始数据空间中不能区分的数据在相对空间中区分开来,从而提高了数据之间的可区分性.

同时相对变换对抑制噪音或识别孤立点都非常有用,从而可提高机器学习的鲁棒性. 例如图 3 中的 x_4 可能是孤立点,但是在原始数据空间中 $d(x_3, x_1) = d(x_3, x_4)$,这使 x_1 和 x_4 有相同的机会成为点 x_3 的近邻,这与我们的直觉不一致. 而在相对空间中,$d(y_3, y_1) < d(y_3, y_4)$,这意味着孤立点 y_4 更加远离正常数据点.

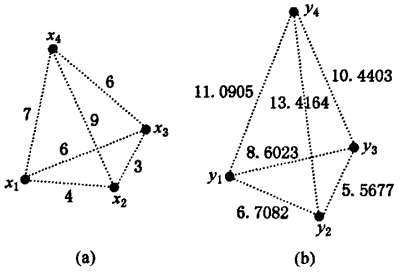

Fig. 3 Relative transformation can weaken the influence of noise. (a) The original space and (b) The constructed relative space.

图 3 相对变换能抑制噪音影响. (a)原始空间;(b)构造的相对空间

2 相对空间的构造

相对变换中的距离可以采用常用的距离如 Euclidean 距离、Chebychev 距离、Manhattan 距离、Minkowsky 距离、加权的 Minkowsky 距离、Mahalanobis 距离等,也包括为图像等特殊数据定义的特殊距离. 它们分别定义如下:

1) $d_e(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$;

2) $d_c(x,y) = \max_{i=1}^{n} |x_i - y_i|$;

3) $d_{\text{man}}(x,y) = \sum_{i=1}^{n} |x_i - y_i|$;

4) $d_{\min}(x,y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$;

5) $d_{w\min}(x,y) = \sqrt[p]{\sum_{i=1}^{n} w_i |x_i - y_i|^p}$;

6) $d_{\text{mah}}(x,y) \sqrt[n]{|\Delta C|}(x-y)^{\mathrm{T}} C^{-1}(x-y)$,

其中 C 为协方差矩阵,若 C 为单位阵,那么 Mahalanobis 就退化为平方欧氏距离.

每种距离度量都有各自适用的问题,我们利用它们构造的相对空间也是如此,但能将问题解决得更好. 为此我们提出一个一般性的评估准则. 对任意给定的一个数据集,我们为每种距离度量建立距离矩阵 D,然后考察该矩阵中相同元素个数的平均值 r 来评估该距离度量就该数据集的合理性:

$$r = \sum_{d_{ij} \in D, num(d_{ij}) > 2} num(d_{ij})/2m,$$

其中 $num(d_{ij})$ 是 d_{ij} 在距离矩阵 D 中出现的个数,m 为满足条件 $num(d_{ij}) > 2$ 的个数. 在近邻查询中,距离矩阵中相同的元素越少,该距离度量的区分能力就越强.

因此若 r 随着样本数的增加而迅速增加,此时样本之间的区分就越困难,从而说明该距离度量的区分能力迅速降低. 我们比较几种距离度量在原始空间和相对空间中的 r 值变化,来考察相对变换的合理性. 我们选择 2500 个中文专利摘要[8],向量化后生成 2500 个文本向量,维数为 6437,这是典型的高维数据. 然后从中分别选择 100,200,…,800 形成 8 个样本数不断增加的数据集. 最后分别在原始空间和相对空间中计算这些数据集的 r 值,结果如图 4 所示. 可以看出:1)不同距离度量在原始空间中的区分能力是不同的,Euclidean 距离最好,这与现有研究是一致的[9],说明了我们定义的评估准则是合理的. 2)每种距离度量对应的 r 都随着样本数的增加而增加,区分能力不断减弱. 3)在实验中发现,每种距离度量在相对空间中的区分能力都显著增强,形成的距离矩阵中没有任何相同的元素,能够全部区别开来,说明了相对变换的重要性.

