一种缓解分类面交错的样本点扩散方法

梁 路 龚奔龙 黎 剑 滕少华
(广东工业大学计算机学院 广州 510006)

摘 要 固定的相似性度量使得学习器无法结合先验信息揭示数据本身固有的统计规律,对于分类面交错严重的数据集,难以取得较好的学习效果。为了缓解分类面交错,提高分类准确度,将边界和样本点扩散结合起来,通过统计样本标签信息和位置信息得到边界点,以边界点为中心选取合适的控制函数对周边样本点进行扩散,使得分类面更加清晰,从而提高分类算法的精度。在多个分类面交错的数据集上,使用不同分类器验证所提方法,结果表明,其准确率有不同程度的提升。与3种经典的有监督度量学习方法进行比较,实验结果表明所提方法适合处理交错程度高的数据集,而且能有效提升SVM的性能。

关键词 度量学习,样本点扩散,数据预处理
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.09.053

Diffusion Method of Sample Points for Alleviating Staggered Situation of Classification

LIANG Lu GONG Ben-long LI Jian TENG Shao-hua
(School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China)

Abstract The fixed similarity measurement makes learner difficult to reveal the inherent statistical rules of the data itself with the priori information, and it is difficult to get good effect for the data set with a staggered classification. In order to improve the classification accuracy of the data set with a staggered classification, this paper combined the boundary and sample diffusion method. The method applies the statistical sample label information and position information to obtain boundary point, which is treated as the center. Then we selected appropriate control function to spread neighboring sample points to make the classification more clear, so as to enhance the learning accuracy. Different classifiers are used to validate the method, and the accuracy of the proposed method is improved in different degrees. Compared with three classical supervised distance metric learning method, the experimental results show that this method is suitable for processing high degree of interleaving data sets, and can effectively improve the performance of SVM.

Keywords Distance metric learning, Sample point dispersion, Data preprocessing

1 引言

在相似度距离计算方面,一些固定的距离度量常被用于计算,例如余弦距离、欧氏距离、曼哈顿距离等。距离函数度量了不同样本点之间的相似性,显著地影响着大部分机器学习算法的性能。然而,固定的距离度量无法有效地反映样本的空间关系,在很多场景下不适用,只能通过不断地增强分类器来进行弥补[1]。在模式识别的分类问题中,分类器的设计已经取得了非常丰富的研究成果。利用原始样本学习出一个更适于分类的输入样本空间,从而提高分类器的性能,是新的着力点,具有越来越重要的研究价值。

一些研究工作[2-4]表明,当先验信息缺失时,通过训练数据学习出的距离度量能够比固定的距离度量带来更大的性能提升。最早,Xing等[5]提出了一种经典的距离相似度量算法,其基本思想是:在最小化相似对之间的马氏距离平方和的同时,约束非相似对之间的马氏距离和,即令其大于某个阈值。用这种方式建立目标优化函数,使得在新的度量空间中同类样本的分布更加紧凑,不同类样本的分布更加松散。距离度量学习算法可分为无监督的距离度量学习算法和有监督的距离度量学习算法。无监督距离度量学习在一定程度上提高了样本的区分性,但实际中有些标签相同的样本在特征空间中却离得很远,一些标签不同的样本在特征空间却很近[6]。为了解决这个问题,需要使用标签信息进行训练才能得到合适的距离度量。相比于非监督的距离度量学习,利用携带标签信息的训练数据进行距离度量学习可以更好地提升分类性能。Weinberger等[7]利用类别信息作为先验知识,提出了最大边际的最近邻算法(Large Margin Nearest Neighbor, LMNN),该算法利用最大化不同类别的边际距离的思想来学

到稿日期:2016-08-18 返修日期:2016-12-12 本文受国家863计划重大项目(2013AA01A212),国家自然科学基金资助项目(61272067,61104156,61402118),广东省自然科学基金(9451009001002777)资助。

