CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 37 卷第 10 期 2015 年 10 月
Vol. 37, No. 10, Oct. 2015

文章编号:1007-130X(2015)10-1965-06

一种高效的 K 值自适应的 SA-KNN 算法*

孙　可[1,3],龚永红[1,2],邓振云[1,3]

(1.广西师范大学广西多源信息挖掘与安全重点实验室,广西 桂林 541004;2.桂林航天工业学院,广西 桂林 541004)
3.广西师范大学计算机科学与信息工程学院,广西 桂林 541004)

摘　要:传统的 K 近邻(KNN)分类算法在实际应用过程中存在一些缺陷:没有考虑去除噪声样本,也没有考虑到在样本数据空间变换过程中保持样本数据本身的流形学结构,并且没有使用样本间属性的相关性。为此,提出引入稀疏学习理论,利用训练样本重构测试样本的方法,重构过程使用了样本间的相关性,也用到局部保持投影 LPP 保持数据结构不变,同时引入 $l_{2,1}$ 范数用于去除噪声样本的方法来寻找投影变换矩阵 $\boldsymbol{W}$,进而利用 $\boldsymbol{W}$ 确定 KNN 算法中 K 值的 SA-KNN 算法。在 UCI 数据集上的仿真实验结果表明,该方法比传统的 KNN 分类算法和 Entropy-KNN 算法有更高的分类准确度。

关键词:K 近邻分类;相关性;去除噪声样本;局部保持投影;稀疏学习

中图分类号:TP181　　**文献标志码**:A

doi:10.3969/j.issn.1007-130X.2015.10.025

An efficient SA-KNN algorithm with adaptive K value

SUN Ke[1,3], GONG Yong-hong[1,2], DENG Zhen-yun[1,3]

(1. Guangxi Key Laboratory of Multi-Source Information Mining & Security, Guangxi Normal University, Guilin 541004;
2. Guilin University of Aerospace Technology, Guilin 541004;
3. College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China)

Abstract: Traditional K Nearest Neighbors (KNN) classification method has drawbacks such as no elimination of noise samples, no manifold structure preservation of the samples, and no consideration of the correlation between samples. To solve these problems, we propose an efficient SA-KNN algorithm with adaptive K value. Sparse learning theory is introduced and we reconstruct each test sample with the training samples for KNN classification. We introduce an $l_{2,1}$ norm to remove the noisy samples, employ the Locality Preserving Projections (LPP) to keep the data structures, and makes the best use of the correlation between the samples in the reconstruction process. With these technologies we can get the transformation matrix $\boldsymbol{W}$ and in turn determine the value of K. Simulation results on the UCI data sets demonstrate a better classification accuracy than the traditional KNN and the Entropy-KNN method.

Key words: K nearest neighbor (KNN) classification; correlation; removal of noise samples; locality preserving projection; sparse learning

* 收稿日期:2014-09-22;修回日期:2014-11-26
基金项目:国家自然科学基金资助项目(61170131 和 61263035);国家 863 计划资助项目(2012AA011005);国家 973 计划资助项目(2013CB329404);广西自然科学基金资助项目(2012GXNSFGA060004);广西八桂创新团队和广西百人计划资助;广西研究生教育创新计划项目(YCSZ2015095,YCSZ2015096)
通信作者:龚永红(1210385063@qq.com)
通信地址:541004 广西桂林市七星区育才路 15 号广西师范大学计算机科学与信息工程学院
Address: College of Computer Science and Information Technology, Guangxi Normal University, 15 Yucai Rd, Qixing District, Guilin 541004, Guangxi, P. R. China

1 引言

分类[1]是数据挖掘和机器学习等领域[2]分析数据的一种重要手段。分类是一种有监督的学习方法,通过对已知有属性描述的样本进行训练,形成一个分类器或者分类模型,然后利用所得到的分类模型,将未知类别的数据映射到相应的类空间中,然后得到未知类别样本所属的类。而基于实例的惰性分类方法——KNN(K Nearest Neighbors)分类方法,因为其简单、易于操作且性能优越的特点,在机器学习、数据挖掘等领域得到了广泛应用,是公认的数据挖掘"十大经典算法"之一[3]。

