第 34 卷 第 1 期
2016 年 3 月
广西师范大学学报(自然科学版)
Journal of Guangxi Normal University(Natural Science Edition)
Vol. 34 No. 1
Mar. 2016

doi:10.16088/j.issn.1001-6600.2016.01.008

基于局部相关性的 kNN 分类算法

邓振云[1],龚永红[1,2],孙 可[1],张继连[1]

(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)

摘 要:kNN 算法作为一种简单、有效的分类算法,在文本分类中得到广泛的应用。但是在 k 值(通常是固定的)的选取问题上通常是人为设定。为此,本文引入了重构和局部保持投影(locality preserving projections,LPP)技术用于最近邻分类,使得 k 值的选取是由样本间的相关性和拓扑结构决定。该算法利用 l_1-范数稀疏编码方法使每个测试样本都由它的 k(不固定)个最近邻样本来重构,同时通过 LPP 保持重构前后样本间的局部结构不变,不仅解决了 k 值的选取问题,并且避免了固定 k 值对分类的影响。实验结果表明,该方法的分类性能优于经典 kNN 算法。

关键词:kNN;保局投影;重构;稀疏编码

中图分类号:TP181 **文献标志码**:A **文章编号**:1001-6600(2016)01-0052-07

0 引言

在数据挖掘的研究与应用中,分类是用于预测分析的重要方式之一。它是一种有监督的学习,通过对训练样本的分析,从中归纳出分类规则,以此来预测测试样本的类别。目前常见的分类算法主要有:决策树、关联规则、贝叶斯、神经网络、遗传算法、kNN 算法等。本文集中在 kNN 分类算法的研究。

由于 kNN 算法在样本较大以及特征属性较多时,分类的效率就将大大降低,因此近年来学者们提出了许多针对 kNN 的改进算法。如文献[1]提出将粗糙集理论应用到 kNN 算法中,实现属性约简,解决了 kNN 算法分类效率低的缺点;文献[2]提出一种基于密度的样本裁剪方法,降低 kNN 的计算量,提高分类的性能;文献[3]提出一种基于类别的 kNN 改进模型,解决 k 近邻选择时大类别、高密度样本占优问题等;文献[4]提出一种充分利用已知类别标签数据进行自训练的半监督分类算法,显著地提高了分类的准确率。这些算法主要通过一定的优化策略或降维方法减少样本之间相关性的计算,以提高分类的效率。而本文直接考虑样本间的相关性、样本的拓扑结构,对于 kNN 算法中 k 值的选取直接由数据本身驱动,避免了人为设定 k 值对分类的影响。

kNN 算法是一种基于实例的学习方法[5],最初用于解决文本分类的问题。其基本思想是:在训练样本中找到待测样本的 k(定值)[6-7]个最近邻样本,然后根据这 k 个最近邻样本的类别进行投票,以此来决定测试样本的类别。但是在实际应用中,这种采取固定 k 值的方法经常是不合理的,如图 1。

当 $k=5$ 时,根据 kNN 分类算法将得到两个样本空间(图 1 中两个实线圆圈)。其中待测样本 1 的 k 个最近邻样本选取合理,但对于待测样本 2 而言,实际上令 $k=3$ 更为合理(如虚线圆圈所示),因为下方两个训练样本距离待测样本 2 较远,如果把它们也作为最近邻样本,很可能会影响分类准确率。因此本文也

收稿日期:2015-06-16

基金项目:国家自然科学基金资助项目(61573270,61263035,61363009);国家 973 计划项目(2013CB329404);广西自然科学基金资助项目(2012GXNSFGA060004,2015GXNSFCB139011);中国博士后科学基金资助项目(2015M570837);广西多源信息挖掘与安全重点实验室开放基金资助项目(MIMS13-08)

通信联系人:张继连(1977—),男,广西桂林人,广西师范大学教授,博士。E-mail:jiliangzhang@sina.com;龚永红(1970—),女,广西永福人,桂林航天工业学院副教授。E-mail:zysjd2015@163.com

认为:对于测试样本取相同的 k 值是不合理的[8];k 值的选取应该由数据的分布或特点决定,即 k 值是从数据中学习的,对不同的测试样本它的取值应该是不固定的。为了更加准确地根据数据分布或特点学习 k 值,应当充分考虑数据中的先验知识,例如样本间的相关性、样本的拓扑结构等。