梁 路(1980-),女,博士,副教授,CCF会员,主要研究方向为数据挖掘、协同计算,E-mail:lianglu@gdut.edu.cn(通信作者);龚奔龙(1991-),男,硕士生,主要研究方向为数据挖掘;黎 剑(1989-),男,硕士生,主要研究方向为数据挖掘、机器学习;滕少华(1962-),男,博士,教授,主要研究方向为大数据、数据挖掘与协同计算、网络安全。

习度量矩阵。此外,经典的有监督度量学习方法还包括近邻成分分析(Neighborhood Components Analysis,NCA)、线性判别式分析(Linear Discriminant Analysis,LDA)[8] 等。此外,王裴岩等[9]提出了一种基于核距离的核度量方法,何进荣等[10]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。

然而,度量学习算法在整个数据集中构造全部相似对和非相似对,当数据量很大时,构造出的相似对集合和非相似对集合也会过大,故其难以处理大规模的数据集问题。在高维数据中由于矩阵的元素个数是数据维度的平方,直接求解的消耗很大,且需要大量的样本数据进行训练。很多应用场景无法收集足够的数据进行训练,导致训练不充分。此外,如果要处理非线性变化,只能定义一些特定的非线性度量,如文献[11],再结合凸优化进行处理。这种方法要求已知数据非线性分布的先验知识,才能定义一个合理的非线性度量,但该条件在实际情况中很难满足。如果采用核方法,同时优化矩阵和核方法参数将使得系统异常复杂且不稳定[12-13]。因此,对更加简单直观、场景适应性更强的度量学习方法的研究依然是很有价值的。文献[14]提出了一种非均匀分布数据的非线性标准化方法,该方法是无监督的,致力于解决数据局部挤压现象,逐个维度使用一种基于数据拟合的非线性变换,在拉伸数据稠密区间的同时压缩数据稀疏区间;而本文对样本的处理是全空间的,致力于扩大类间隔,是有监督的。因此本文方法与文献[14]中的方法可以作为样本集处理的系列方法。

本文借鉴度量学习的思想,使同类样本趋于紧凑,异类样本趋于松散,从边界的角度直观地提出了一种改进数据集分布的样本点扩散方法,并且无需构造全部相似对和非相似对。该方法利用样本的标签信息,统计并计算出各个样本点被异类样本点包裹的程度,即边界值,我们认为边界值较大的样本点的周边区域有较大的类别混淆程度,以这些点为中心点,对其他样本点在全空间进行扩散,扩散的长度与样本点的位置和标签有关。经过扩散处理后的样本集的类间距离增大,类间间隔更加清晰。将本文提出的方法与具有代表性的有监督度量学习算法(LDA,LMNN 和 NCA)进行比较,实验结果表明本文方法能够有效提高数据集的识别准确度,与 3 种经典方法相比亦有不同程度的提升效果。

边界值极好地融合了样本集的类别信息和位置信息,其取值大小为样本点扩散的中心点位置和扩散尺度提供了重要参考。本文第 2 节介绍如何求解边界值;第 3 节介绍得到边界之后样本扩散的方法;第 4 节给出实验结果并进行分析;最后总结全文。

2　边界值的计算

万韩永等[15]提出了一种基于样本重要性的改进的 KNN 方法,利用样本的标签信息和空间信息学习出各个样本点的边界值,即邻域内被异类样本点包裹的程度,进而算出每个样本点的重要性,即样本点对其所属类别的隶属程度,最后利用样本点的重要性对 KNN 算法的投票策略进行优化。本文借鉴了其求解样本点边界值的方法,下面进行简单介绍。