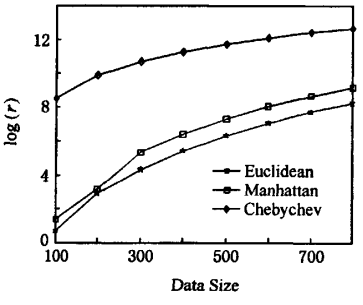

Fig. 4 Comparisons among distinguishing abilities of several distances.

图 4 几种距离度量的区分能力比较

3 实验分析

利用相对变换增强机器学习有两种基本途径,

第1种是组合运用原始空间和相对空间协同完成机器学习任务,我们选择流形学习为应用范例.第2种是将数据转换到相对空间后,在相对空间中完成全部机器学习分析,我们选择支持向量机分类器为应用范例.考虑到欧氏距离是目前公认的具有较好效果的度量,这与前面的讨论是一致的,我们以它为例,构造相对空间并应用于机器学习.

3.1 数据降维

流形学习是实现非线性数据降维的一类新机器学习方法,我们对其有一些前期研究[10-12]因此选它为例.具体选择基于Hessian的局部线性嵌入算法HLLE[13],说明相对变换能够显著提高流形学习算法的性能.方法是在相对空间中确定数据的邻域结构,而嵌入仍然在原始空间中完成,为便于区别,将此方法记为R-HLLE.实验中的参数采用HLLE在其实验中所用的参数.实验数据采用Swiss roll surface,其是广泛采用的标准数据集[1-6],类似于一个长方形纸片卷起的三维图形,机器学习的目的就是从卷起的三维图形还原成二维长方形.我们采用HLLE的采样方法,从Swiss roll surface随机采样多个800个点的长方形但同时从其中心移去一个小的长方形以使得数据集不再是凸的.这是一个很有挑战性的数据集,很多算法都不能得到理想的结果.

实验1. 噪音数据.

我们从Swiss roll surface上随机采样800个点,然后叠加均值为0和方差为0.4的高斯噪音.按此方法采样多次并测试几种方法.分析发现HLLE在少部分情况下能够将数据嵌入在二维空间.ISOMAP总是将去除的区域强烈膨胀,并扭曲其余的数据点.LLE在绝大多数情况下都得不到正确结果.R-HLLE也受噪音的影响,在部分情况下也不能正确嵌入,但相对稳定,在较多情况下都能够较完美地将数据嵌入在二维空间,其中心移去的一个小长方形也能在嵌入的二维空间中正确体现.图5是其中的一个结果,可以看出R-HLLE表现最好.

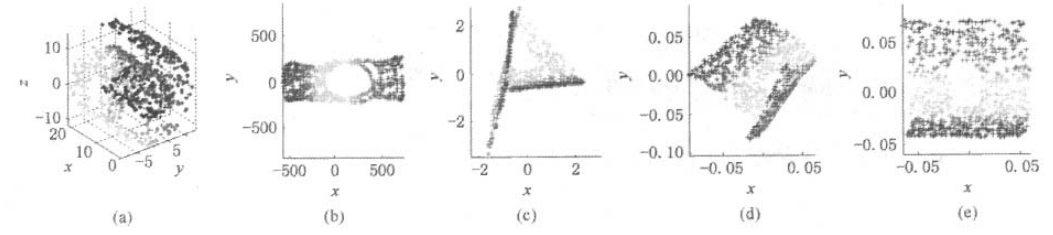

Fig. 5 Embedding results on noisy data set. (a) Original data; (b) ISOMAP; (c) Regular LLE; (d) HLLE; and (e) R-HLLE.

图5 噪音数据集上的嵌入结果. (a) Original data; (b) ISOMAP; (c) Regular LLE; (d) HLLE; and (e) R-HLLE

实验2. 稀疏数据.

