Journal of Computer Applications
计算机应用,2012,32(12):3311-3314
ISSN 1001-9081
CODEN JYIIDU
2012-12-01
http://www.joca.cn

文章编号:1001-9081(2012)12-3311-04
doi:10.3724/SP.J.1087.2012.03311

流形上的 k 最近邻分类方法

文志强*,胡永祥,朱文球
(湖南工业大学 计算机与通信学院,湖南 株洲 412007)
(*通信作者电子邮箱 zhqwen20001@163.com)

摘　要: 针对分类数据中存在噪声样本和维数问题,提出了流形上的 k 最近邻方法。首先,利用贝叶斯公式对经典 k 最近邻方法进行扩展,并采用核概率密度方法估计样本的局部联合概率密度;其次,建立噪声样本点对模型,并构建改进的边际本征图和相应的权值矩阵,通过定义目标函数寻找最优降维映射矩阵;最后,提出一个完整的流形上 k 最近邻算法。与6种经典方法在12个常用数据集上的实验比较表明,在大多数情况下所提方法的分类性能要优于其他方法。

关键词: k 最近邻;噪声样本;降维;分类器;流形
中图分类号: TP391　　**文献标志码:** A

k-nearest neighbors classifier over manifolds

WEN Zhi-qiang*, HU Yong-xiang, ZHU Wen-qiu
(*School of Computer and Communication, Hunan University of Technology, Zhuzhou Hunan* 412007, *China*)

Abstract: For resolving the problem of the existing noise sample and large number of dimensions, the k-nearest neighbors classifier over manifolds was presented in this paper. Firstly the classic k-nearest neighbors was extended by Bayes theorem and local joint probability density was estimated by kernel density estimation in classifier. In addition, after building the noise sample model, an objective function was defined via improved marginal intrinsic graph and its weight matrix for searching the optimal dimension reduction mapping matrix. At last, details about k-nearest neighbors algorithm over manifolds were provided. The experimental results demonstrate that the presented method has lower classification error rate than six kinds of classic methods in most cases on twelve data sets.

Key words: k-Nearest Neighbors (kNN); noise sample; dimensionality reduction; classifier; manifold

0　引言

k 最近邻(k-Nearest Neighbors, kNN)方法是一种简单而高效的监督分类方法,其思想是找到测试样本在训练样本集中的 k 个最近邻对象,然后根据这些对象的类别属性进行投票,决定测试样本的类别属性。kNN 方法的优势在于无须事先知道训练样本的属性值分布,而且不要求获取显式规则。一般来说,kNN 的分类准确率要高于其他分类方法,且有坚实理论基础,包括误差估计[1]和误差边界[2]等理论,目前在文本分类、模式识别、图像分类等领域得到广泛应用。然而,kNN 算法也有一些不足:首先,由于存在维数灾难问题,kNN 算法不适合用于高维数据的分类;其次,当类条件概率分布为不可分或重叠时,邻域中的训练样本将会具有不同类标记。当这些具有不同标记的样本数相当时,kNN 的分类错误率将会变大。本文从上述两点不足出发,研究改善 kNN 算法性能的方法。为了简单起见,称落入贝叶斯决策边界错误区内的样本为噪声样本,这种噪声样本的存在会导致分类误差率增大。显然,如果训练样本中不存在噪声样本,这就意味着分类中具有不同标记的样本重叠问题就不存在,就不会出现错判。目前有很多针对 kNN 算法的研究:1)有一些研究样本距离的选择问题,如选用加权距离[3]、语义距离[4]、Mahalanobis 距离[5]、Chi-squared 距离[6]、最大最近距离[7]等来构造 kNN 算法。这些方法能提供平滑后的类条件概率估计,有一定的去噪功能,能改善 kNN 算法的分类性能。2)研究基于降维策略的最近邻规则[8-9],这种策略能降低维数影响,改善高维数据的最近邻分类性能。3)考虑噪声样本对分类的影响,引入相应的评价方法,减少噪声样本对分类的影响,如 Li 等[10]提出了局部概率中心(Local Probability Center, LPC)的最近邻方法,利用测试样本的局部 k 最近邻样本后验概率获取测试样本在每类标记中局部概率中心,根据测试样本到局部概率中心的距离实现分类。Hotta 等[11]提出了 kNN 分类均值模式(Categorical Average Pattern, CAP)方法,计算被分类对象在每类中的局部概率中心,然后根据被分类对象与这些局部类中心的距离实现分类。另外,近来有一些新进展,出现了基于代表点选择主特征的分类方法[12]、类条件最近邻方法[13]、k 最近邻均值分类方法[14]等,这些方法都被证明能改善 k 最近邻算法的分类性能。然而,据所查文献了解,很少有针对减少高维数据中噪声样本影响的 kNN 方法研究。