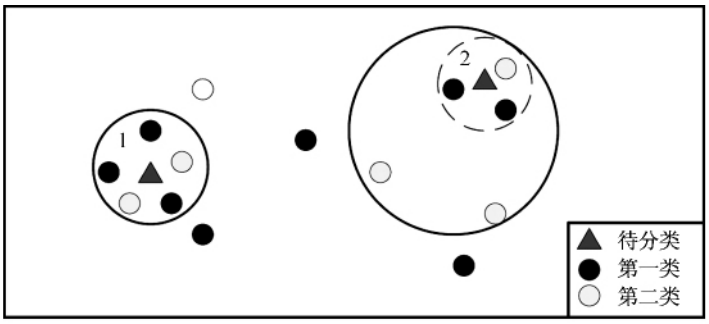

图 1　$k=5$ 时测试样本分类情况

Fig. 1　An example of kNN classification task with $k=5$

根据以上理论,本文用训练样本重构所有测试样本生成相关性矩阵,并通过 l_1-范数稀疏编码方法[9-11]对矩阵进行重构[12],使得每个测试样本都由它的 k 个相关的最近邻样本预测,同时通过 LPP 保持了样本的局部结构而不受重构影响。这样便使得每个测试样本都由它的 k(不固定)个相关的最近邻样本预测,分类的性能得到很大的提高。本文将提出的算法简称为 LS-kNN(LPP-Sparse-kNN)算法。

1　LS-kNN 算法

1.1　局部保持投影(LPP)

LPP 是非线性方法拉普拉斯特征映射(Laplacian eigenmap)的线性近似,其特点是获得原始数据在低维度上的投影,且保持数据的非线性流形特征不变。LPP 算法的基础是构造一个模拟局部结构的最近邻图,本算法应用此特性即构造一个模拟测试样本的最近邻的矩阵 $\boldsymbol{W}$。通常采用最小化目标函数式(1)来确定这个最近邻矩阵 $\boldsymbol{W}$:

$$\min_{\boldsymbol{W}} \sum_{i,j} (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_j)^2 S_{ij}。 \tag{1}$$

其中:$\boldsymbol{x}_i$ 为训练集 $\boldsymbol{X}=\{\boldsymbol{x}_i\}_{i=1}^{l}\in \mathbf{R}^{D\times L}$ 中的训练样本,$S_{ij}=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/\sigma)$ 是一个由热核估计定义的权值矩阵。D 是训练样本的维数,S_{ij} 代表 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$ 两个点之间的近邻程度。

对公式(1)进行代数变化操作后发现,特征空间保持了原始的局部结构。操作如下:

$$\frac{1}{2}\sum_{i,j}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_j)^2 S_{ij} = \sum_i \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i d_{ij}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W} - \sum_{i,j}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}S_{ij} =$$
$$\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}) - \mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{S}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}) = \mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W})。 \tag{2}$$

$\boldsymbol{D}=[d_{ij}]$ 表示一个对角矩阵,$\boldsymbol{D}$ 中第 i 个元素被计算为 $\boldsymbol{S}$ 的第 i 列的总和,即 $d_{ii}=\sum_j S_{ij}$。显然,$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$ 是一个拉普拉斯矩阵。

1.2　重构

假设训练集 $\boldsymbol{X}\in\mathbf{R}^{n\times d}$,测试集 $\boldsymbol{Y}\in\mathbf{R}^{m\times d}$,其中 m、n、d 分别是测试样本的个数、训练样本个数和样本的维数。本文算法的核心是通过训练样本去重构测试样本,得到一个相关性矩阵 $\boldsymbol{W}\in\mathbf{R}^{n\times m}$,使得 $\boldsymbol{Y}$ 与 $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$ 尽可能差距最少。这可以通过最小二乘损失函数[13] 实现,即:

$$\min_{\boldsymbol{W}} \|\boldsymbol{Y}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\|_F^2。 \tag{3}$$