对于具有类标签的样本集,首先在不同类别的样本点之间建立一个加权无向图,权重值与样本之间的距离成反比关系;然后在建立的无向图上生成马尔科夫转移矩阵;最后在整个无向图上执行随机游走算法,计算出每个样本点的边界值。边权重的计算方法为:

$$r(j,i)=\exp\{-\lambda\, dist(j,i)\}$$

其中,$dist(j,i)$为样本点 x_i 和 x_j 之间的欧氏距离,λ 为参数。由于当样本维度增大时,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逐渐减小,这会使得基于样本间距离度量的学习算法的性能大大降低。因此本文实验对其进行改进,使用分数范数[10]来度量样本点间的差异:$\|x\|_r=(\sum_{i=1}^{d}(x^i)^r)^{\frac{1}{r}}$。其中,$d$ 为样本维度,本文实验中 r 取 0.5。然后生成马尔科夫转移矩阵:

$$P(j,i)=\begin{cases}0, & \sum\limits_{k\in N_j} r(j,k)I(j,k)=0\\ \dfrac{r(j,i)I(j,i)}{\sum\limits_{k\in N_j} r(j,k)I(j,k)}, & \text{其他}\end{cases}$$

其中,$I(j,i)=\begin{cases}1, & y_i\neq y_j \text{ 且 } (x_i)\in N_j\\ 0, & \text{其他}\end{cases}$,$y$ 为类标签,N_j 为样本 x_j 的 k 个最近邻,在本文实验中求解最近邻时亦使用 $r=0.5$的分数范数。对于一个样本,它的边界值由最近邻中不同类标签样本的边界值决定。$P(j,i)$可以看作是样本 x_j 到样本 x_i 的转移概率。由此生成马尔科夫链的平稳分布:

$$\pi(j)=\sum_{i\in X}\pi(i)P(i,j)$$

其中,$\pi(j)$即为样本 x_j 的边界值,根据随机游走理论,该算法最终收敛。算出所有样本点的边界值组成边界值向量 B,$B[i]$表示样本 x_i 的边界值。

3　扩散方法

边界值大的样本点的周边区域是样本集中相对稠密的区域,而且该区域内样本点的类别有较大的不纯度,这将导致分类面模糊交错,从而极大地影响各类机器学习算法的执行效果。对这些区域的样本点的位置进行合理变换,使得在不改变样本集整体分布规律的情况下能够改善分类面交错区域的样本点分布,扩大数据集交错区域的样本点间距,是本文所提方法的基本思想。下面介绍具体的样本扩散方法。

定义 1(扩散的中心点)　按元素值从大到小对边界值向量 B 进行降序排列,前 n 个即为扩散的中心点,记为(C_1,C_2,…,C_n),n 为参数。

定义 2　使用分数范数来计算样本 x_i 到中心点 C_j 的距离 $d(x_i,C_j)$。

定义 3　中心点 C_j 对样本点 x_i 的推送距离 $P(x_i,C_j)$的计算公式为:

$$P(x_i,C_j)=\varepsilon\times B(j)\times f(d(x_i,C_j))$$

其中,ε 为参数,$f(x)$为推力控制函数,将在定义 4 中解释。本文将数据集中的对象看成空间中的质点,并且将牛顿力学的合力思想引入到样本点扩散方法中。推送距离 $P(x_i,C_j)$类似于质点间的斥力,$B(j)$所表征的边界值对推送距离有正向影响,其意义在于:样本点的边界值越大,其周边区域的异类样本点越密集,则这一区域的样本点需要挪动的距离越大,边界值类似于求解万有引力时的质点质量,其值越大,产生的力越大。与扩散距离相关的另一个重要因素是样本点与扩散

中心点的距离，为了打散交错的分类面，距离越近，则需推送得越远，因此，$P(x_i,C_j)$应与间距 $d(x_i,C_j)$负相关，$f(d(x_i,C_j))$即为表现这种负相关关系的函数。