流形学习就是从高维采样数据中恢复低维流形结构,即找到高维空间中的低维流形,并求出相应的嵌入映射,以实现维数约简或者数据可视化。自2000年以来,大量流形学习算法被先后提出,最具有代表性的无监督流形学习算法主要有等距映射(ISOmetric MAPping, ISOMAP)算法[15],局部线性嵌入(Locally Linear Embedding, LLE)[16]、拉普拉斯特征映射

收稿日期:2012-06-13;**修回日期:**2012-07-21。　**基金项目:**国家自然科学基金资助项目(61170102);湖南省自然科学基金资助项目(11JJ3070, 10JJ3002, 11JJ4050);湖南省教育厅科研资助项目(12A039,12A042)。

作者简介:文志强(1973-),男,湖南湘乡人,副教授,博士,CCF 会员,主要研究方向:图像处理、视觉跟踪;　胡永祥(1973-),男,湖南安化人,副教授,博士,主要研究方向:图像配准、模式识别;　朱文球(1969-),男,湖南攸县人,教授,主要研究方向:数字图像处理、模式识别。

(Laplacian Eigenmap)算法[17]、局部切空间排列(Local Tangent Space Alignment, LTSA)[18]等。在流形学习算法中引入类别信息,则可以获得监督流形学习算法,典型算法有监督LLE(Supervised-LLE, SLLE)[19]、局部Fisher嵌入(Local Fisher Embedding, LFE)[20]、局部判别嵌入(Local Discriminant Embedding, LDE)[21]等。融入类别信息的监督流形学习算法更适合于分类,特别是最近Yan等[22]给出基于图表达的通用数据嵌入框架,并提出了边际Fisher分析(Marginal Fisher Analysis, MFA)算法,定义了描述类别内紧密性和类别间分离性准则,以保持类内特征向量间距离不变,扩展类间特征向量之间的距离,以提高本征样本的可分性。本文在LDE和MFA方法的基础上,考虑类间样本重叠区域中的噪声样本,提出监督流形学习上的kNN方法。

1 k最近邻规则

设$z_i \in \mathbf{R}^d (i = 1,2,\cdots,n)$是$d$维样本,$y_i \in \{1,2,\cdots,c\}$是相应的类标记,这里$n$是样本数,$c$是类别数。$n_l$表示类别为$l$的样本数,满足$\sum_{l=1}^{c} n_l = n$。让$Z$为所有样本集合$Z = (z_1 \mid z_2 \mid \cdots \mid z_n)$,其标记集$Y = (y_1 y_2 \cdots y_n)$,$z$为测试样本点。$P(l \mid z,Z,Y)$为给定样本集$Z$及相应标记集$Y$下,测试样本$z$属于类别$l$的概率($l = 1,2,\cdots,c$),则$z$的类别是:

$$w = \arg\max_l P(l \mid z,Z,Y) \tag{1}$$

如果把一个体积放在z周围,并且能够包含进k个样本,其中第l类的样本数为k_l,满足$k = \sum_{l=1}^{c} k_l$,$P(l \mid z,Z,Y) = k_l/k$,则z的类别是$w = \arg\max_l k_l$。也即将一个测试数据点z分类为与它最接近的k个近邻中出现最多的那个类别。而最近邻规则是k近邻规则的特殊情形($k = 1$),即对于测试样本点z,在样本集Z中距离它最近的点记为z',那么最近邻规则的分类方法是把样本点z判为z'所属的类别。

2 k最近邻规则的扩展

考虑z的局部邻域样本集U,$\varphi_l(z)$表示样本z在标记为l的样本子集中k-最近邻集,满足$Z = Z - U \cup U$且$U = \varphi_1(z) \cup \varphi_2(z) \cup,\cdots,\cup \varphi_c(z)$。利用贝叶斯公式

$$P(l \mid z,U,Z-U,Y) = \frac{p(z,U \mid l)P(l \mid Z-U,Y)}{p(z,U \mid Z-U,Y)}$$

则

$$P(l \mid z,Z,Y) = P(l \mid z,U,Z-U,Y) \propto p(z,U \mid l)P(l \mid Z-U,Y)$$

$P(l \mid Z-U,Y)$为先验概率,与训练样本无关且假设先验概率相等,因此可以转化为:

$$P(l \mid z,Z,Y) \propto p(z,U \mid l)$$

对于给定类别标记l,似然函数$p(z,U \mid l)$仅与$\varphi_l(z)$相关,因此得:

$$P(l \mid z,Z,Y) \propto p(z,\varphi_l(z) \mid l)$$

上述概率中l,$\varphi_l(z)$都声明了l类别,所以有:

$$P(l \mid z,Z,Y) \propto p(z,\varphi_l(z))$$

概率$p(z,\varphi_l(z))$表示z和$\varphi_l(z)$的联合概率密度,则式(1)转化成式(2):