其中 $\|\boldsymbol{Y}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\|_F^2=\mathrm{tr}(\boldsymbol{Y}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X})$ 为 Frobenius 范数,$\boldsymbol{W}$ 表示 $\boldsymbol{Y}$ 与 $\boldsymbol{X}$ 之间的相关性,如 w_{ij} 即表示第 i 个训练样本与第 j 个测试样本之间的相关性大小。若 $w_{ij}>0$,表示 $\boldsymbol{X}$ 的第 i 个训练样本与 $\boldsymbol{Y}$ 的第 j 个测试样本是正相关;若 $w_{ij}=0$,则表示它们不相关;若 $w_{ij}<0$,则表示它们负相关。把这个特性应用到

kNN 分类算法中,即可找到每个测试样本与之相关的样本,也就是满足 $w_{ij} \neq 0$ 的训练样本。同时,由于每个样本之间的相关性大小都不同,因此重构后每个测试样本 j 对应的 $w_{ij} \neq 0$ 的个数都不一样,正好解决了 kNN 固定 k 值问题,使得每个测试样本的类别都由不固定的最近邻训练样本数预测。

1.3 LS-kNN 算法的描述

由于公式(3)是凸函数,易知其解为 $\boldsymbol{W}^* = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$。但 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 在实际应用中不一定可逆,对此通常会考虑岭回归从而引入一个 l_2 -范数使得函数可逆:

$$\min_{\boldsymbol{W}} \| \boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}}\boldsymbol{X} \|_F^2 + \rho \| \boldsymbol{W} \|_2^2 。 \tag{4}$$

其中 $\| \boldsymbol{W} \|_2^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} | w_{ij} |^2$,$\rho$ 是 l_2 -范数的正则化因子参数,其优化解为 $\boldsymbol{W}^* = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \rho\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$。但是研究已经证明,目标函数(4)得到的 $\boldsymbol{W}$ 不一定稀疏,若利用到 kNN 分类中,即 $k = n$ 选取所有训练样本作为最近邻数,这显然是不合理的。对此,本算法引用了 l_1 -范数[14] 和 LPP 正则化因子替换公式中的 l_2-范数,目标函数(4)转换如下:

$$\min_{\boldsymbol{W}} \frac{1}{2} \| \boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}}\boldsymbol{X} \|_F^2 + \rho_1 \mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}) + \rho_2 \| \boldsymbol{W} \|_1 , \tag{5}$$

其中 $\| \boldsymbol{W} \|_1 = \sum_{i=1}^{m}\sum_{j=1}^{n} | w_{ij} |$,$\rho_2$ 为 l_1 -范数的正则化因子参数,控制着矩阵 $\boldsymbol{W}$ 的稀疏性。因为 l_1 -范数[9] 的基本思想是在回归系数的绝对值之和小于一个常数的约束条件下,使其残差平方和最小化,从而能够产生某些严格等于 0 的系数。通过调试 ρ_2 控制矩阵中 0 的个数,可以很好地控制矩阵的稀疏性,ρ_2 越大,矩阵也就越稀疏性;反之越密集。而保局投影参数 ρ_1 用于调控矩阵 $\boldsymbol{W}$ 的局部结构,使其数量级与最小二乘损失函数 $\| \boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}}\boldsymbol{X} \|_F^2$ 保持一致,若 ρ_1 值越大,则 LPP 所占比重就越大;反之越小。

通过目标函数(5)进行重构将得到如下形式的投影矩阵 $\boldsymbol{W}$:

$$\boldsymbol{W} = \begin{pmatrix} 0.11 & 0.32 & 0 \\ 0 & 0.20 & 0 \\ 0 & 0.40 & 0 \\ 0.73 & 0 & 0.51 \\ 0 & 0.69 & 0 \end{pmatrix} 。$$

根据重构技术可知 $w_{ij} \neq 0$ 即为相关的样本,且第 j 列中 $w_{ij} \neq 0$ 的个数即为第 j 个测试样本最近邻样本的个数。因此从矩阵 $\boldsymbol{W}$ 的形式可判断,第一个测试样本(第一列)的最近邻样本数为 2,即 $k = 2$,第二个测试样本的最近邻数为 4。以此类推,可发现测试样本没有采用固定的最近邻数即没有用固定 k 值去寻找训练样本。

综上可知,LS-kNN 算法通过考虑样本间的相关性,有效地解决了 kNN 分类存在的问题,并通过 LPP 保证了数据内部分布的完整性。注意,l_2 -范数就无法实现这个功能,因为它得到的矩阵不稀疏。