定义 4 推力控制函数为：$f(x)=xe^{-\frac{x^2}{2}}$

在选择控制函数时，类比于万有引力定律公式 $F=\frac{GMm}{r^2}$ 的形式，选择与距离的平方负相关的形式，即：

$$f_1(x)=\frac{1}{x^2}$$

经过推理发现，当样本点与扩散中心点的间距非常小即 x 趋近于 0 时，推送距离 $P(x_i,C_j)$将发生剧烈跳变，这时会产生严重的后果，样本点经过推送后的相对位置发生改变，从而导致误分类，说明选择这个函数是不恰当的。

为了解决 0 点附近的跳变现象，同时保证距离越近推送越远以及距离越远推送距离趋向于 0，考虑类似于高斯分布函数的拱形曲线形式，引入因子 $\exp(\frac{x^2}{2})$，因子 x 也是为了在实验中缓解跳变现象并提升分类准确率而引入的。

定义 5 样本 x_i 受到的各个扩散中心点的合力矢量为：

$$\varphi=\sum_{j=1}^{n}P(x_i,C_j)\times\frac{\overrightarrow{C_jx_i}}{|\overrightarrow{C_jx_i}|}$$

其意义为每个扩散中心点都对非中心点的样本点施加一定的影响，样本挪动的矢量为各个扩散中心点作用力的合力。当某个中心点 C_j 距离样本点 x_i 很远时，由推力控制函数 f 控制的推送距离 $P(x_i,C_j)$将会趋于零，从而消除较远的扩散中心点对样本点施加的影响，保持了原有数据集的空间分布规律，这是合理的。

定义 6 样本 x_i 的新坐标为 $x_i'=x_i+\varphi$，式中的 x_i 为样本的特征向量，与 φ 做矢量加法。对每个样本点进行同样的操作，样本集中的样本被映射到一个新的空间，实验证明在新空间中分类算法的效果有明显提升。

将基于边界点的样本扩散方法描述如下。

输入：带有标签信息的数据集 X，中心点个数 n，推送距离参数 ε

输出：经过样本点扩散后的新数据集 X′

步骤 1 在不同类别的样本点之间建立加权无向图。

步骤 2 在建立的无向图上生成马尔科夫状态转移矩阵。

步骤 3 生成平稳分布，算法[15]收敛后得到样本集的边界值向量 B。

步骤 4 对向量 B 中的元素进行降序排列，选取边界值前 n 位的样本点作为扩散中心点。

步骤 5 对于样本点 x_i，计算其所受扩散中心点的合力 φ，进而计算得到其在新空间中的坐标 x_i'。

步骤 6 对其余样本点进行步骤 5 的操作，得到新的特征空间。

4 实验设计与结果分析

4.1 数据集及评价标准

实验所采用的数据集均存在不同程度的交错现象，由于 SVM 的软间隔和核技巧可以较好地解决非线性数据集分类问题，本文采用 SVM(不加任何预处理)进行分类的准确率来衡量数据集的交错程度。准确率越低，说明交错越严重。

在 UCI 数据集上选择 Wine，Pima，Biodegradation(Bide)，Spambase 4 个常用的数据集，并在其上直接使用 SVM 进行分类，得到的准确率分别为 39.3258%，68.4783%，72.0183%，57.2488%。这说明数据集是合理、有效的，均存在不同程度的交错现象，Wine 的交错程度最严重，Bide 的交错程度相对轻微。数据集的详细信息如表 1 所列。

表 1 数据集信息

	样本数	维度	类别数
Wine	178	13	3
Pima	768	8	2
Bide	1055	41	2
Spambase	4601	57	2

全空间的交错难以可视化，其程度可用前文所述的 SVM 分类准确度近似衡量。为了展示交错现象，以 Wine 为例，随机抽取两个维度并添加类标签，图 1 直观展示了其交错程度。