我们从Swiss roll surface上随机采样数据规模为400点的多个稀疏数据集,然后测试几种方法.分析发现在很多情况下HLLE和LLE是混乱的.R-HLLE在部分情况下也不能正确嵌入,但相对而言,R-HLLE表现最好,在较多情况下都能够较完美地将数据嵌入在二维空间,其中心移去的一个小长方形也能在嵌入的二维空间中正确体现,图6是其中的一个结果,这证实了相对变换能使原始数据空间中的稀疏数据变得相对密集.

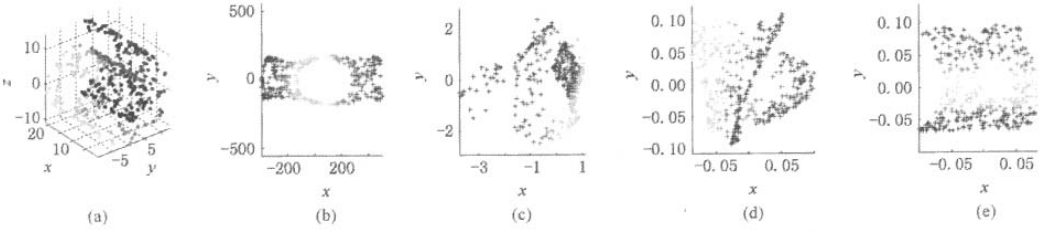

Fig. 6 Embedding results on sparse data set. (a) Original data; (b) ISOMAP; (c) Regular LLE; (d) HLLE; and (e) R-HLLE.

图6 稀疏数据集上的嵌入结果. (a) Original data; (b) ISOMAP; (c) Regular LLE; (d) HLLE; and (e) R-HLLE

3.2 数据分类

对非线性数据实现快速分类,Vapnik提出的支持向量机SVM(support vector machine)是一个很好的选择,理论基础好,并已被广泛应用到语音处理、

图像检索文本分类等领域[14]. 因此我们选择 SVM 分类器,比较其在数据的原始空间和相对空间中的分类效果来检验相对变换的有效性. 为方便,将在相对空间中完成分类的 SVM 记为 R-SVM.

实验 3. 噪音数据.

实验数据采用螺旋状分布的两类数据集[14],如图 7 所示:

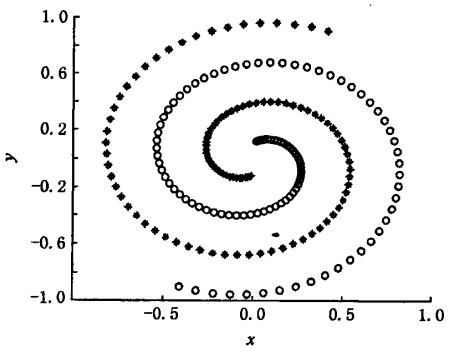

Fig. 7 Two spiral classes data set.

图 7 螺旋分布的两类数据集

我们抽取 200 个记录的数据集. 实验中,核函数采用多项式,其中惩罚系数 $C=1$,核函数参数 σ 范围为 1～20,测试采用"leave one out"策略,我们分别运行 SVM 和 R-SVM 寻找最大分类正确率的核函数参数,结果如图 8 所示:

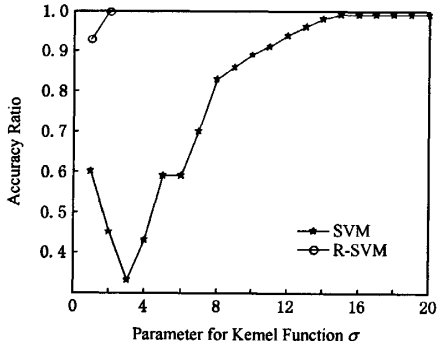

Fig. 8 Classification accuracy against kernel parameter.