$$w = \max_l p(z,\varphi_l(z)) \tag{2}$$

概率函数$p(z,\varphi_l(z))$采用核概率密度估计方法,给定在d维空间$\mathbf{R}^d$中有k个数据点$\boldsymbol{b}_{lj} \in \varphi_l(z)$,$j = 1,2,\cdots,k$,$l = 1,2,\cdots,c$。点集$\varphi_l(z)$关于核函数$K(z)$和$d \times d$的带宽矩阵$\boldsymbol{H}$的多元核函数密度估计为:

$$\hat{p}(z,\varphi_l(z)) = \frac{1}{k}\sum_{j=1}^{k} K_H(z - b_{lj})$$

式中$K_H(z) = |\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{-1/2}z)$。$d$元核函数$K(z)$为具有紧支集的有界函数,满足下列四个条件:

① $\int_{R^d} K(z)\,dz = 1$

② $\lim_{\|z\|\to\infty} \|z\|^d K(z) = 0$

③ $\int_{R^d} zK(z)\,dz = 0$

④ $\int_{R^d} zz^{\mathrm{T}}K(z)\,dz = c_{\mathrm{kernel}}\boldsymbol{I}$

式中:c_{kernel}为常数,$c_{\mathrm{kernel}}\boldsymbol{I}$是核函数$K(z)$的协方差矩阵。为简化处理,通常采用一类特殊的径向对称核函数满足:

$$K(z) = c_{\mathrm{kernel},d}\,\mathrm{kernel}(\|z\|^2)$$

式中系数$c_{\mathrm{kernel},d}$选取的原则是保证$K(z)$的积分为1。核函数kernel(·)为非负、递减、连续且有界函数。完整的参数表示$\boldsymbol{H}$会增加估计的复杂性,实际中,$\boldsymbol{H}$可以为对角矩阵$\boldsymbol{H} = \mathrm{diag}[h_1^2,h_2^2,\cdots,h_d^2]$或为$h^2\boldsymbol{I}$,$\boldsymbol{I}$为$d \times d$为单位矩阵。为了简单起见,使用后一种$\boldsymbol{H}$可以进一步降低密度估计的复杂度,这样只需确定一个带宽参数$h > 0$即可。值得注意的是,首先需要保证特征空间具有有效的欧几里得尺度。带宽为$h^2\boldsymbol{I}$时,核密度估计函数就可写成如下形式。

$$\hat{p}(z,\varphi_l(z)) = \frac{1}{kh^d}\sum_{j=1}^{k} K\left(\frac{z - \boldsymbol{b}_{lj}}{h}\right)$$

这就是著名的Parzen窗方法。联合概率函数$p(z,\varphi_l(z))$估计为式(3):

$$\hat{p}(z,\varphi_l(z)) = \frac{c_{\mathrm{kernel},d}}{kh^d}\sum_{j=1}^{k}\mathrm{kernel}\left(\left\|\frac{z - \boldsymbol{b}_{lj}}{h}\right\|^2\right) \tag{3}$$

函数kernel(·)采用高斯函数。式(3)可以解释为,将在测试样本z为中心的局部函数的平均值作为该样本的概率密度函数估计值,参数h就决定了局部区域的大小。

3 流形上的k近邻规则

假设M是一个嵌入在m维空间的d维流形,具有产生函数$f(\cdot)$,$z \in \mathbf{R}^d$且$\boldsymbol{x} \in \mathbf{R}^m$,这里$d \ll m$。给定一系列$n$个$m$维向量样本集构成的矩阵$\boldsymbol{X} = [\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]$,降维后的特征向量$z_i \in \mathbf{R}^d$,则有非噪声模型:$z_i = f(\boldsymbol{x}_i)$。设有测试样本$\boldsymbol{x} \in \mathbf{R}^m$,则有:

$$w = \max_l p(f(\boldsymbol{x}),\varphi_l(f(\boldsymbol{x}))) \tag{4}$$

考虑由矩阵$\boldsymbol{X}$中列构成的样本子集$\boldsymbol{\Phi}$中向量$\boldsymbol{x}_j$,满足线性关系$z_j = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}_j$,其中,$\boldsymbol{V}$表示子集$\boldsymbol{\Phi}$中样本的映射矩阵($m \times d$),T表示矩阵转置操作。设$\boldsymbol{a}_{lj} \in \boldsymbol{\Phi}$表示第$l$类中的第$j$个样本向量,其中,$l = 1,2,\cdots,c$并且$j = 1,2,\cdots,k_l'$。$k_l'$为集合$\boldsymbol{\Phi}$中第$l$类样本总数,实验中取$k_1' = k_1' = \cdots = k_c' = k'$,则$p(f(\boldsymbol{x}),\varphi_l(f(\boldsymbol{x})))$转化为$p(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x},\varphi_l(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}))$。考虑集合$\boldsymbol{\Phi}$中样本$\boldsymbol{x}$需实现$m$维嵌入空间映射到$d$维本征空间,定义目标函数如式(5)所示,在该式中,$\boldsymbol{x}_i,\boldsymbol{x}_j \in \boldsymbol{\Phi}$,$w_{ij}$和$w'_{ij}$分别表示由样本集$\boldsymbol{\Phi}$中样本点(看成是图的节点)构成的邻域关系本征图$G$权值矩阵$\boldsymbol{W}$和惩罚图$G'$权值矩阵$\boldsymbol{W}'$中的元素。易知,目标函数(5)可以转化为式(6)泛化本征值问题。即相当于寻找式(6)的泛化本征向量$\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{N_c}$,对应于$N_c$个最大特征值。

$$\max_{\boldsymbol{V}} J(\boldsymbol{V}) = \sum_{i,j}\|\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{x}_j)\|^2 w'_{ij} \tag{5}$$

$$\text{s.t.}\ \sum_{i,j}\|\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{x}_j)\|^2 w_{ij}=1$$
$$\boldsymbol{A}\boldsymbol{v}=\lambda\boldsymbol{B}\boldsymbol{v} \tag{6}$$

其中,$\boldsymbol{A}=\boldsymbol{X}(\boldsymbol{D}'-\boldsymbol{W}')\boldsymbol{X}^{\mathrm{T}}$,$\boldsymbol{B}=\boldsymbol{X}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{X}^{\mathrm{T}}$。$\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)$,$N$ 为集合 $\boldsymbol{\Phi}$ 中样本总数。矩阵 $\boldsymbol{D}$ 和 $\boldsymbol{D}'$ 是对角矩阵,满足 $D_{ii}=\sum_j w_{ji}$ 且 $D'_{ii}=\sum_j w'_{ji}$,则 $\boldsymbol{D}'-\boldsymbol{W}'$ 和 $\boldsymbol{D}-\boldsymbol{W}$ 称为图 G' 和 G 的 Laplacian 矩阵。

目前有一些研究从不同的角度来设置相应的本征矩阵和惩罚矩阵。文献[21]为了保持邻近同类节点之间的距离不变,而不同类邻近节点之间的距离增大,设置矩阵 $\boldsymbol{W}$ 中,当同类样本点对时 $w_{ij}=1$,否则为 $w_{ij}=0$;设置矩阵 $\boldsymbol{W}'$ 中,当不同类样本点对时 $w'_{ij}=1$,否则为 $w'_{ij}=0$。文献[22]考虑保持邻近同类节点间的距离不变,而类间邻近节点间的距离增大,设置矩阵 $\boldsymbol{W}$ 与文献[21]相同,而设置矩阵 $\boldsymbol{W}'$ 中,当不同类边际近邻样本点时 $w'_{ij}=1$,否则为 $w'_{ij}=0$。FDA 方法中考虑类间样本协方差矩阵作为本征图的权值矩阵,而类内样本协方差矩阵作为惩罚图的权值矩阵[23-24]。这些方法可改变样本点间样本点分布,减少样本点分布区域的重叠,提供样本点的可分性。尽管这样,由于噪声样本的影响,样本点分布区域重叠的减少是相对有限的,甚至有时会增加分布区域的重叠。因此,本节将介绍这些方法的缺陷,探讨解决方法。

3.1　噪声样本点对

影响分类性能最关键因素是不同标记样本之间有一个重叠区域,如图1所示的 A 和 B 区域,这主要是由于样本噪声影响或样本分布模型的不确定性造成的。因此,在进行降维时要考虑重叠区域特征。