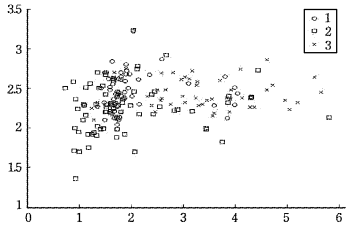

图 1 Wine 的样本交错示意图

本文使用常用的分类评价指标，TP 表示正类样本被判为正类的数目，TN 表示负类样本被判为负类的数目，FP 表示被错判的负类样本的数目，FN 表示被错判的正类样本的数目。实验以准确率为考查指标，其计算公式为：

$$\frac{TP+TN}{TP+FP+FP+TN}$$

4.2 实验设计及参数设置

在各个数据集上采用基于边界的样本扩散方法进行预处理，得到新的样本空间，然后使用逻辑斯蒂回归(LR)、支持向量机(SVM)、k 近邻(KNN) 3 种分类器对数据集进行分类，统计分类准确率并计算 F1 值。使用有监督的度量学习方法 LMNN 和 LDA 对数据集进行预处理并执行同样操作，最后比较准确率和 F1 值。对比实验所使用的 LMNN 和 LDA 两种度量方法的代码来自 Laurens van der Maaten[16] 提供的度量学习工具箱 drtoolbox。

在求取样本集边界向量的过程中，λ 为生成数据集加权无向图时计算边权重的参数，表征衰减程度，常取大于 1 的整数，若其取值太小则不能体现出样本间距离与权重的反比关系，太大则会使得权重衰减过快，本文实验中 λ 取 2。为保证不可约条件，对马尔科夫转移矩阵 P 做调整：

$$P=\alpha\times P+(1-\alpha)\times\frac{U}{m}$$

其中，$\alpha\in[0,1]$，U 为单位矩阵，m 为样本条数，本文调整参数 α 的取值为 0.95。近邻数 k 的合理取值约为 5～30，可根据数据集大小适当调整，本文对 Wine 和 Pima 的取值为 6，对 Bide 和 Spambase 的取值分别为 10 和 12。当某个样本的近邻样本点中没有该样本的任何异类样本点或者异类样本点所占比例极小时，将该样本点的边界值直接取为 0，算法中比例的阈值为 β，实验中取为 0.01。随机游走过程中，将前后两次迭代的边界值差值的阈值设为 0.001。边界值随维度的增大有趋于 0 的趋势，因此对于高维数据，计算推送距离的参数 ε 常取较大值，本文实验中 ε 的取值分别为 20，20，200，200。扩散中

心点的个数 $n=0.2\rho$,ρ 为边界值向量中大于 0 的元素的个数。

4.3 结果及分析

在数据集 Wine 上的实验结果如表 2 所列。其中 Bb-D (Border based diffusion method)是本文提出的方法的简称。

表 2　在数据集 Wine 上的实验结果

	Bb-D	LMNN	LDA	NCA
LR	**0.9775**	**0.9775**	0.9662	0.9213
SVM	**0.9663**	0.4157	0.9213	0.4270
KNN	0.9551	0.8876	**0.9663**	0.6966

从实验结果可以看出,本文提出的方法 Bb-D 在 LR 和 SVM 这两个分类器上的准确率是最高的,在 LR 上 Bb-D 与 LMNN 方法的正确率均为最高,在 KNN 上 Bb-D 的准确率略低于 LDA,但整体上,本文方法对各个分类器的分类效果是非常不错的。其中,4 种方法在运用 KNN 分类器进行分类时,k 均经过多次实验取 20 以内的最优值,下同。

在数据集 Pima 上的实验结果如表 3 所列。

表 3　在数据集 Pima 上的实验结果

	Bb-D	LMNN	LDA	NCA
LR	**0.7976**	0.7857	0.7857	0.7679
SVM	**0.7917**	0.6726	**0.7917**	0.5714
KNN	0.7917	0.7321	**0.7976**	0.7321