图 8 分类准确率随核函数参数的变化趋势

R-SVM 测试参数到 2 时已经取得 100% 的正确率,因此后面的参数不再测试. 而 SVM 需要测试完所有的参数,到 $\sigma=15$ 时取得最大值,此后的参数值不再增加分类正确率,因此在后面的实验中,R-SVM 设置 $\sigma=2$,而 SVM 设置 $\sigma=15$. 从分类结果来看,R-SVM 的分类正确率比 SVM 高,但不明显,只高 1%,这主要是生成的数据是无噪音的人工合成数据,而相对变换主要在于能够更好地处理噪音数据,因此我们对此实验数据依次添加均值为 0,方差分别为 0.02 0.04 0.06 0.08 0.10 的随机高斯噪音形成 5 个噪音数据集,然后分别运行 SVM 和 R-SVM,分类结果如表 1 所示:

Table 1 Classification Accuracy on Noisy Data Sets

表 1 数据噪音对分类准确率的影响 %

Noise Variance	SVM	R-SVM
0.02	95	100
0.04	92.5	99
0.06	89	96
0.08	86	91
0.1	77	82.5

不难看出 SVM 受噪音的影响非常明显,即使是少量的噪音,分类的准确率都明显降低,而 R-SVM 则不同,在噪音方差为 0.02 时不受任何影响,准确率仍然为 100%,噪音方差为 0.04 时影响不明显. 但是随着噪音的进一步增大,R-SVM 的分类正确率都明显降低,说明它抗噪音能力仍然需要提高. 但在任何情况下,R-SVM 在噪音数据集上的分类正确率都明显比 SVM 高很多,大于 5%.

实验 4. 稀疏数据.

支持向量机能够处理小样本问题,但分类的准确率仍受影响. 我们采用分析 SVM 的常用数据集 ringnorm[15]做实验,该数据集是 Leo Breiman 生成的用于两类划分的样本集,每一类都是取自一个 20 维的多变量正态分布 . 类 1 的均值为 0,方差为单位元的 4 倍. 类 2 的均值为$(a,a,\cdots,a)$,方差为单位元,其中 $a=2/20$. 实验中,核函数采用多项式,其中惩罚系数 $C=1$,核函数参数 σ 范围为 1～20. 首先在样本规模为 100 的样本上分别运行 SVM 和 R-SVM 寻找取得最大分类正确率的核函数参数,发现 $\sigma=2$ 对两个分类器都是最佳参数,后继实验均采用此值. 然后我们分别在样本规模为 100,200,…,1000 的数据集上运行 SVM 和 R-SVM,测试采用"leave one out"策略,分类结果如表 2 所示,不难看出 SVM 受数据稀疏的影响非常明显,在样本规模小于 300 时,SVM 的分类准确率比 R-SVM 的差距非常明显,差距大于 10%,而 R-SVM 则在小样本上表现最好. 随着样本规模的增大,SVM 的分类准确率不断提高,并停留在较高的水平,但有波动. 而 R-SVM 则一直稳定在高水平,它在任何规模数据集上的分类准确率都比 SVM 高.

Table 2 Classification Accuracy Against Different Data Sizes

表 2 数据稀疏对分类准确率的影响 %

Data Size	SVM	R-SVM
100	87	99
200	87.5	98
300	92.67	96.33
400	92.75	97.75
500	93	96.8
600	95.67	97.33
700	92.57	97.43
800	97.25	98.75
900	97.22	97.78
1000	95.40	97.50

4 结 论

解决数据稀疏和噪音是机器学习的共性问题,我们根据认知的相对性规律,提出了相对变换方法,其特征在于:1) 提出了一种新的思路,尝试通过将一些认知规律的几何化来发现一些新的计算原理和方法. 2)相对变换能抑制噪音,并将稀疏数据变得相对密集. 3)相对变换具有简单性、普适性和可操作性特点. 4)相对变换与核函数不同. 核函数属于黑箱方法,缺乏可解释性,升维后是否能够抑制噪音和数据稀疏不清楚. 而相对变换属于白箱方法,空间维的物理意义清晰,可解释性好. 在将相对变换用于SVM(其已经用到核函数)的实验中证实了相对变换能够增强SVM处理噪音和数据稀疏的能力,说明了相对变换比核函数更加优越,至少具有互补性. 目前相对变换在机器学习中的初步应用是成功的,后继工作是对其拓展研究,包括理论研究和扩展应用. 例如对大规模数据,相对变换将构造很高维的空间,不仅计算复杂,也可能面临新的维数灾难,探索局部相对变换就是我们正在展开的一项很重要的工作.