最后,给出 LS-kNN 分类算法的具体步骤。

算法 1:LS-kNN 分类算法。

输入:样本集,并作规范化处理。

输出:测试集的类标签。

1:由 ADMM 优化算法(6)得到最优解 $\boldsymbol{W}$;

2:根据 $\boldsymbol{W}$ 找到每个测试样本对应的 k 个最近邻训练样本;

3:对每个测试样本,用与其对应的 k 个最近邻样本的类标签进行分类投票;

4:输出测试样本的类标签。

2 LS-kNN 算法的优化

由于目标函数(5)是凸且非平滑的,因此本文采用 ADMM(alternating direction method of

multipliers)[15] 算法来优化此函数。目标函数(5) 可等效于求解以下 N 个独立的子问题:

$$\underset{\boldsymbol{w}_i}{\operatorname{argmin}} \frac{1}{2} \| \boldsymbol{y}_i - \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{x}_i \|_F^2 + \rho_1 \operatorname{tr}(\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{x}_i \boldsymbol{L} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w}_i) + \rho_2 \| \boldsymbol{w}_i \|_1 。 \tag{6}$$

其中 $\boldsymbol{w}_i$ 向量是 $\boldsymbol{W}$ 的第 i 个列子向量且 $\operatorname{vec}(\boldsymbol{W}) = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_n]^{\mathrm{T}}$。对目标函数(6) 进行 ADMM 优化,在目标函数(5) 上增加一个虚拟变量(dummy variable)C 使函数转换为:

$$\underset{\boldsymbol{W},\boldsymbol{C}}{\operatorname{argmin}} \frac{1}{2} \| \boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}} \boldsymbol{X} \|_F^2 + \rho_1 \operatorname{tr}(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{L} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}) + \rho_2 \| \boldsymbol{C} \|_1 + \frac{\rho_3}{2} \| \boldsymbol{W} - \boldsymbol{C} \|_F^2, \text{s. t.} \boldsymbol{W} = \boldsymbol{C} 。 \tag{7}$$

公式(7) 较为复杂,因此可以使用扩展拉格朗日进行替代,使其变成如下形式:

$$L(\boldsymbol{W},\boldsymbol{C},\boldsymbol{\Lambda}) = \frac{1}{2} \| \boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}} \boldsymbol{X} \|_F^2 + \rho_1 \operatorname{tr}(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{L} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}) + \rho_2 \| \boldsymbol{C} \|_1 +$$

$$\frac{\rho_3}{2} \| \boldsymbol{W} - \boldsymbol{C} \|_F^2 + \operatorname{vec}(\boldsymbol{\Lambda})^{\mathrm{T}} \operatorname{vec}(\boldsymbol{W} - \boldsymbol{C}) 。 \tag{8}$$

ADMM 算法采用迭代法思想,包括如下的迭代步骤:

$$\begin{aligned} &(\text{ⅰ}) \boldsymbol{W}^{k+1} = \underset{\boldsymbol{W}}{\operatorname{argmin}} L(\boldsymbol{W}, \boldsymbol{C}^k, \boldsymbol{\Lambda}^k); \\ &(\text{ⅱ}) \boldsymbol{C}^{k+1} = \underset{\boldsymbol{C}}{\operatorname{argmin}} L(\boldsymbol{W}^{k+1}, \boldsymbol{C}, \boldsymbol{\Lambda}^k); \\ &(\text{ⅲ}) \boldsymbol{\Lambda}^{k+1} = \boldsymbol{\Lambda}^k + \rho_3 (\boldsymbol{W} + \boldsymbol{C}) 。 \end{aligned} \tag{9}$$

通过 ADMM 算法,可以使目标函数(5) 划分为两个子问题。首先分析子问题 1:如果只最小化目标函数(8) 中的 $\boldsymbol{W}$,那么当 l_1 -范数惩罚 $\| \boldsymbol{C} \|_1$ 使得 $\boldsymbol{W}$ 从目标函数里消失时,这就将子问题 1 转化为一个非常有效且简单的最小二乘问题。然后分析子问题 2:如果只最小化目标函数(8) 中的 $\boldsymbol{C}$,那么当 $\| \boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}} \boldsymbol{X} \|_F^2$ 消失时,允许 $\boldsymbol{C}$ 通过每个元素独立求解。这使本文能够有效地使用软阈值 ρ_2。利用 $\boldsymbol{W}$ 与 $\boldsymbol{C}$ 的当前估计和 ADMM 算法的第 3 步相结合,以此来更新当前估计的拉格朗日乘数矩阵 $\boldsymbol{\Lambda}$。注意,此处引入的惩罚参数 ρ_3 起着特殊的作用,因为它利用一个近似的估计来求解 $\boldsymbol{W}$ 和 $\boldsymbol{C}$。