结果显示,Bb-D 在 Pima 数据集上的表现与在数据集 Wine 上的表现相似,即在 LR 和 SVM 上均有最优的效果,在 KNN 上其准确率略低于 LDA 方法。总体上,本文方法在各个分类器上都有较高的准确率,表现较为均衡,而 LMNN 和 NCA 在分类器 SVM 上的准确率偏低,LDA 在 LR 上的表现则不如 Bb-D。

在数据集 Bide 上的实验结果如表 4 所列。

表 4　在数据集 Bide 上的实验结果

	Bb-D	LMNN	LDA	NCA
LR	0.8420	**0.8670**	0.8532	0.8578
SVM	0.8303	0.7523	**0.8624**	0.7340
KNN	0.8394	0.8440	0.8624	**0.8670**

结果显示,在 Bide 数据集上,LMNN 在分类器 LR 上有最好的效果,但在 SVM 上的准确率较低;类似地,NCA 在 KNN 分类器上有最高的准确率,但在 SVM 上的表现不佳,本组数据中表现最好的是 LDA,其在各个分类器上都有较高的准确率。Bb-D 在 3 个分类器上的表现均衡,没有明显短板。

在数据集 Spambase 上的实验结果如表 5 所列。

表 5　在数据集 Spambase 上的实验结果

	Bb-D	LMNN	LDA	NCA
LR	0.8451	**0.8591**	0.8402	0.8287
SVM	**0.8583**	0.5725	0.8353	0.6713
KNN	0.8097	0.7125	**0.8254**	0.8023

实验结果显示,Bb-D 方法在 SVM 分类器上的准确率最高,在 LR 和 KNN 上亦有较高的准确率,LMNN 只在 LR 分类器上有较好的效果。

综合以上实验结果可以看出,本文提出的 Bb-D 方法在各个数据集上使用不同分类器时都有较高的准确率,即使在数据集 Bide 上针对 3 个分类器都没有取得最高准确率,但综合来看 Bb-D 仍然是适用性最强的。在数据集 Wine 上,使用 SVM 分类的准确率只有 39.3258%,这说明数据集存在相当严重的交错现象,使用 Bb-D 后,在 LR 和 SVM 这两个分类器上的准确率均达到 97%左右,提升非常明显,且其表现优于对比方法 LMNN,LDA,NCA 的表现。随着 SVM 在原始数据集上的分类准确率的提升,Wine,Spambase,Pima,Bide 使用 Bb-D 的提升效果呈下降趋势,这说明 Bb-D 擅长处理的是交错程度大的数据集。进一步观察发现,当 Bb-D 与 SVM 结合时,SVM 的分类准确率几乎均达到最高,除了在交错情况相对轻微的数据集 Bide 上稍低于 LDA 方法。

不同的度量学习方法在不同数据集、不同分类器上的性能有较大的波动,这是由度量学习本身的特点决定的,本文实验也反映了这种现象。不同的距离度量算法适用于不同的任务和应用场景,很难找到一种能适用于所有问题的距离度量学习算法,在实际应用中,应该根据不同的应用场合选择不同的算法[17]。

综上所述,Bb-D 没有对应用场景做出特殊要求,它擅长处理交错严重的数据集,理论上适用于任何基于距离的机器学习算法;而且它更适合与 SVM 相结合,能有效提升其分类准确度。追其理论根源,我们认为:SVM 的思想是找到一个线性可分超平面,并使用最大间隔来正确地区分二类训练数据,如此即可有效地降低对测试样本分类错误的风险,这种 SVM 称作硬间隔 SVM。但是在训练阶段搜索最优超平面时,由于受到可分边界附近一些错分样本的影响,原有硬间隔 SVM 中最大间隔的原则将失效。为解决该问题,软间隔 SVM 利用松散阈值来得到一个近似线性可分超平面。本文提出的方法恰恰是作用于间隔的,大间隔降低了分类超平面的学习难度。

