



DOI:10.12404/j.issn.1671-1815.2308656

引用格式:李蒙蒙,杨中良,岳彩通,等.基于最小成分本征向量量子空间投影的近邻分类算法[J].科学技术与工程,2024,24(36):15511-15517.

Li Mengmeng, Yang Zhongliang, Yue Caitong, et al. Nearest neighbor classification algorithm based on minimum component eigenvector subspace projection[J]. Science Technology and Engineering, 2024, 24(36): 15511-15517.

自动化技术、计算机技术

基于最小成分本征向量量子空间投影的近邻分类算法

李蒙蒙^{1,2}, 杨中良³, 岳彩通^{1,2}, 万红^{1,2}, 李志辉^{1,2}, 尚志刚^{1,2*}

(1. 郑州大学电气与信息工程学院, 郑州 450001; 2. 河南省脑科学与脑机接口技术重点实验室, 郑州 450001;
3. 三一重工股份有限公司, 苏州 215300)

摘要 近邻法是模式识别中的经典算法之一,其分类性能高度依赖样本间的距离度量方式。适当的距离度量方式有助于提高近邻法的分类性能。然而,当前此类算法多从判别模型的角度寻找最大化分类效果的度量,忽略了各类样本集的类聚集属性。鉴于此,基于最小成分本征向量提出一种量子空间投影近邻分类算法(nearest neighbor classification algorithm based on minimum component eigenvector subspace projection, NN_MCESP)。该算法结合了经典的主成分分析和近邻法,能够有效地实现基于最小成分本征向量投影的各类样本聚集属性分析,并完成基于量子空间近邻投票准则的分类。在多组分类数据集上通过与其他分类算法的实验对比,验证了 NN_MCESP 算法的有效性和稳定性。

关键词 近邻法(NN);主成分分析(PCA);最小成分;子空间投影

中图分类号 TP181; 文献标志码 A

Nearest Neighbor Classification Algorithm Based on Minimum Component Eigenvector Subspace Projection

LI Meng-meng^{1,2}, YANG Zhong-liang³, YUE Cai-tong^{1,2}, WAN Hong^{1,2}, LI Zhi-hui^{1,2}, SHANG Zhi-gang^{1,2*}

(1. School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China;
2. Henan Key Laboratory of Brain Science and Brain-Computer Interface Technology, Zhengzhou 450001, China;
3. Sany Heavy Industry Co., Ltd., Suzhou 215300, China)

[Abstract] The nearest neighbor algorithm is one of the most classical pattern recognition algorithms, which classification performance highly depends on the distance metric between samples. Appropriate distance metric can help improve the classification performance of the algorithm. However, such algorithms mostly seek metrics to maximize classification effectiveness from the perspective of discriminant models currently, ignoring the aggregation properties of various sample sets belonging to different classes. In view of this, a nearest neighbor classification algorithm based on minimum component eigenvector subspace projection (NN_MCESP) was proposed. This algorithm combined classic principal component analysis (PCA) and nearest neighbor algorithm, which can effectively implement aggregation properties analysis of various sample clusters based on minimum component eigenvector projection, and complete classification based on subspace nearest neighbor voting criteria. The effectiveness and stability of NN_MCESP were validated by comparing with other classification algorithms on multiple data sets.

[Keywords] nearest neighbor(NN); principal component analysis(PCA); minimum component; subspace projection

近邻法(nearest neighbor, NN)是一种常用的经典监督学习方法,其工作机制非常简单:对于给定的测试样本,以某种距离度量方式为标准计算其与训练样本的距离并进行排序,根据与其最近的样本

收稿日期:2023-11-06; 修订日期:2024-09-30

基金项目:国家自然科学基金(62301496, 62106230);国家资助博士后研究人员计划(GZC20232447);中国博士后科学基金特别资助项目(2021T140616);中国博士后科学基金面上资助项目(2021M692920);河南省自然科学基金(242300421411, 242300420277);重庆邮电大学大数据重点实验室开放基金(BDIC-2023-A-007, BDIC-2023-B-005)

第一作者:李蒙蒙(1990—),男,汉族,河南商丘人,博士,讲师。研究方向:机器学习、生物信号检测处理与脑机接口。E-mail:limengmeng@zzu.edu.cn。

*通信作者:尚志刚(1975—),男,汉族,甘肃兰州人,博士,教授,博士研究生导师。研究方向:机器学习与脑机接口。E-mail:zhigang_shang@zzu.edu.cn。