通过以上分析,本文将目标函数(5) 拆分为 N 个独立的子问题,再使用 ADMM 算法求解最优化的 $\boldsymbol{W}$。

算法 2: ADMM 算法,$\boldsymbol{W}$ 的优化。

Input: 数据集、惩罚函数 ρ_3。

Output: $\boldsymbol{W}$、$\boldsymbol{\Lambda}$。

1: Initialize $\boldsymbol{W}^0, \boldsymbol{C}^0, \boldsymbol{\Lambda}^0$;

2: repeat

3: $\boldsymbol{W}^{k+1} \leftarrow \boldsymbol{W}^k$;

4: $\boldsymbol{C}^{k+1} \leftarrow \boldsymbol{C}^k$;

5: $\boldsymbol{\Lambda}^{k+1} \leftarrow \boldsymbol{\Lambda}^k + \rho_3 (\boldsymbol{W} + \boldsymbol{C})$;

6: $k = k + 1$;

7: Until $\boldsymbol{W}$ 最优。

3　实验与结果分析

为了验证算法的有效性,本文以分类准确率作为评价指标。用 kNN 分类算法、LMNN 算法以及本文的 LS-kNN 改进算法,在 LIBSVM 数据集[16] 上选取 4 个数据集 ionosophere、seeds、heart、cheveland 进行实验(为了使数据特征之间具有可比性,已对数据集进行了标准化处理)。

为了保证 3 个算法比较的公平性,本实验对样本事先采取相同的预处理,并在每个数据集上用十折交叉验证法做 10 次实验来看算法效果。对于 kNN 和 LMNN 算法,我们令 $k = 5$;而 LS-kNN 算法的 k 值是由数据驱动产生,无需事先处理。最后对 3 种算法得到的实验结果进行比较。

实验结果的优劣由分类准确率评定。多次实验得到的分类准确率结果可以反映分类的可信度以及算法的稳定性。分类准确率的均值越高,可信度也就越高,稳定性越强。分类准确率的计算公式如下:

$$\text{accuracy} = \frac{n_{\text{match}}}{n}, \tag{11}$$

其中 n_{match} 为测试样本中正确分类的个数,n 为测试样本的总数。

本文提出的 LS-kNN 算法的目标函数(5)中引入了 ρ_1、ρ_2 两个重要的参数,其中 ρ_1 通过控制 LPP 正则项因子 $\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W})$,使其数量级与损失函数 $\|\boldsymbol{Y}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\|_F^2$ 保持一致;而 ρ_2 为 l_1 -范数的正则化因子参数,控制着矩阵 $\boldsymbol{W}$ 的稀疏性。通过对目标函数(5)进行实验分析,发现 ρ_1、$\rho_2 \in \{10^{-5},\cdots,10^{-1}\}$ 能使得矩阵重构 $\boldsymbol{W}$ 产生较好的稀疏结构,因此我们对 ρ_1、ρ_2 在该取值范围内进行了实验,如图 2 所示。图 2 中分别为 ionosophere、seeds、heart、cheveland 这 4 个数据集在不同 ρ_1、ρ_2 取值下的分类准确率。