传统的KNN分类算法在利用样本学习的过程中存在一些缺陷,例如K值的选取存在偏好和不确定性的问题,通常情况下K值是由用户给定或者利用十折交叉法得到的,并不是通过学习样本本身特点确定的,这使得得到的K值存在一定的不合理性。而且,KNN分类算法在使用已知属性样本形成分类器时,只是孤立地使用各个样本数据,并没有考虑到样本之间的相关性[4]。事实上,样本之间存在相关性,如果分类模型能利用这种相关性,就能够获得更好的学习效果,提高分类性能[5,6]。并且在利用样本进行投影变换时,传统KNN算法没有考虑到样本的流形学结构,局部保持投影LPP(Locality Preserving Projections)[7]认为,保持算法前后数据自身具有的流形学结构,能保留样本更多的信息,通常可以取得更显著的学习效果。流形学结构认为传统的欧氏空间难以度量真实世界的非线性数据和数据的结构分布,所谓流形(Manifold)是局部具有欧氏空间性质的空间,包括各种维度的曲线曲面,流形的局部和欧氏空间是同构的,流形学习假设所处理的数据点分布在嵌入于外维欧氏空间中的一个潜在的流形体上[7]。另外,在学习过程中传统KNN分类算法没有考虑到清除噪声样本的问题,但是实际的数据是存在着大量噪声的[8,9],如果我们在学习分类器的过程中没有清除这些噪声样本,势必会影响分类器模型形成规则的准确性。所以,清除噪声样本避免他们对分类器规则形成的影响是必要的[10~12]。

为此,本文根据稀疏学习理论,利用训练样本重构测试样本的方法寻找投影变换矩阵W,然后利用得到的W确定测试样本分类所需的K值,进而进行最近邻分类。该过程中考虑样本之间存在的关系,充分利用样本间的相关性;考虑保持数据结构不变,保留更多的样本信息使用LPP算法;考虑到噪声样本会对分类器产生的影响,应用$l_{2,1}$范式的稀疏性来去除噪声样本。我们定义这种利用上述技术、通过学习得到K值的最近邻方法为Self-Adaption KNN,简称SA-KNN方法。

2 相关理论背景LPP算法简介

2.1 LPP算法简介

LPP又叫局部保持投影,是非线性方法,它的目标是保证在高维空间的原始数据所存在的相邻关系,在投影后的较低维的空间上也保持相应的相邻关系。它是通过寻找一个投影变换矩阵W,来实现上述的样本在新的空间上的相邻的关系的,同时它又能保持原始数据非线性的流形特征。可以说,它是一种能较好地保持非线性流形中局部数据特征的线性流形学习算法[7]。

假设输入样本为:$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_n\}$,其中$\boldsymbol{x}_i\in\mathbf{R}^d$,$\boldsymbol{x}_i$代表$d$维空间里的一个点;其在低维空间$\mathbf{R}^c$上的投影变换的样本为:$\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_i,\cdots,\boldsymbol{y}_n\}$,$\boldsymbol{y}_i\in\mathbf{R}^c$,其中$c<<d$,这样,通过投影变换矩阵就可以将高维的空间投影到低维的空间上。

投影矩阵为:$\boldsymbol{W}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_i,\cdots,\boldsymbol{w}_d\}$,其中$\boldsymbol{w}_i\in\mathbf{R}^c$表示低维空间,彼此之间独立且不为零。

变换矩阵$\boldsymbol{W}$求解通常使用最小化下面的目标函数:

$$\min\sum_{ij}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_j)^2S_{ij} \tag{1}$$

其中,$\boldsymbol{S}$为权值矩阵,它的取值方法有两种,可以简单地取值为1或者为0,也可以使用下式确定$\boldsymbol{S}$中每个元素:

$$S_{ij}=\mathrm{e}^{\frac{\|x_i-x_j\|^2}{t}}$$

其中,t是一个大于零的常量。

对式(1)进行变换,其中令$\boldsymbol{y}_i=\boldsymbol{W}\boldsymbol{x}_i$,则有:

$$\sum_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2S_{ij}=\sum_{ij}\boldsymbol{y}_i^2S_{ij}-2\sum_{ij}\boldsymbol{y}_i\boldsymbol{y}_iS_{ij}+\sum_{ij}\boldsymbol{y}_i^2S_{ij}=2\sum_i\boldsymbol{y}_i^2\boldsymbol{D}_{ii}-2\sum_{ij}\boldsymbol{y}_i\boldsymbol{y}_iS_{ij}=2\boldsymbol{y}(\boldsymbol{D}-\boldsymbol{S})\boldsymbol{y}^{\mathrm{T}}=2\boldsymbol{y}\boldsymbol{L}\boldsymbol{y}^{\mathrm{T}} \tag{2}$$