结束语　本文针对类别交错的数据集提出了一种基于边界的样本扩散方法,该方法在样本空间中学习到各个样本点的边界值,即得到分类面异类样本的交错程度,采取合理的推送距离函数直接作用于交错的分类面,从而缓解了分类面的交错,得到了更清晰的分类面,提高了分类算法的精度。在不同数据集上使用不同的分类器均取得了较高的准确率,而且在不同数据集上针对不同的分类器,本文方法与经典的有监督度量学习方法相比仍有一定的优势。Bb-D 适合处理交错程度高的数据集,而且能有效提升 SVM 的性能。然而,本文提出的方法涉及的参数较多,无法手动寻求最优组合。设置合适的目标函数,使用最优化理论学习出部分参数可能会得到更优的结果,可作为针该方法的另一个研究方向。另外,针对边界值的计算和扩散,可能有其他更好的方法,值得继续探索。

参 考 文 献

[1] CHEN K Z, LE C P, ZHONG S P. Fusion Method of Distance Metric Learning and SVM for Image Matching[J]. Journal of Chinese Computer Systems, 2015, 36(6): 1353-1357. (in Chinese)
陈开志,乐承沛,钟尚平. 融合距离度量学习和 SVM 的图像匹配算法[J]. 小型微型计算机系统,2015,36(6):1353-1357.

(下转第 295 页)

gent Computing. IEEE，2014：1-5.

[16] WEN R，LI D W，LUAN X F，et al. Robot path planning based on ant colony algorithm[J]. Computer and Digital Engineering，2012，40(5)：20-22. (in Chinese)
温瑞，李大伟，栾孝丰，等. 基于蚁群算法的机器人路径规划[J]. 计算机与数字工程，2012，40(5)：20-22.

[17] QIU L L. Robot path planning based on improved ant colony algorithm [D]. Shanghai：Donghua University，2015. (in Chinese)
邱莉莉. 基于改进蚁群算法的机器人路径规划[D]. 上海：东华大学，2015.

[18] LIU K，YOU X M，LIU S. Improved ant colony algorithm for path planning of mobile robot in complex environment[J]. Computer Engineering and Application，2016，52(13)：60-63. (in Chinese)
刘锴，游晓明，刘升. 复杂环境移动机器人路径规划的改进蚁群算法[J]. 计算机工程与应用，2016，52(13)：60-63.

[19] PENG L. Path planning of mobile robot based on genetic algorithm[D]. Changsha：Changsha University of Science and Technology，2013. (in Chinese)
彭丽. 基于遗传算法的移动机器人路径规划[D]. 长沙：长沙理工大学，2013.

[20] LI T X，CHEN G D. Path planning of indoor mobile robot based on improved genetic algorithm [J]. Manufacturing Automation，2015(20)：31-35. (in Chinese)
李天旭，陈广大. 基于改进遗传算法的室内移动机器人路径规划[J]. 制造业自动化，2015(20)：31-35.

[21] ZHANG Y，DAI E C，LUO Y. Path planning of mobile robot based on improved genetic algorithm[J]. Computer Measurement and Control，2016，24(1)：313-316. (in Chinese)
张毅，代恩灿，罗元. 基于改进遗传算法的移动机器人路径规划[J]. 计算机测量与控制，2016，24(1)：313-316.

[22] ZHU Y Y. Path planning of mobile robot based on hybrid particle swarm optimization[D]. Shanghai：Shanghai University of Engineering Science，2015. (in Chinese)
朱莹莹. 基于混合粒子群算法的移动机器人路径规划研究[D]. 上海：上海工程技术大学，2015.

[23] WANG Y，CAO W. A global path planning method for mobile robot based on a three-dimensional-like map[J]. Robotica，2013，32(4)：611-624.

[24] HOU Z W，JIA Y L，WANG Z H，et al. Research on jewelry positioning technology based on the minimum bounding rectangle [J]. Computer Engineering，2016，42(2)：254-260. (in Chinese)
侯占伟，贾玉兰，王志衡，等. 基于最小外接矩形的珠宝定位技术研究[J]. 计算机工程，2016，42(2)：254-260.