的类别信息实现分类判别^[1]。判别时通常选用投票法,即选择所有近邻样本中出现最多的类别标签作为该样本的判别结果。经典的近邻法直接使用欧几里得距离作为样本间的距离度量方式,而不考虑样本中蕴含的统计规律,这使其算法性能受到明显限制。

研究表明,对于特定的数据集,选择合适的度量方式对于提高模式识别和数据挖掘算法的性能至关重要^[2],而学习合适的距离度量可以显著提高经典 NN 的分类性能^[3]。早期学习距离函数的信息论算法采用 Bregman 投影,在距离函数约束下最小化两个多变量 Gaussian 间的差分相对熵,取得了很好的效果。之后,基于投影子梯度下降方法的求解器,使相距最近的样本总是属于同一类,而来自不同类的样本间距离很大,从而显著改进了传统近邻法的精度^[4]。Li 等^[5]整合两种基于本征值的方法构建一个统一框架,直接寻找类内距离和类间距离之间的平衡,基于对本征分解度量的线性变换进行直接优化,有效调整了数据在变换空间的分布,快速有效解决了分类问题。

针对近邻分类中的距离度量问题,来自根特大学的 Bernard De Baets 教授课题组开展了一系列研究。2017 年,该课题组的 Nguyen 等^[6]提出了一种最大化 Universum 样本不一致数量的策略,解决近邻分类任务中等价函数复杂性度量的问题,表现出良好的性能。2018 年,Nguyen 等^[7]基于最大间距原则的启发,使用坡道损失函数构造边距最大化目标函数,提出了高泛化性的距离度量分类算法,并通过实验证明了其在避免异常值影响方面的优越性。

然而,上述算法多基于判别模型的思路学习得到度量矩阵,并通过最大化类内距离和类间距离间的差异实现判别。它们基于条件概率分布构造预测模型描述各类间的差异,不能很好地反映每个类的类聚集特性,往往造成因不顾样本结构而强行分类带来的性能提升瓶颈问题。而基于生成模型的思路则通过学习数据的联合概率分布,考虑样本所在的类属性构造生成空间并完成预测,基于本征分解构造生成空间用于距离度量就是其中的一种典型思路。根据这一思路,利用子空间^[8]投影对原始样本所有特征的张成空间进行特定操作以降低空间维数,并在学习到的样本子空间完成投影映射实现数据降维,为解决上述瓶颈提供了可能。

绝大多数的维数约简算法都可以溯源到子空间投影思想,如主成分分析(principal component analysis, PCA)^[9]、线性判别分析(linear discriminant analysis, LDA)^[10]、局部保留投影(locality preserving

projections, LPP)和局部线性嵌入(locally linear embedding, LLE)等^[11]。基于子空间投影改进距离度量方式的分类思路已被验证。

较早研究提出的基于最小成分分析投影和最小似然决策规则的分类算法,受到先验概率和高斯分布假设以及计算速度的限制,有待进一步改进;使用本征值优化的思路在每次迭代计算中只计算最大本征值和相应本征向量的方式,保证了结果的严格收敛,提高了算法的效率,但仍有进一步改进的空间^[12]。在此基础上,Luo 等^[13]提出了一种基于组成轮廓的形状模型,结合多个距离度量提出一种子空间生成学习算法,在信息投影原理下实现对最具代表性特征的提取,实验结果验证了算法性能,但其在解决大规模问题时的性能仍有待提升。Nguyen 等^[14]利用协方差矩阵的本征分解结果,将选定的本征向量组成变换矩阵,以最大化 Jeffrey 散度,构建了近邻距离度量学习的优化框架,保留了局部判别信息,提升分类性能。这些尝试为基于本征分解构建子空间进行分类判别提供了解决方案,进一步为基于生成模型思想的近邻分类奠定了基础。

针对上述问题,基于在主成分分析本征向量投影子空间内实现近邻法度量学习的思路,提出了基于最小成分本征向量子空间投影的近邻分类算法(nearest neighbor classification algorithm based on minimum component eigenvector subspace projection, NN_MCESP),旨在提高分类的有效性和稳定性。与其他算法不同的是,NN_MCESP 在每个类内计算样本集的协方差矩阵,并基于本征分解得到最小成分及其对应的本征向量,以生成投影子空间,最终结合各子空间内的近邻分类结果综合投票完成预测分类。

1 基于最小成分本征向量子空间投影的近邻分类算法

在子空间投影中,重要的是找到合适的投影矩阵,该矩阵可以视作对数据集新表征的坐标系集合,亦即在样本空间中张成一个子空间。鉴于 PCA 可以在降低维度的同时保留原始数据集的结构信息,而各类样本的最小成分表征又可以用来描述数据的类聚集特性,因此结合 PCA 的最小成分子空间投影和子空间内的近邻投票分类,提出一种基于最小成分本征向量子空间投影的近邻分类算法,该算法由两部分组成,具体如下。