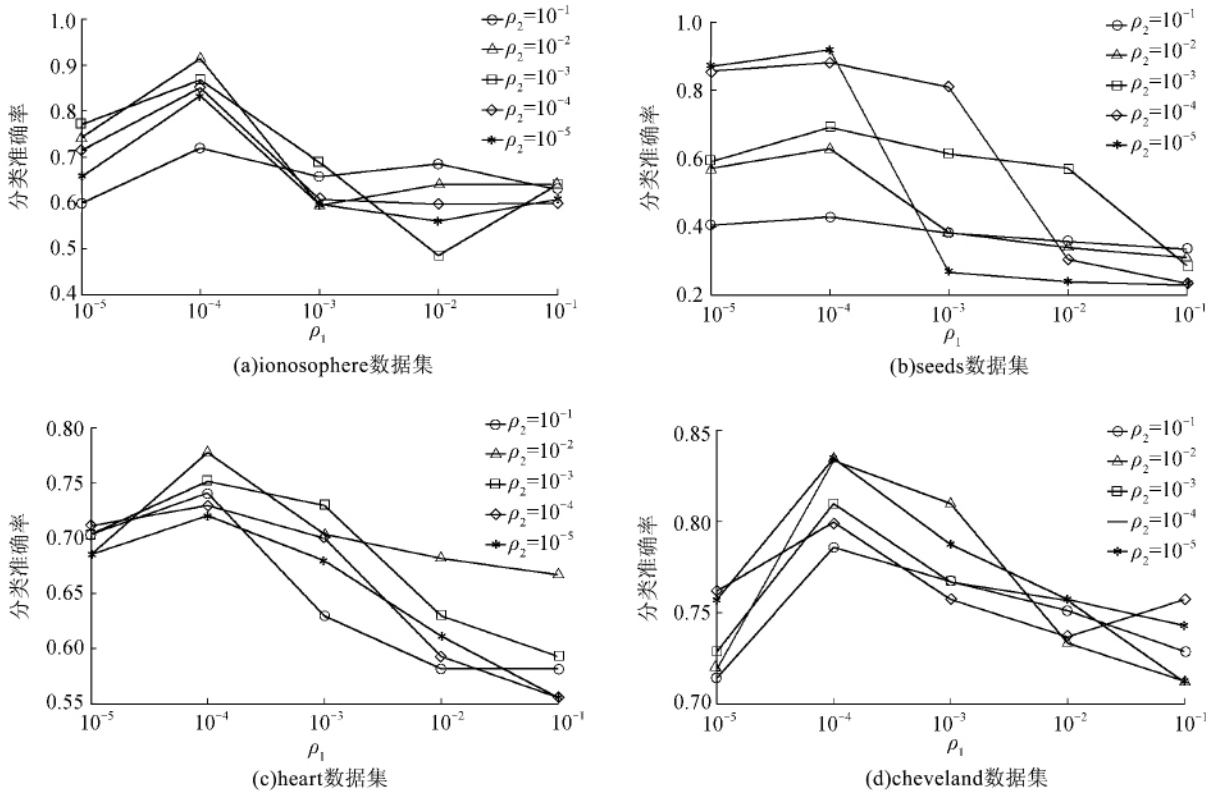

图 2 LS-kNN 算法在不同数据集不同参数取值下的分类准确率

Fig. 2 Classification accuracy of LS-kNN on the different datasets with different parameter setting

从图 2 可以看出,当参数 $\rho_1 \in (10^{-5},10^{-4})$,$\rho_2 \in (10^{-5},10^{-4})$ 范围时,LS-kNN 算法的分类准确率达到最优,因此我们将在后面的对比实验中,令 ρ_1、ρ_2 在 $(10^{-5},10^{-4})$ 范围内随机取值。

通过对 ρ_1、ρ_2 进行调参使 LS-kNN 达到最优,我们将其与 kNN 算法和 LMNN 算法进行对比,实验过程中对每个数据集都重复 10 次,不但报告每次的实验结果而且报告 10 次结果的均值和方差,如表 1。

表 1 kNN、LMNN 与 LS-kNN 算法的分类准确率

Tab. 1 Classification accuracy of kNN, LMNN and LS-kNN

数据集	kNN	LMNN	LS-kNN
ionosophere	0.768 6±0.003 2	0.840 0±0.004 0	0.874 3±0.003 1
seeds	0.757 1±0.002 7	0.819 0±0.002 9	0.900 0±0.003 3
heart	0.718 5±0.004 6	0.774 1±0.004 4	0.811 1±0.004 7
cheveland	0.761 9±0.004 5	0.804 8±0.004 3	0.861 9±0.005 8

为了更加直观地比较 kNN、LMNN 和 LS-kNN 的分类性能，我们做了对比图，如图 3 所示。由图 3 中 4 个数据集得出的实验结果可知，LS-kNN 算法的分类准确率明显高于传统 kNN 和 LMNN 算法。