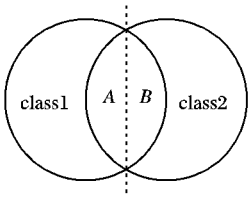

图1　两类样本的重叠区域

在 MFA 方法[22]中,利用类内的样本构建本征图如图 2(a),且利用近邻图的类间边际样本构建惩罚图如图2(b),并利用 Fisher 准则实现线性降维,但这方法在实际中存在一些限制。如图 2(c)中,$(a,b)(c,d)(e,f)(g,h)$ 互为邻域节点,但点对 (c,d) 与其他点对有一些区别,如点 c 偏离 class1 的密度中心,d 偏离 class2 的密度中心,而其他点 bfh 及 aeg 都靠近其各自的密度中心。在文献[22]中只考虑了 $(a,b)(e,f)(g,h)$ 互为邻域节点,而没有考虑 (c,d) 这种近邻节点。实际上这种 (c,d) 近邻节点是实实在在存在的,而且将会对分类之前的学习及分类准确率有较大影响。同样 LDE 和 FDA 方法不能处理这种问题。因此在样本特征降维时,考虑这样的点对,使得类似于 (c,d) 点对之间的距离更近一些。为了方便,称这样的点对为噪声样本点对(Noise Sample Pairs, NSP)。那么如何确定这样的 NSP 集呢? 以两类问题为例,如图2(d)所示,设 o_1 和 o_2 分别是两类的中心。对于 $NSP(c,d)$ 应满足:$|co_2|<|do_2|$ 或 $|co_1|>|do_1|$,其中,$|\cdot|$ 表示距离。因此求 NSP 的方法是:先使用训练样本获取各类的中心 $\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_c$,然后获取集合 $\boldsymbol{\Phi}$ 中任意 $\boldsymbol{x}_i$ 的 g'- 近邻集合 $\psi(\boldsymbol{x}_i)$,寻找满足式(9)条件的噪声点对构成集合 $\boldsymbol{\Pi}$。$d(x,y)$ 表示 x 到点 y 的欧氏距离,具有类标记的样本集 $\boldsymbol{\Phi}$ 为 $\{\boldsymbol{x}_i,y_i\}$,$\boldsymbol{x}_i\in\mathbf{R}^m$,$y_i\in\{1,2,\cdots,c\}$。

$$\boldsymbol{\Pi}=\{\forall(\boldsymbol{x}_i,\boldsymbol{x}_j)\mid \boldsymbol{x}_i\in\boldsymbol{\Phi}\wedge\boldsymbol{x}_j\in\psi(\boldsymbol{x}_i)\wedge d(\boldsymbol{x}_i,\boldsymbol{\mu}_{y_i})>d(\boldsymbol{x}_j,\boldsymbol{\mu}_{y_i})\wedge y_i\neq y_j\} \tag{8}$$

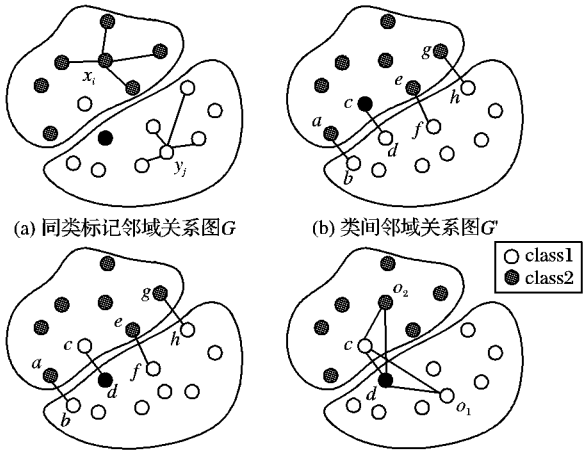

图2　邻域关系图

3.2　流形上的 k 最近邻算法

嵌入空间的维数为 m,$\boldsymbol{x}$ 为测试样本点。$\boldsymbol{\pi}(\boldsymbol{x})$ 表示在集合 $\boldsymbol{\Phi}$ 中与 $\boldsymbol{x}$ 在同类标记中的 g- 近邻集合,$\psi(\boldsymbol{x})$ 表示集合 $\boldsymbol{\Phi}$ 中的 g'- 近邻集合,d 表示本征维数。算法的步骤如下所示。

步骤 1　分别提取测试样本 $\boldsymbol{x}$ 在类别 i 中 s- 邻域集,组合成样本子集 $\boldsymbol{\Phi}$。构造集合 $\boldsymbol{\Phi}$ 中本征图 G 和惩罚图 G',即获取 $\pi(\boldsymbol{x}_i)$ 和 $\psi(\boldsymbol{x}_i)$,$\forall\boldsymbol{x}_i\in\boldsymbol{\Phi}$。

步骤 2　计算图 G 权值矩阵 $\boldsymbol{W}$(i,j 表示集合 $\boldsymbol{\Phi}$ 样本序号):

$$w_{ij}=\begin{cases}1, & \boldsymbol{x}_j\in\boldsymbol{\pi}(\boldsymbol{x}_i)\\ 0, & \text{其他}\end{cases}$$

其中 $w_{ji}=w_{ij}$。

步骤 3　计算图 G' 权值矩阵 $\boldsymbol{W}'$:

$$w'_{ij}=\begin{cases}1, & \boldsymbol{x}_j\in\psi(\boldsymbol{x}_i)\ \&\ y_i\neq y_j\ \&\ (\boldsymbol{x}_i,\boldsymbol{x}_j)\notin\boldsymbol{\Pi}\\ b, & \boldsymbol{x}_j\in\psi(\boldsymbol{x}_i)\ \&\ y_i\neq y_j\ \&\ (\boldsymbol{x}_i,\boldsymbol{x}_j)\in\boldsymbol{\Pi}\\ 0, & \text{其他}\end{cases}$$