(1) 利用最小成分本征向量构建生成子空间。结合训练数据集中的类别标签信息完成对各类数

据的主成分分析,按照本征值由小到大得到对应的成分排序,进一步地设置累计贡献率阈值确定入选生成子空间的最小成分个数,得到每类样本对应的最小成分本征向量量子空间。

(2) 基于子空间投影与近邻法的投票分类。将训练样本和测试样本依次投影到各类数据对应的生成子空间中,并综合各类子空间内近邻分类的结果按照特定准则投票完成判别,确定每个测试样本的最终标签。

1.1 利用最小成分本征向量构建生成子空间

PCA 是一种最经典常用的降维算法。对含有 m 个样本, d 维特征的数据集 $X = \{x_1, x_2, \dots, x_m\} \in \mathbf{R}^{m \times d}$, 首先进行数据中心化处理,可表示为

$$\sum_{i=1}^m x_i = 0 \quad (1)$$

之后,计算样本间的协方差矩阵,计算公式为

$$C = XX^T \quad (2)$$

求协方差矩阵 C 的本征值 $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_m)$ 和相对应的本征向量 $W = (w_1, w_2, \dots, w_m)$ 。

原始的 PCA 不考虑类别信息,基于最大方差理论,将所有本征值进行由大到小的排序,并提取排序靠前的特定数量的本征值对应的本征向量组成投影空间,将原始数据投影到该空间中实现高维数据的低维映射。

对于多类训练数据集,取其中任何一类,首先对数据进行主成分分析得到本征值以及对应的本征向量,之后对本征值进行从小到大的排序,并按顺序计算本征值的累积贡献率,当累积本征值之和大于本征值总和的 $T\%$ 时,记录对应的本征值数量,取该数量相对应的本征向量组成最小成分本征矩阵,即为生成子空间。基于最小成分本征向量构建生成子空间的算法伪代码如下。

算法: 基于最小成分本征向量构建生成子空间
输入: 数据集 $X = \{x_1, x_2, \dots, x_m\}$; 类别数 M ; 小本征值累计贡献率 $T\%$
过程:
for $i = 1, 2, \dots, C = 1$ to do
计算第 i 类数据 X_i 的协方差矩阵 $C_i = X_i X_i^T$
对协方差矩阵进行本征值分解
本征值排序,计算得到最小本征值累计贡献率达到 $T\%$ 时对应的个数 d'
取最小的 d' 个本征值所对应的本征向量 $w_1, w_2, \dots, w_{d'}$
end for
输出: 各类样本的生成子空间 $W_i = (w_{i1}, w_{i2}, \dots, w_{id'})$

1.2 基于子空间投影与近邻法的投票分类

在完成各类样本集的主成分分析并得到对应的生成子空间后,将训练集和测试集数据分别投影

到各类的生成子空间。基于每个子空间内的近邻法分类结果进行综合投票,最终确定测试样本的类别标签。其中投票准则为子空间结果优先,若子空间结果存在冲突则采信原空间结果。

以二元分类问题为例,测试样本被投影在 A 类子空间中,则可实现对该样本类别的 A 或非 A 判别;类似地,该样本被投影在 B 类子空间则实现 B 或非 B 类判别。总结可能出现的 4 种判别结果组合,如表 1 所示。子空间优先准则为:若在两类子空间投影分类后分别获得标签 A 与非 B,则将该样本判为 A 类;获得标签非 A 与 B,则将该样本判为 B 类。如不符合上述准则,即在 A 类子空间投影获得分类标签为 A 且 B 类子空间获得分类标签 B,或 A 类子空间投影分类结果获得分类标签为非 A 且 B 类子空间获得分类标签非 B 时,则采信原始空间中的分类结果。

表 1 基于投票的二分类判别准则

Table 1 Voting-based binary classification criteria			
A 类子空间 判别结果	B 类子空间 判别结果	原始空间 判别结果	最终 判别结果
A	非 B	—	A
非 A	B	—	B
A	B	A	A
		B	B
非 A	非 B	A	A
		B	B

2 实验分析

2.1 实验数据与对比算法

在公开数据集上进行分类性能对比实验,使用数据集的详细信息如表 2 所示。将所提的 NN_MCESP 算法与支持向量机 (support vector machine, SVM)、Logistic 回归 (logistic regression, LR)、朴素贝叶斯 (naive Bayes, NB)、经典近邻法 NN 进行了比较。其中, SVM 使用了线性核函数。NN_MCESP 和 NN 算法中近邻参数 K 的默认初始设置为 1, 并使用欧几里得距离作为距离度量方式。NN_MCESP 算法中的最小成分累计贡献率阈值参数的默认初始设置为 $T=5$ 。