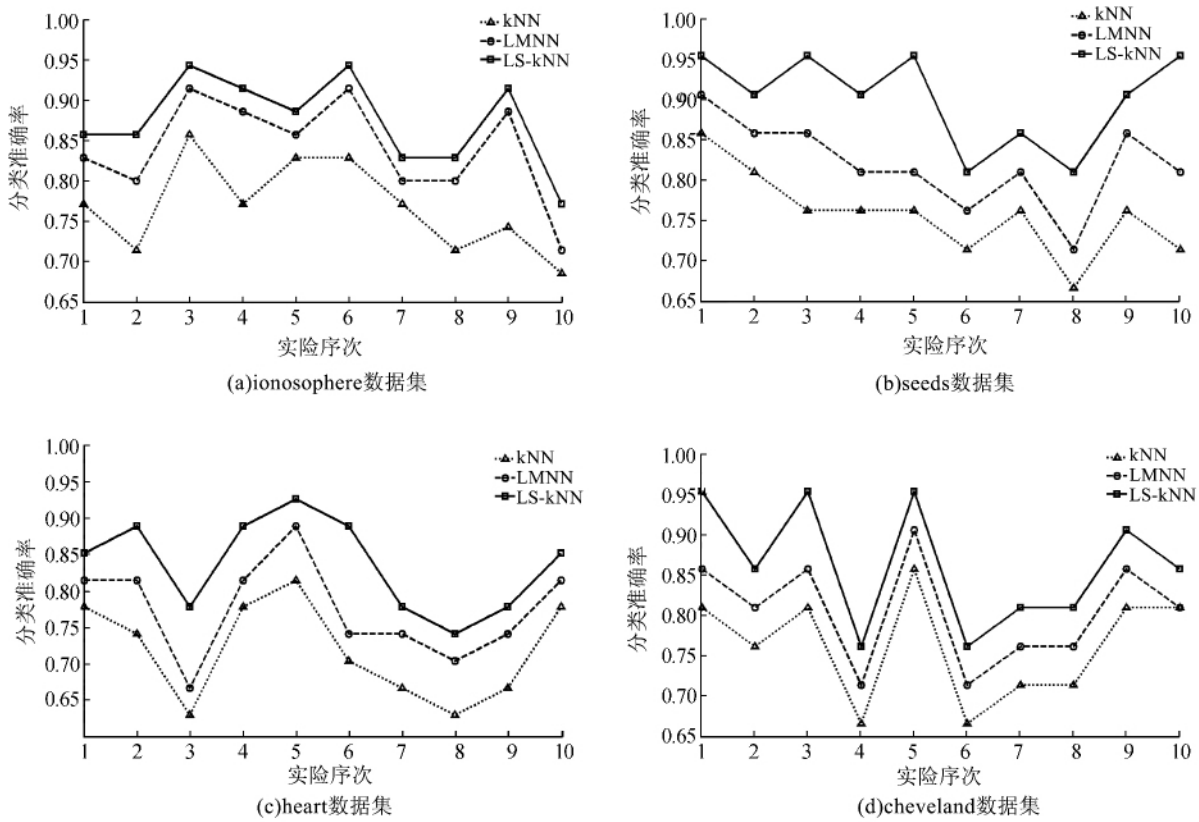

图 3　kNN，LMNN 与 LS-kNN 算法在 4 个数据集上的分类准确率

Fig. 3　Classification accuracy of kNN，LMNN and LS-kNN on four datasets

4　结束语

本文提出了一种基于 LPP 和 l_1-范数的 kNN 改进算法 LS-kNN，有效解决了传统 kNN 分类算法中的两个缺陷：k 值需事先给定且固定；未考虑样本之间的相关性。该算法思想是根据样本的相关性和拓扑结构，用训练样本对所有测试样本进行重构，以实现 k 值的选取由数据本身驱动，而无需人为设定。因此，本算法可用于无法通过经验或者需要长时间实验选定 k 值的情况，大幅度减少选取 k 值的时间。最后通过与两种算法以分类准确率为评价标准进行对比实验，实验结果表明，LS-kNN 算法的分类准确率优于 kNN 分类算法。

参　考　文　献：

[1]　张著英，黄玉龙，王翰虎. 一个高效的 KNN 分类算法[J]. 计算机科学，2008，35(3)：170-172.