其中 $w'_{ji}=w'_{ij}$。

步骤 4　求向量方程式(6)最大的 d 个广义特征值及其特征向量 $\boldsymbol{v}_1\boldsymbol{v}_2\cdots\boldsymbol{v}_d$。

步骤 5　计算 $\boldsymbol{x}$ 的本征值 $\boldsymbol{z}$,$\boldsymbol{x}_i$ 的本征值 z_i,即 $\boldsymbol{z}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}$ 和 $z_i=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}_i$,其中 $\boldsymbol{V}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_d)$。

步骤 6　构成 k 近邻的局部集合 $U=\varphi_1(z)\cup\varphi_2(z)\cup\cdots\cup\varphi_c(z)$,并使用式(3)估计似然概率 $p(\boldsymbol{z},\varphi_i(\boldsymbol{z}))$。

步骤 7　使用式(2)估计 $\boldsymbol{x}$ 的类别。

4　实验结果及比较

实验用的计算机性能为 Intel 芯片,2.2 GHz CPU 主频,3 GB内存,操作系统为 Windows XP,采用 Matlab 7.0 作为编程工具。采用 UCI 数据库(http://www.ics.uci.edu/~mlearn/MLRepository.html)中常用的 12 个数据集,与一些经典方法进行了实验比较,以验证本文方法的有效性。数据集的相关参数如表 1 所示,这些数据集包括各种范围的数据集大小和样本特征维数。在进行实验之前,先对所有样本的各个特征维度进行归一化处理,使训练数据集的中心为 0,方差为 1,同样,测试数据特征使用相应的训练中心和方差进行归

一化。相应参数都由交叉验证实验来决定。参数 k 对于所有分类器来说是一个常用的参数，考虑在所有的实验中让 $k \in [3,16]$，本文方法及相关方法涉及的特征本征维数 d 由全局主成分分析(Principal Component Analysis, PCA)保留方差的99%获得，具体见表1中的参数。本文方法中的 s 设为16，g 和 g' 设为8。b 一般设置在 -1 到1之间，实验中 $b = -1$。

表1 数据集及参数描述

数据集名称	样本数	类数	特征数	d
iris	150	3	4	3
Bupa	345	2	6	6
ionosphere	351	2	34	28
Pima	768	2	8	8
Sonar	208	2	60	39
balance scale	625	3	4	4
abalone	4177	3	8	6
Cloud	2048	2	10	7
glass	214	6	9	7
segment	2310	7	19	12
wine	178	3	13	12
Letter recognition	20000	26	16	15

另外，不同训练数据集和测试数据集会导致分类器获得不同的分类器性能，不同的 k 值也会获得不同的分类器性能。因此，为了很好地评价分类方法，采取当 k 取遍3 ~ 16所有值时的平均值，并计算10次这样的值再一次求平均值作为分类器性能的度量标准。另外，每次计算均值时，随机选取每个数据集中50%数据作为训练样本，其他作为测试样本，以此来保证分类方法的分类性能测定的稳定和可靠。

分类性能评价由分类误差率来描述。就分类误差率来说，对于多分类问题，有两个方面的参数：平均分类误差率 Err 及分类误差率方差 div。其中分类误差率方差表明分类误差率分布的均匀程度，其值越大，说明分类误差率分布越不均匀。为了验证分类算法的性能，本文提出的流形上 k 近邻算法(k-Nearest Neighbors over Manifolds, kNNM)与三个经典 kNN及其改进方法进行比较，这些方法是：1)经典 kNN算法，是最基本的，广泛使用于模式识别的分类方法；2)CAP算法[11]，使用分类 kNN样本的平均模式实现分类；3)使用局部概率中心的最近邻算法[10]。另外与三个经典降维及分类方法进行比较，这些方法表示为局部Fisher鉴别分析[24]及 kNN方法(LFDA + kNN)，局部判别嵌入[21]及 kNN邻方法(LDE + kNN)，边际Fisher分析[17]及 kNN方法(MFA + kNN)。表2分别为12个数据集使用上述6个分类方法及本文方法的平均分类误差率及分类误差率方差。从表2可看出，本文方法的平均分类误差率对于大多数样本集来说都是最低的，而且12个数据集的平均分类误差率也是最小的(21.96%)。由此可见本文方法比其他六种方法更有效。