参 考 文 献

[1] J B Tenenbaum, V de Silva, J C Langford. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290(5500): 2319-2323

[2] M Balasubramanian, E L Schwartz. The ISOMAP algorithm and topological stability [J]. Science, 2002, 295(5552): 7

[3] Heeyoul Choi, Seungjin Choi. Robust kernel ISOMAP [J]. Pattern Recognition, 2007, 40(3): 853-862

[4] Wang Heyong, Zheng Jie, Yao Zhengan, *et al*. Application of dimension reduction on using improved LLE based on clustering [J]. Journal of Computer Research and Development, 2006, 43(8): 1485-1490 (in Chinese)
(王和勇, 郑杰, 姚正安, 等. 基于聚类和改进距离的LLE方法在数据降维中的应用[J]. 计算机研究与发展, 2006, 43(8): 1485-1490)

[5] S T Roweis, L K Saul. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323-2326

[6] H S Sung, D D Lee. The manifold ways of perception [J]. Science, 2000, 290(5500): 2268-2269

[7] Li Deyi, Liu Changyu, Du Yi, *et al*. Artificial intelligence with uncertainty [J]. Journal of Software, 2004, 15(11): 1583-1594 (in Chinese)
(李德毅, 刘常昱, 杜鹢, 等. 不确定性人工智能[J]. 软件学报, 2004, 15(11): 1583-1594)

[8] Wen Guihua, Jiang Lijun, Wen Jun, *et al*. Generating creative ideas through patents [G]. In: LNAI 4099. Berlin: Springer, 2006. 681-690

[9] He Ling, Wu Lingda, Cai Yichao. Similarity measurement of data in high-dimensional spaces [J]. Mathematics in Practice and Theory, 2006, 36(9): 189-194 (in Chinese)
(贺玲, 吴玲达, 蔡益朝. 高维空间中数据的相似性度量[J]. 数学的实践与认识, 2006, 36(9): 189-194)

[10] Wen Guihua, Jiang Lijun, Wen Jun, *et al*. Performing locally linear embedding with adaptive neighborhood size on manifold [G]. In: LNAI 4099. Berlin: Springer, 2006. 985-989

[11] Wen Guihua, Jiang Lijun, Wen Jun, *et al*, Clustering-based nonlinear dimensionality reduction on manifold [G]. In: LNAI 4099. Berlin: Springer, 2006. 444-453

[12] Wen Guihua, Jiang Lijun, Nigel R Shadbolt. Using graph algebra to optimize neighborhood for isometric mapping [C]. The 20th Int'l Joint Conf on Artificial Intelligence (IJCAI-07), Hyderabad, India, 2007

[13] D L Donoho, C Grimes. Hessian eigenmaps: Locally linear embedding, techniques for high-dimensional data [C]. Proceeding of the National Academy of Sciences of the United States of America, 2003, 100(10): 5591-5596

[14] Nikolaos Nasios, Adrian G Bors. Kernel-based classification using quantum mechanics [J]. Pattern Recognition, 2007, 40(3): 875-889

[15] Li Honglian, Wang Chunhua, Yuan Zongbao, *et al*. A learning strategy of SVM used to large training set [J]. Chinese Journal of Computers, 2004, 27(5): 713-719 (in Chinese)
(李红莲, 王春花, 袁宗宝, 等. 针对大规模训练集的支持向量机的学习策略[J]. 计算机学报, 2004, 27(5): 713-719)

Wen Guihua, born in 1968. Received his Ph. D. degree in artificial intelligence from South China University of Technology, Guangzhou, China. He has been associate professor of South China University of Technology since 2001. His main research

interests are computational creativity, data mining and knowledge discovery, machine learning, and cognitive geometry.