[2]　李荣陆，胡运发. 基于密度的 kNN 文本分类器训练样本裁剪方法[J]. 计算机研究与发展，2004，41(4)：539-545.

[3]　张孝飞，黄河燕. 一种采用聚类技术改进的 KNN 文本分类方法[J]. 模式识别与人工智能，2009，22(6)：936-940.

[4]　陆广泉，谢扬才，刘星，等. 一种基于 KNN 的半监督分类改进算法[J]. 广西师范大学学报(自然科学版)，2012，30(1)：45-49. DOI：10.16088/j.issn.1001-6600.2012.01.004.

[5]　HAN Jiawei，KAMBER M. Data mining：concepts and techniques[M]. Waltham，MA：Morgan Kanfmann Publishers，2000.

[6] ZHANG Shichao. Cost-sensitive classification with respect to waiting cost[J]. Knowledge-Based Systems, 2010,23(5): 369-378. DOI: 10.1016/j.knosys.2010.01.008.

[7] LALL U, SHARMA A. A nearest neighbor bootstrap for resampling hydrologic time series[J]. Water Resouarces Research,1996,32(3): 679-693. DOI: 10.1029/95WR02966.

[8] LIU Huawen, ZHANG Shichao. Noisy data elimination using mutual k-nearest neighbor for classification mining[J]. Journal of Systems and Software,2012,85(5): 1067-1074. DOI:10.1016/j.jss.2011.12.019.

[9] WU Xindong, ZHANG Chengqi, ZHANG Shichao. Database classification for multi-database mining[J]. Information Systems, 2005,30(1): 71-88. DOI:10.1016/j.is.2003.10.001.

[10] JENATTON R, GRAMFORT A, MICHEL V, et al. Multi-scale mining of fMRI data with hierarchical structured sparsity[J]. Siam Journal on Imaging Sciences,2012,5(3): 835-856. DOI:10.1137/110832380.

[11] ZHU Xiaofeng, HUANG Zi, SHEN Hengtao, et al. Dimentionality reduction by mixed-kernel canonical correlation analysis[J]. Pattern Recognition,2012,45(8): 3003-3016. DOI:10.1016/j.patcog.2012.02.007.

[12] ZHU Xiaofeng, HUANG Zi, CHENG Hong, et al. Sparse hashing for fast for fast multimedia search[J]. ACM Transaction on Information System,2013,31(2): 9. DOI:10.1145/2457465.2457469.

[13] KANG P, CHO S. Locally linear reconstruction for instance-based learning[J]. Pattern Recogntion,2008,41(11): 3507-3518. DOI: 10.1016/j.patcog. 2008.04.009.

[14] LIANG Jinjin, WU De. Sparse least square support vector machine with L1 norm[J]. Computer Engineering and Design,2014,35(1): 293-296,338.

[15] BOYD S. Alternating direction method of multipliers[EB/OL]. [2015-03-25]. http://stanford.edu/class/ee364b/lectures/admm_slides.pdf.

[16] CHANG C C, LIN C J. LIBSVM: A library for support vector machine[J]. ACM Transactions on Intelligent Systems and Technology,2011,2(3):27. DOI: 10.1145/1961189.1961199.

A kNN Classification Algorithm Based on Local Correlation

DENG Zhenyun[1], GONG Yonghong[1,2], SUN Ke[1], ZHANG Jilian[1]

(1. Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin Guangxi 541004, China; 2. Guilin University of Aerospace Technology, Guilin Guangxi 541004, China)

Abstract: As a simple and effective classification algorithm, kNN algorithm is widely used in text classification. However, the k value (usually fixed) is usually set by users. For this purpose, the reconstruction and locality preserving projections (LPP) technology is introduced into the nearest neighbor classification, which makes the selection of the k value to be determined by the correlation between the samples and the topology structure. The algorithm uses l_1-norm sparse coding method to reconstruct the test sample by its k (not fixed) nearest neighbor samples and LPP keeps the local structure of the sample after the reconstruction, which not only solves the problem of choosing k value, but also avoids the influence of fixed k value on classification. Experimental results show that the classification performance of the proposed method is better than that of the classical kNN algorithm.

Keywords: k-nearest neighbor; locality preserving projections; reconstruction; sparse coding

(责任编辑 黄 勇)