其中,$\boldsymbol{D}_{ii}=\sum_jS_{ij}$为对角阵,$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$,叫做拉普拉斯矩阵。

2.2 稀疏学习理论简介

稀疏学习(Sparse Learning)理论最初用来解

决图像视觉,表现出了很强的内在理论价值和技术潜力,目前稀疏学习理论[13]发展迅速,已经在机器学习、模式识别领域得到广泛应用。

在机器学习中,通过对模型参数向量 $w \in \mathbf{R}^n$ 进行稀疏性假设,实现稀疏正则化,使用训练数据对参数 w 进行拟合,这种参数拟合过程往往是通过最小化某个经验风险函数(这个最小化风险函数是平滑的凸函数)和导致稀疏性的正则化项来实现的:

$$\min_w g(w) = f(w) + \rho\ (w) \tag{3}$$

通过调节参数 ρ 可以控制数据拟合项和稀疏正则化项之间的平衡,改变 w 的稀疏性。

稀疏学习理论将样本之间的系数权重作为鉴别信息引入模型,通过对输入数据使用稀疏约束,使之变得稀疏,这样数据中一些无关项就会变为零,而主要信息得到保存,所以对噪声样本存在很强的鲁棒性。而且稀疏学习的正则化因子选取范数的优化问题是一个凸优化,能保证得到唯一的全局最优解[14,15]。

稀疏学习模型通常有回归和分类两种应用,本文使用了分类技术,而回归即指优化样本空间中类标签和条件属性之间的关系[16]。重构技术可以充分利用样本间的属性关系,重构是指优化测试样本和训练样本之间的关系。本文正是利用重构的方法学习出测试样本和训练样本之间的关系矩阵,并以此学习出分类所需的 K 值。

目标函数(3)中 $f(w)$为损失函数,常用的损失函数有绝对损失函数、对数损失函数、最小二乘损失函数等,(w) 是正则化项,常用的包括 l_1 范数、$l_{2,1}$ 范数和 l_F 范数等。

3　SA-KNN 算法描述和优化方法

3.1　算法描述

假设给定训练集 $\boldsymbol{X} \in \mathbf{R}^{n\times d}$ 和测试集 $\boldsymbol{Y} \in \mathbf{R}^{m\times d}$,其中,$d$ 是样本维数,n、m 是样本数量。本文希望寻找到一个投影变换矩阵 $\boldsymbol{W} \in \mathbf{R}^{n\times m}$,通过 $\boldsymbol{W}$ 确定分类所需的 K 值,显然,损失函数模型选择最小二乘损失模型表示更加合理,即:

$$\min_{\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\|_F^2 \tag{4}$$

考虑到式(4)是一个凸函数,它的解可表示为:$\boldsymbol{W}^* = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$,但是 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$ 存在不可逆的问题,通常情况下是考虑利用岭回归引入一个 l_2 范数来解决[17]:

$$\min_{\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\|_F^2 + \rho\|\boldsymbol{W}\|_2^2 \tag{5}$$

此时公式(5)的解为:$\boldsymbol{W}^* = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \rho\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$,然而 $\boldsymbol{W}$ 是实数矩阵,没有稀疏项,应用到 KNN 分类中时,K 通常取样本数目 n,这显然是不合理的。因为,正如本文引言提到的数据中通常存在的噪声样本的问题,$K=n$ 就是选取了所有的样本数据,并没有将噪声样本去除。为此,本文将基于岭回归的 l_2 范数改为具有稀疏性的 $l_{2,1}$ 范数,这样就可以利用稀疏学习的原理使得噪声样本变为零,去除非相关性噪声样本。同时,引入局部保持投影(LPP)算法,保持样本数据在空间投影变换过程中的流形学结构不变。利用测试样本 $\boldsymbol{X}$ 的条件属性重构测试样本 $\boldsymbol{Y}$,寻找 $\boldsymbol{Y}$ 与 $\boldsymbol{X}$ 之间的相关性函数关系,即得到相关性矩阵 $\boldsymbol{W}$,这个 $\boldsymbol{W}$ 矩阵就代表了测试样本和训练样本之间的相关关系,$\boldsymbol{W}$ 矩阵相应位置上值的大小反映了相关程度的大小,这样就利用了样本之间的相关性。故我们所用的模型如下:

$$\min_{\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\|_F^2 + \rho_1\|\boldsymbol{W}\|_{2,1} + \rho_2 * \mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}) \tag{6}$$

其中,$\boldsymbol{X}$ 为训练样本,$\boldsymbol{Y}$ 为测试样本,$\boldsymbol{L}$ 是如公式(2)所描述的拉氏矩阵,ρ_1 和 ρ_2 是两个调整参数。

我们的目标是把 $\boldsymbol{X}$ 投影到 $\boldsymbol{Y}$ 的空间中去寻找 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的关系,也就是要找到一个优化矩阵 $\boldsymbol{W}$ 使得 $\boldsymbol{Y}$ 与 $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$ 尽可能接近。为了提高分类的效果,根据流形学原理[7],$\boldsymbol{X}$ 的局部结构显然需要在新的空间中得到保持。根据公式(1)可知,此正则化项为 $\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W})$,$\rho_2$ 调控此正则化项,具体来说,ρ_2 是用来控制 LPP 部分的数量级和最小二乘损失模型部分的数量级保持一致的,这样可以保持样本 $\boldsymbol{X}$ 在新的空间的流形学结构不变。

模型中,ρ_1 控制 $\boldsymbol{W}$ 矩阵行的稀疏性,即它的值越大时 $\boldsymbol{W}$ 中行为零的数量增加,值越小时行为零的数量减少。通过合适的 ρ_1 产生合适的 $\boldsymbol{W}$ 矩阵,这样使得噪声样本对应的 $\boldsymbol{W}$ 的行稀疏,即使得这些噪声样本为零。举例说明如下,假设通过模型得到一个 5×4 的矩阵:

$$\boldsymbol{W} = \begin{pmatrix} 0.2 & 0.1 & 0.3 & 0.8 \\ 0 & 0 & 0 & 0 \\ 0.1 & 0.5 & 0.7 & 0.6 \\ 0.6 & 0.4 & 0.8 & 0.9 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

这里行代表的是训练样本,列代表的是测试样本,而其中的值代表的是训练样本和测试样本的相

关程度,值越大相关程度越高,重构过程就要使用这些样本。例如对于训练样本,我们观察到测试样本1、3、4行是不为零的,也就是说这些测试样本和训练样本存在着相关性,重构过程就要选取第1个、第3个和第4个测试样本。

同时,利用W的稀疏性产生行为零的稀疏结构,这里为零的行对应的训练样本代表的是噪声样本,这些样本与其他测试样本之间不存在相关关系,重构时,利用W为零的行与之相乘来忽略掉这些样本,并利用余下的样本进行重构。根据以上叙述的规则,此时可以确定K的值为3,即不为零的行数,这样就完成了K值根据样本特性自动选定的功能。已有的K值选取方法由用户选取固定的值,没有考虑到数据的特点,十折交叉验证法也没考虑样本间的相关性和样本的局部结构。本文算法的K是通过学习得到的,学习的过程中考虑了样本间的相关性和局部结构,而且去除了噪声的影响。

程序的伪代码如下:

输入:训练集、测试集;

输出:分类准确率。

(1)依据所选择的模型:

$\min\limits_{W}\|\boldsymbol{Y}-\boldsymbol{XW}\|_F^2+\rho_1\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{XLX}^{\mathrm{T}}\boldsymbol{W})+\rho_2\|\boldsymbol{W}\|_{2,1}$求解优化问题得到投影变换矩阵$W$。

(2)利用W确定K值,进而形成SA-KNN分类器。

(3)将得到的分类器应用到测试集,将所属类数量最多的类标签作为测试集所属类,并计算分类正确率。

3.2 优化分析求解

虽然我们所选用的模型是一个凸函数,但后面两项都是非光滑的,无法直接求得解析解,为此,本文提出一种有效的优化算法来求解目标函数[18,19]。

具体说,首先对$w_i(1\leqslant i\leqslant m)$求导并令其为0,可得:

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{w}_i-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}_i+\rho_1\boldsymbol{L}\boldsymbol{w}_i+\rho_2\widetilde{\boldsymbol{D}}_i\boldsymbol{w}_i=0 \quad (7)$$

这里$\widetilde{D}_i$也是对角矩阵,第k个对角元素为$\frac{1}{2\|w_k\|_2}$。所以有下式:

$$\boldsymbol{w}_i=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\rho_1 L+\rho_2\widetilde{\boldsymbol{D}})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}_i \quad (8)$$

注意到$\widetilde{D}$依赖于W,因此它也是未知的。本文接下来提出一种迭代算法去求解最优值W,即下面的算法1。

算法1 目标函数优化算法

输入:X、Y;

初始化$W^{(t)}\in\mathbf{R}^{n\times m}$,$t=1$;

do{

(1)计算对角矩阵$\widetilde{D}^{(t)}$,这里$\widetilde{D}^{(t)}$第k个对角元素为$\frac{1}{2\|(w^{(t)})^k\|_2}$;

(2)For 每个 i($1\leqslant i\leqslant m$),

$w_i^{(t+1)}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\rho_1\boldsymbol{L}+\rho_2\widetilde{\boldsymbol{D}}^{(t)})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}_i$;

(3)$t=t+1$;

}until 收敛

输出:$W^{(t)}\in\mathbf{R}^{n\times m}$。

定理1 算法1在每次迭代中目标值减小。

证明 根据算法里的第(2)步可得到:

$$\boldsymbol{W}^{(t+1)}=\min_{W}\mathrm{tr}(\boldsymbol{Y}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X})+\rho_1\boldsymbol{L}+\rho_2 Tr(\boldsymbol{W}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{(t)}\boldsymbol{W}) \quad (9)$$

因此,可以将公式(9)做如下变化:

$$\mathrm{tr}(\boldsymbol{Y}-(\boldsymbol{W}^{(t+1)})^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}(\boldsymbol{Y}-(\boldsymbol{W}^{(t+1)})^{\mathrm{T}}\boldsymbol{X})+\rho_1\boldsymbol{L}+\rho_2\ \mathrm{tr}((\boldsymbol{W}^{(t+1)})^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{t}\boldsymbol{W}^{(t+1)})\leqslant\mathrm{tr}(\boldsymbol{Y}-(\boldsymbol{W}^{(t)})^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}(\boldsymbol{Y}-(\boldsymbol{W}^{(t)})^{\mathrm{T}}\boldsymbol{X})+\rho_1\boldsymbol{L}+\rho_2\,\mathrm{tr}(\boldsymbol{W}^{(t)})^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{t}\boldsymbol{W}^{(t)}\Rightarrow$$

$$\mathrm{tr}(\boldsymbol{Y}-(\boldsymbol{W}^{(t+1)})^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}(\boldsymbol{Y}-(\boldsymbol{W}^{(t+1)})^{\mathrm{T}}\boldsymbol{X})+\rho_1\boldsymbol{L}+\rho_2\sum_{k=1}^{d}\left(\frac{\|(\boldsymbol{w}^{(t+1)})^k\|_2^2}{2\|(\boldsymbol{w}^{(t)})^k\|_2}-\|(\boldsymbol{w}^{(t+1)})^k\|_2+\|(\boldsymbol{w}^{(t+1)})^k\|_2\right)\leqslant\mathrm{tr}(\boldsymbol{Y}-(\boldsymbol{W}^{(t)})^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}(\boldsymbol{Y}-(\boldsymbol{W}^{(t)})^{\mathrm{T}}\boldsymbol{X})+\rho_1\boldsymbol{L}+\rho_2\sum_{k=1}^{d}\left(\|(\boldsymbol{w}^{(t)})^k\|_2+\frac{\|(\boldsymbol{w}^{(t)})^k\|_2^2}{2\|(\boldsymbol{w}^{(t)})^k\|_2}-\|(\boldsymbol{w}^{(t)})^k\|_2\right)\Rightarrow\mathrm{tr}(\boldsymbol{Y}-(\boldsymbol{W}^{(t+1)})^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}(\boldsymbol{Y}-(\boldsymbol{W}^{(t+1)})^{\mathrm{T}}\boldsymbol{X})+\rho_1\boldsymbol{L}+\rho_2\sum_{k=1}^{d}\|(\boldsymbol{w}^{(t+1)})^k\|_2\leqslant\mathrm{tr}(\boldsymbol{Y}-(\boldsymbol{W}^{(t)})^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}(\boldsymbol{Y}-(\boldsymbol{W}^{(t)})^{\mathrm{T}}\boldsymbol{X})+\rho_1\boldsymbol{L}+\rho_2\sum_{k=1}^{d}\|(\boldsymbol{w}^{(t)})^k\|_2$$