(上接第 289 页)

[2] CHOI S. Robust Learning From Demonstration Using Leveraged Gaussian Processes and Sparse Constrained Opimization[C]// IEEE Conference on Robotics and Automation. 2016：470-475.

[3] YANG P，HUANG K，LIU C L. Geometry preserving multitask metric learning[J]. Machine Learning，2013，92(1)：133-175.

[4] JIN R，WANG S，ZHOU Y. Regularized Distance Metric Learning：Theory and Algorithm[C]// Conference on Neural Information Processing Systems 2009. Vancouver，British Columbia，Canada，2009：862-870.

[5] XING E P，NG A Y，JORDAN M I，et al. Distance Metric Learning，With Application To Clustering With Side-Information[J]. Advances in Neural Information Processing Systems，2003，15：505-512.

[6] WANG W. Global and Locality Incorporation in Distance Metric Learning[D]. Hefei：University of Science and Technology of China，2014. (in Chinese)
王微. 融合全局和局部信息的度量学习方法研究[D]. 合肥：中国科学技术大学，2014.

[7] WEINBERGER K Q，SAUL L K. Distance metric learning for large margin nearest neighbor classification[J]. Journal of Machine Learning Research，2009，10：207-244

[8] HUA Y，JIE Y. A direct LDA algorithm for high-dimensional data-with application to face recognition[J]. Pattern Recognition，2001，34(10)：2067-2070.

[9] WANG P Y，CAI D F. Distance-based Kernel Evalution Measure [J]. Computer Science，2014，41(2)：72-75. (in Chinese)
王裴岩，蔡东风. 一种基于核距离的核函数度量方法[J]. 计算机科学，2014，41(2)：72-75.

[10] HE J R，DING L X，HU Q H，et al. Properties of High-dimensional Data Space and Metric Choice[J]. Computer Science，2014，41(3)：212-217. (in Chinese)
何进荣，丁立新，胡庆辉，等. 高维数据空间的性质及度量选择[J]. 计算机科学，2014，41(3)：212-217

[11] KEDEM D，TYREE S，WEINBERGER K，et al. Non-linear Metric Learning[J]. Advances in Neural Information Processing Systems，2012，4：2582-2590.

[12] HE Y，CHEN W，CHEN Y，et al. Kernel Density Metric Learning[C]// 2013 IEEE International Conference on Data Mining (ICDM). 2013：271-280.

[13] JAIN P，KULIS B，DAVIS J V，et al. Metric and Kernel Learning Using a Linear Transformation[J]. Journal of Machine Learning Research，2009，13(1)：519-547.

[14] LIANG L，LI J，HUO Y X，et al. A nonlinear normalization for non-uniformly distributed data [J]. Computer Science，2016，43(4)：264-269. (in Chinese)
梁路，黎剑，霍颖翔，等. 一种非均匀分布数据的非线性标准化方法[J]. 计算机科学，2016，43(4)：264-269.

[15] WAN H Y，ZUO J L，WAN J Y，et al. The KNN Text Classification Based on Sample Importance Principals[J]. Journal of Jiangxi Normal University(Natural Science Edition)，2015(3)：297-303. (in Chinese)
万韩永，左家莉，万剑怡，等. 基于样本重要性原理的 KNN 文本分类算法[J]. 江西师范大学学报(自然科学版)，2015(3)：297-303.

[16] Laurens van der Matten[OL]. http：//lvdmaaten. github. io/drtoolbox.

[17] SHEN Y Y，YAN Y，WANG H Z. Recent Advances on Supervised Distance Metric Learning Algorithms[J]. Acta Automatica Sinica，2014，40(12)：2673-2686. (in Chinese)
沈媛媛，严严，王菡子. 有监督的距离度量学习算法研究进展[J]. 自动化学报，2014，40(12)：2673-2686.