文贵华,1968年生,博士,副研究员,主要研究方向为创新计算、数据挖掘与知识发现、机器学习、认知几何.

Research Background

Recently developed manifold learning approaches can nicely deal with nonlinear manifolds. These approaches are intuitive, well understood, good theoretical studies, and produces good embeddings on the clean data sets even if these data sets are highly folded, twisted, or curved. The support vector machine is another competitive machine learning approach for classification. This approach has good generalization performance and only small training samples are required. However, these approaches are still much sensitive to noises or outliers contained in the data set, as these noises or outliers easily distort the constructed neighborhood graph that should faithfully represent the underlying data manifold and distort the hyperplane for support machine learning. Some approaches can attack the problem by deleting the noises or outliers from the input data set. However sometimes the outliers are useful to data analysis such as outlier mining so that they cannot be removed simply. Another problem is that these approaches are also unsuitable for dealing with much sparse data sets.

Inspired by the cognitive relativity, we present a relative transformation to build the relative space from the original space of data. The relative transformation is simple but efficient to deal with the outliers and noises. Furthermore, in the relative space the distances among points vary nonlinearly. Possibly it makes closer the points belonging to the same surface of the manifold while it makes further away the points located at the different surfaces. This is useful to the sparse data sets. The relative transformation is simple, general and easy to implement. It also has clear physical meaning and does not add any parameter. The experimental results validate the proposed approach.

面向机器学习的相对变换

作者： 文贵华， Wen Guihua
作者单位： 华南理工大学计算机科学与工程学院,广州,510641
刊名： 计算机研究与发展 ISTIC EI PKU
英文刊名： JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT
年，卷(期)： 2008,45(4)
被引用次数： 2次

参考文献(15条)

1. Heeyoul Choi;Seungjin Choi Robust kernel ISOMAP[外文期刊] 2007(03)

2. M Balasubramanian;E L Schwartz The ISOMAP algorithm and topological stability 2002(5552)

3. J B Tenenbaum;V de Silva;J C Langford A global geometric framework for nonlinear dimensionality reduction[外文期刊] 2000(5500)

4. 贺玲;吴玲达;蔡益朝 高维空间中数据的相似性度量[期刊论文]-数学的实践与认识 2006(09)

5. Wen Guihua;Jiang Lijun;Wen Jun Generating creative ideas through patents 2006

6. 李德毅;刘常昱;杜NFDB4 不确定性人工智能[期刊论文]-软件学报 2004(11)

7. H S Sung;D D Lee The manifold ways of perception[外文期刊] 2000(5500)

8. 李红莲;王春花;袁宗宝 针对大规模训练集的支持向量机的学习策略[期刊论文]-计算机学报 2004(05)

9. Nikolaos Nasios;Adrian G Bors Kernel-based classification using quantum mechanics[外文期刊] 2007(03)

10. D L Donoho;C Grimes Hessian eigenmaps:Locally linear embedding,techniques for high-dimensional data 2003(10)

11. Wen Guihua;Jiang Lijun;Nigel R Shadbolt Using graph algebra to optimize neighborhood for isometric mapping 2007

12. Wen Guihua;Jiang Lijun;Wen Jun Clustering-based nonlinear dimensionality reduction on manifold 2006

13. Wen Guihua;Jiang Lijun;Wen Jun Performing locally linear embedding with adaptive neighborhood size on manifold 2006

14. S T Roweis;L K Saul Nonlinear dimensionality reduction by locally linear embedding[外文期刊] 2000(5500)

15. 王和勇;郑杰;姚正安 基于聚类和改进距离的LLE方法在数据降维中的应用[期刊论文]-计算机研究与发展 2006(08)

引证文献(2条)

1. 文贵华.陆庭辉.江丽君.文军 基于相对流形的局部线性嵌入[期刊论文]-软件学报 2009(9)

2. 文贵华.陆庭辉.江丽君.文军 基于相对流形的局部线性嵌入[期刊论文]-软件学报 2009(9)