根据文献[20]可知,对于任意向量w和w_0,我们有:

$$\|\boldsymbol{w}\|_2-\frac{\|\boldsymbol{w}\|_2^2}{2\|\boldsymbol{w}_0\|_2}\leqslant\|\boldsymbol{w}_0\|_2-\frac{\|\boldsymbol{w}_0\|_2^2}{2\|\boldsymbol{w}_0\|_2}$$

因此,最后一步成立,即算法在每次迭代过程中都减小目标值。□

$W^{(t)}$和$\widetilde{D}^{(t)}$在收敛处满足等式(9)。但是,由于公式(7)是一个凸优化问题,满足等式(8)意味着W对于公式(7)来说是一个全局最优解。因此算法1将收敛到公式(7)的全局最优解。因为,我们在每一次迭代运算时都有封闭形式的解,所以我们的算法收敛非常快。

4 实验结果分析

4.1 实验、数据集和评价指标介绍

本文选用的数据集全部来自于UCI中有关分

类的数据集,考虑到选取数据集的一般性,我们选取两个二类数据集和两个多类数据集,并对数据做了必要的修改,例如将样本进行正规化处理。实验采用 Matlab 2010b 软件,在 PC 机上进行编程操作。数据集信息统计如表 1 所示,为叙述方便,用缩写代表某些数据集。其中 BT 为数据集 Blood Transfusion 的缩写,CMC 为数据集 Contraceptive Method Choice 的缩写,GI 为数据集 Glass Identification 的缩写。

Table 1　Data set information statistics

表 1　数据集信息统计

名称	BT	German	CMC	GI
数据类型	多变量	多变量	多变量	多变量
属性类型	实数	实数	实数	实数
样本个数	740	1000	790	214
属性个数	5	25	10	11
类别数	2	2	3	6
来源	UCI	UCI	UCI	UCI

为体现 SA-KNN 算法的优越性能,我们选用 KNN 算法和基于属性值信息熵的 Entropy-KNN 算法[21]作比较,本实验采用的评价指标是分类的准确率,即分类正确样本占总样本的比例,正确率越高表明分类的效果越好。

4.2　实验结果和分析

本文将每组数据同时使用传统的 KNN 算法、Entropy-KNN 算法和 SA-KNN 算法进行比较,在相同的条件下比较三种算法的分类准确率,为使所得结果更加准确,我们采用十折交叉法,每次实验得到 10 组不同的分类准确率结果,并将这些结果统计绘制成图表。

四个数据集使用三种算法所得的分类准确率的比较结果图如 1～图 4 所示。

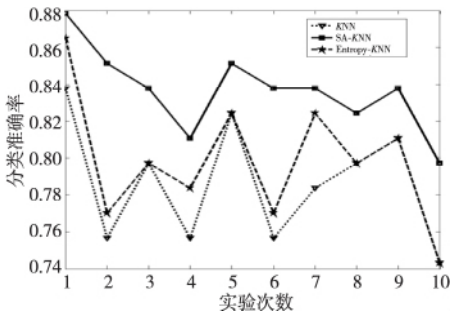

Figure 1　Classification accuracy comparison of chart of BT among the three algorithms

图 1 数据集 BT 分类准确率比较图

从图 1～图 4 的统计结果可以看出,SA-KNN 算法因为在重构过程中充分使用了样本间的相关性,进行样本空间变换时使用局部保持投影算法,利用 $l_{2,1}$ 范数正则化项去除噪声样本,所以在相同的实验条件下比传统的 KNN 算法和基于属性信息熵的 Entropy-KNN 算法有更高的分类准确率。