表2 分类算法的分类误差比较

数据集	LPC		CAP		kNN		kNNM		LFDA + kNN		MFA + kNN		LDE + kNN	
	Err/%	div	Err/%	div	Err/%	div	Err/%	div	Err/%	div	Err/%	div	Err/%	div
iris	5.44	5.37	5.27	4.99	5.47	4.7200	5.08	4.16	3.60	1.89	5.20	5.66	5.73	4.77
Bupa	38.93	18.93	41.83	15.40	40.28	6.5500	30.94	6.18	44.15	7.22	40.57	6.31	39.51	6.81
ionosphere	16.12	13.17	15.34	16.99	20.01	17.6000	16.07	14.46			14.21	12.50	18.58	15.50
Pima	27.97	6.00	28.93	10.49	32.74	11.5800	23.96	9.36	34.92	4.18	33.83	13.33	33.96	10.89
Sonar	17.48	5.53	18.17	9.03	17.38	7.2000	17.02	2.51	27.19	6.02	32.18	4.07	17.62	9.26
balance scale	30.61	31.12	30.68	36.40	42.27	40.8400	29.29	36.27	22.47	23.08	43.10	38.82	43.01	40.30
abalone	46.38	9.03	50.06	4.29	50.43	10.8900	36.46	3.42	50.27	10.42	50.14	10.71	50.43	10.28
Cloud	51.09	1.77	51.20	2.05	50.59	0.0124	51.14	2.10	51.69	1.27	51.89	2.73	52.41	1.10
glass	41.50	25.84	44.07	27.10	36.18	21.3400	30.66	17.39	40.84	22.31	36.68	24.23	36.91	21.84
segment	6.17	6.78	6.20	6.05	4.88	4.6200	4.92	7.73			11.20	11.08	10.49	9.48
wine	2.69	2.35	2.75	3.09	4.47	5.8400	2.18	2.35	2.76	3.31	5.78	5.41	4.89	62.20
Letter recognition	15.79	6.14	16.28	6.62	13.86	6.6000	15.86	7.23	11.84	10.00	14.00	4.99	13.24	5.77
Avg	25.01	11.00	25.89	11.87	26.54	11.4800	21.96	9.43	28.97	8.97	28.23	11.65	27.23	16.51

5 结语

本文针对分类数据中存在噪声样本和高维数问题，运用贝叶斯公式和流形降维方法，对常用 k 最近邻方法进行了扩展，提出了流形上的 k 最近邻方法，其目的是通过分析噪声样本点对的特征，利用有监督降维映射方法，以改善降维后样本特征的可分性，从而达到提高样本分类准确率的目的。文中大量实验和比较验证了本文方法的有效性。虽然本文方法提高了分类的准确性，但实际上还有很多需要进一步研究的地方，如本文方法的期望风险分析、误差估计等。

参考文献：

[1] LOIZOU G, MAYBANK S J. The nearest neighbor and the Bayes error rates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1987, 9(2): 254 - 262.

[2] BAX E. Validation of k-nearest neighbor classifiers[J]. IEEE Transactions on Information Theory, 2012, 58(5): 3225 - 3234.

[3] PAREDES R, VIDAL E. Learning weighted metrics to minimize nearest-neighbor classification error[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(7): 1100 - 1110.

[4] 杨立，左春，王裕国. 基于语义距离的K-最近邻分类方法[J]. 软件学报，2005，16(12)：2054 - 2062.

[5] VERDIER G, FERREIRA A. Adaptive Mahalanobis distance and k-nearest neighbor rule for fault detection in semiconductor manufacturing[J]. IEEE Transactions on Semiconductor Manufacturing, 2011, 24(1): 59 - 68.

[6] DOMENICONI C, PENG J, GUNOPULOS D. Locally adaptive metric nearest neighbor classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(9): 1281 - 1285.

[7] SAMET H. K-nearest neighbor finding using MaxNearestDist[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(2): 243 - 252.

(下转第3352页)

4 结语

为了有效提高服务计算的效率,本文将情感引入到蚁群算法中,形成一种情感行为的蚁群算法(简称情感蚁群算法),该方法从认知及推理、情感变化、效用和情感与认知映射角度展开,并给出具体的描述与表述方法,然后将这些表述纳入到蚁群算法中形成满足情感与认知映射的新算法,该算法有效增加了传统蚁群算法的智能性和扩展性。最后将情感蚁群算法实验测试到服务组合中表明是可行性的和合理的。

下一步工作将继续研究情感蚁群算法应用的稳定性和收敛性,从而形成一个整体的情感蚁群算法。

参考文献:

[1] CANFORA G, PENTA M D, ESPOSITO R, *et al*. A framework for QoS-aware binding and re-binding of composite Web services[J]. Journal of Systems and Software, 2008, 81(10):1754-1769.

[2] HUANG A F M, LAN C W, YANG S J H. An optimal QoS-based Web service selection scheme[J]. Information Sciences, 2009, 179(9): 3309-3322.

[3] LIN C F, SHEU R K, CHANG Y S, *et al*. A relaxable service selection algorithm for QoS-based Web service composition[J]. Information & Software Technology, 2011, 53(12): 1370-1381.