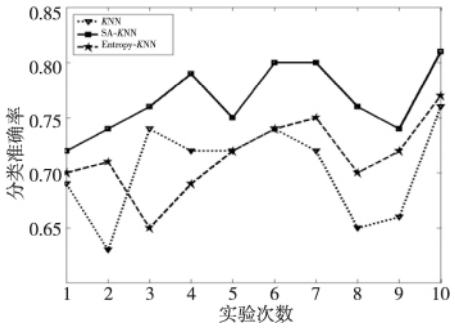

Figure 2　Classification accuracy comparison of chart of German among the three algorithms

图 2　数据集 German 分类准确率比较图

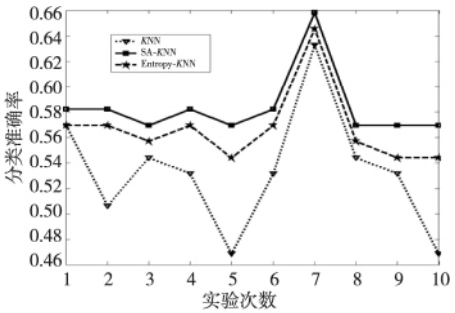

Figure 3　Classification accuracy comparison of chart of CMC among the three algorithms

图 3　数据集 CMC 分类准确率比较图

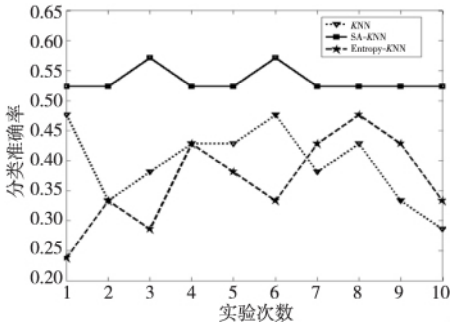

Figure 4　Classification accuracy comparison of chart of GI among the three algorithms

图 4　数据集 GI 分类准确率比较图

数据集分类准确率的均值和方差统计结果如表 2 所示。

从表 2 的统计结果可以看出,SA-KNN 算法不仅在分类准确率的统计上优于传统的 KNN 分类算法以及基于属性信息熵的 Entropy-KNN 算法,而且 SA-KNN 分类算法的方差更小,这表明 SA-KNN 算法的性能更加稳定。

Table 2 Statistical results of the mean and variance of accuracy

表 2 准确率的均值和方差统计结果

名称	SA-*K*NN	*K*NN	Entropy-*K*NN
BT	0.8365±5.1e−4	0.7865±1.0e−3	0.7986±1.2e−3
German	0.7670±9.6e−4	0.7030±1.9e−3	0.7150±1.1e−3
GI	0.5333±4.0e−4	0.3667±5.6e−3	0.3952±4.1e−3
CMC	0.5835±7.3e−4	0.5329±2.3e−3	0.5671±8.8e−4

5 结束语

本文提出了一种基于稀疏学习理论,利用训练样本重构测试样本的 SA-*K*NN 分类方法,在此重构过程中考虑到噪声样本影响,使用具有稀疏性的 $l_{2,1}$ 范式正则化因子来去除噪声样本;考虑到样本之间的关系,充分利用样本间相关关系;考虑到进行样本空间变换时样本流形学结构需要保持的问题,引入局部保持投影(LPP)算法。我们使用此重构技术得到具有稀疏结构的转换矩阵 W,利用 W 可以得到最近邻分类所需的 K 值,与传统的 *K*NN 分类方法以及基于属性值信息熵的 Entropy-*K*NN 算法相比,该方法在使用上述技术的情况下可以通过学习样本自身特点,自动地确定合适的 K 值。实验结果表明,SA-*K*NN 分类方法不仅比传统的 *K*NN 分类方法和基于属性信息熵的 Entropy-*K*NN 算法具有更高的分类准确率,而且分类准确率更稳定。

参考文献:

[1] Han Jia-wei, Kambei M, Pei Jian. Data mining concepts and techniques[M]. California: Morgan Kaufmann, 2011.

[2] Zhang Shi-chao, Zhang Cheng-qi, Yan Xiao-wei. Post-mining: Maintenance of association rules by weighting[J]. Information Systems, 2003, 28(7): 691-707.

[3] Wu X, Kumar V, Quinlan J R, et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems, 2008, 14(1): 1-37.

[4] Wu Xin-dong, Zhang Cheng-qi, Zhang Shi-chao. Efficient mining of both positive and negative association rules[J]. ACM Transactions on Information Systems, 2004, 22(3): 381-405.

[5] Zhu Xiao-feng, Huang Zi, Shen Heng-tao, et al. Dimensionality reduction by mied kernel canonical correlation analysis [J]. Pattern Recognition, 2012, 45(8): 3003-3016.

[6] Wu Xin-dong, Zhang Shi-chao. Synthesizing high frequency rules from different data sources[J]. IEEE Transactions on Knowledge Data Engineering 2003, 15(2): 353-367.

[7] He Xiao-fei, Niyogi P. Locality preserving projections[C]// Proc of NIPS, 2003: 5-25.

[8] Zhu Xiao-feng, Huang Zi, Cheng Hong, et al. Sparse hashing for fast multimedia search[J]. ACM Transactions on Information Systems, 2013, 31(2): 9.

[9] Zhang Shi-chao, Qin Zhen-xing, Charles X. et al. "Missing Is Useful": Missing values in cost-sensitive decision trees[J]. IEEE Transactions on Knowledge Data Engineering, 2005, 17(12): 1689-1693.

[10] Zhu Xiao-feng, Zhang Shi-chao, Jin Zhi, et al. Missing value estimation for mixed-attribute data sets[J]. IEEE Transactions on Knowledge Data Engineering, 2011, 23(1): 110-121.

[11] Zhu Xiao-feng, Huang Zi, Yang Yang, et al. Self-taught dimensionality reduction on the high-dimensional small-sized data[J]. Pattern Recognition, 2013, 46(1): 215-229.

[12] Qin Yong-song, Zhang Shi-chao, Zhu Xiao-feng, et al. Semi-parametric optimization for missing data imputation [J]. Applied Intelligence, 2007, 27(1): 79-88.

[13] Donoho D L. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, 52: 1289-1306.

[14] Baraniuk R G. Compressive sensing[J]. Lecture Notes in IEEE Signal Processing Magazine, 2007, 24: 118-120.

[15] Zhu Xiao-feng, Huang Zi, Cui Jiang-tao, et al. Video-to-shot tag propagation by graph sparse group lasso [J]. IEEE Transactions on Multimedia, 2013, 15(3): 633-646.

[16] Zhu Xiao-feng, Suk Heung-Il, Shen Ding-gang. Matrix-similarity based loss function and feature selection for alzheimer's disease diagnosis[C]// Proc of CVPR, 2014: 3089-3096.

[17] Trevor H, Robert T, Jerome F-RIDMAN. The elements of statistical learning: Data minning, inference, and prediction [M]. New York: Springer-Verlag, 2009.

[18] Zhu Xiao-feng, Zhang Lei, Huang Zi. A sparse embedding and least variance encoding approach to Hashing[J]. IEEE Transactions on Image Processing, 2014, 23(9): 3737-3750.

[19] Zhang Shi-chao. Nearest neighbor selection for iteratively *k*NN imputation[J]. Journal of Systems and Software, 2012, 85(11): 2541-2552.

[20] Zhu Xiao-feng, Huang Zi, Shen Heng-tao, et al. Linear cross-modal hashing for efficient multimedia search[C]// Proc of ACM Multimedia, 2013: 143-152.

[21] Tong Xian-qun, Zhou Zhong-mei. Enhancement of *K*-nearest neighbor algorithm based on information entropy of attribute value[J]. Computer Engineering and Applications, 2010, 46(3): 115-117. (in Chinese)

附中文参考文献:

[21] 童先群,周忠眉.基于属性值信息熵的 *K*NN 改进算法[J].计算机工程与应用,2010,46(3):115-117.

作者简介:

孙可(1987-),男,河南永城人,硕士生,研究方向为数据挖掘和机器学习。**E-mail**:723626584@qq.com

SUN Ke, born in 1987, MS candidate, his research interests include data mining, and machine learning.