[4] 张国丽,李祚泳.基于蚁群算法的情感模型研究[J].计算机应用,2009,29(10):2758-2761.

[5] PICARD R W. Affective computing: Challenges[J]. International Journal of Human-Computer Studies, 2003, 59(1):55-64.

[6] AMMAR M B, NEJI M, ALIMI A M, *et al*. The affective tutoring system[J]. Expert Systems with Applications, 2010, 37(4):3013-3023.

[7] 胡记文,尹全军,陈伟,等.情感影响下的人类认知行为建模研究概述[J].系统仿真学报,2012,24(3):515-519.

[8] 李海芳,何海鹏,陈俊杰.性格、心情和情感的多层情感建模方法[J].计算机辅助设计与图形学学报,2011,23(4):725-730.

[9] ORTONY A, CLORE G L, COLLINS A. The cognitive structure of enotions[M]. Cambridge: Cambridge University Press, 1990.

[10] PICARD W. Affective computing[M]. Cambridge: MIT Press, 1997.

[11] KSHIRSAGAR S, MAGNENAT-THALMANN N. A multilayer personality model[C]// Proceedings of the 2nd International Symposium on Smart Graphics. New York: ACM, 2002: 107-115.

[12] ROSEMAN I J, JOSE P E, SPINDEL M S. Appraisals of emotioneliciting events: Testing a theory of discrete emotions[J]. Journal of Personality and Social Psychology, 1990, 59(5):899-915.

[13] VELASQUEZ J D. Modeling emotions and other motivations in synthetic agents[C]// Proceedings of the 14th National Conference on Artificial Intelligence and 9th Conference on Innovative Applications of Artificial Intelligence. Rhode Island: AAAI Press, 1997: 10-15.

[14] 宋峻峰,张维明,姚莉,等. OWL DL的形式化基础研究[J].小型微型计算机系统,2005,26(2):297-301.

[15] 曹逸,徐德智,陈建二,等.一种带传递关系的认知描述逻辑研究[J].计算机研究与发展,2009,46(3):452-458.

[16] 段海滨.蚁群算法原理及其应用[M].北京:科学出版社,2005.

[17] LÉCUÉ F, MEHANDJIEV N. Seeking quality of Web service composition in a semantic dimension[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(6):942-958.

(上接第3314页)

[8] JING P, HEISTERKAMP D R, DAI H K. LDA/SVM driven nearest neighbor classification[J]. IEEE Transactions on Neural Networks, 2003, 14(4):940-942.

[9] HASTIE T, TIBSHIRANI R. Discriminant adaptive nearest neighbor classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(6): 607-615.

[10] LI BOYU, CHEN YUNWEN, CHEN YANQIU. The nearest neighbor algorithm of local probability centers[J]. IEEE Transactions on Systems, Man, and Cybernetics, PART B: Cybernetics, 2008, 38(1):141-154.

[11] HOTTA S, KIYASU S. Pattern recognition using average patterns of categorical k-nearest neighbors[C]// Proceedings of the 17th International Conference on Pattern Recognition. Piscataway: IEEE, 2004, 4:412-415.

[12] GARCIA S, DERRAC J, CANO J R, *et al*. Prototype selection for nearest neighbor classification: Taxonomy and empirical study[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(3):417-435.

[13] MARCHIORI E. Class conditional nearest neighbor for large margin instance selection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(2): 364-370.

[14] VISWANATH P. An improvement to k-nearest neighbor classifier[C]// Proceedings of IEEE Conference on Recent Advances in Intelligent Computational Systems. Piscataway: IEEE, 2011:227-231.

[15] TENENBAUM J B, SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500):2319-2323.

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

[17] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003, 15(6):1373-1396.

[18] ZHANG Z, ZHA H. Principal manifolds and nonlinear dimension via local tangent space alignment[J]. SIAM Journal Scientific Computing, 2004, 26(1):319-338.

[19] ZHAO L, ZHANG Z. Supervised locally linear embedding with probability-based distance for classification[J]. Computers and Mathematics with Applications, 2009, 57(6): 919-926.

[20] RIDDER D, LOOG M, REINDERS M. Local Fisher embedding[C]// Proceedings of the 17th International Conference on Pattern Recognition. Piscataway: IEEE, 2004, 2: 295-298.

[21] CHEN H, CHANG H, LIU T. Local discriminant embedding and its variants[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2005, 2: 846-853.

[22] YAN SHUICHENG, XU DONG, ZHANG BENYU, *et al*. Graph embedding and extensions: a general framework for dimensionality reduction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(1):40-51.

[23] KIM S, OOMMEN B J. On using prototype reduction schemes to optimize kernel-based Fisher discriminant analysis[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2008, 38(2): 564-570.

[24] SUGIYAMA M. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis[J]. Journal of Machine Learning Research, 2007, 8(5): 1027-1061.