2.2 实验方法与评价指标

在数值实验中,使用的所有数据集在分类之前先进行了归一化操作。同时,将所有数据处理和分类过程进行 100 次重复测试,每次测试中使用 5 倍交叉验证计算分类精度。计算 100 次测试结果的平均值和标准差,并使用 Wilcoxon 秩和检验评估不同算法性能间的差异显著性,显著性水平设置为 0.05。

2.3 实验结果与分析

为了评估 NN_MCESP 的有效性,进行对比实验,结果如表 3 所示。进一步地,针对算法间的统计显著性差异检验,总结算法间的性能比较,结果如表 4 所示。

综合表 3 和表 4 中的结果可知,与其他传统分类算法相比,NN_MCESP 算法有着巨大的分类性能优势。具体讲,NN_MCESP 算法在绝大多数测试数据集上都获得了最佳的效果(10/12),根据 Wilcoxon 秩和检验的结果,NN_MCESP 算法在所有数据集上的分类性能均不显著低于其他四种算法。具体地,与 SVM 相比,NN_MCESP 算法的显著性优势体现在 10 个数据集上;与 NB 和 LR 的对比中,NN_MCESP 算法分别在 9 个数据集上具有显著性优势;与 NN 相比,NN_MCESP 算法在 7 个数据集上具有显著性优势。

由于对比算法尤其是 NN 算法在进行分类模型的训练时,仅在全局所有类别样本集的角度上构建判别式模型,并没有像 NN_MCESP 算法一样充分考虑各类样本集的聚类效应统计规律,结合生成式模型进行处理,直接导致了分类精度的受限。在这种情况下,融合 PCA 最小成分子空间投影和近邻分类思路的 NN_MCESP 算法处理该类问题,尤其是原始空间中可分性不佳的数据集的分类问题时,更具有实用性。

表 3 NN_MCESP 与其他 4 种分类算法性能比较

Table 3 Performance comparisons of NN_MCESP and other four classification algorithms

序号	数据集	SVM	NB	LR	NN	NN_MCESP
1	Vote	93.83 ± 1.95	93.49 ± 2.59	93.90 ± 2.41	91.42 ± 2.33	94.15 ± 0.80
2	Monk-2	80.53 ± 3.96	91.74 ± 3.70	76.22 ± 4.05	75.81 ± 3.86	91.84 ± 4.73
3	ionosphere	83.79 ± 3.61	81.93 ± 4.12	84.34 ± 3.57	86.70 ± 4.03	87.15 ± 1.22
4	Musk1	85.26 ± 4.96	73.15 ± 6.26	80.73 ± 5.89	87.96 ± 2.69	88.02 ± 3.09
5	Magic	79.12 ± 0.53	72.68 ± 0.49	79.10 ± 0.51	81.79 ± 0.51	82.82 ± 0.50
6	FeaSelData	89.89 ± 3.20	87.06 ± 3.67	90.00 ± 3.21	89.89 ± 3.53	90.07 ± 1.61
7	Musk	95.47 ± 0.49	83.70 ± 0.99	95.26 ± 0.35	94.77 ± 0.58	96.91 ± 0.46
8	parkinsons	85.38 ± 2.43	68.46 ± 6.40	83.58 ± 4.55	94.46 ± 1.30	94.60 ± 0.83
9	Iris	96.40 ± 2.87	95.73 ± 2.96	97.86 ± 2.33	93.46 ± 3.66	96.53 ± 2.44
10	movement_libras	79.75 ± 3.53	62.04 ± 5.10	51.67 ± 6.55	84.88 ± 4.27	85.89 ± 4.22
11	segment	94.82 ± 0.96	79.83 ± 1.73	94.03 ± 1.17	96.31 ± 1.12	97.09 ± 0.79
12	Vowel	79.93 ± 1.68	66.26 ± 3.35	67.95 ± 3.07	89.37 ± 0.28	89.05 ± 0.56

注:分类性能最优的算法结果以加粗文字表示。

表 4 算法性能显著性差异分析结果

Table 4 Significant analysis results of the algorithms

算法对比	显著性差异分析结果		
	+	≈	-
NN_MCESP vs SVM	10	2	0
NN_MCESP vs NB	9	3	0
NN_MCESP vs LR	9	3	0
NN_MCESP vs NN	7	5	0

注:“+”表示 NN_MCESP 算法在精度上显著优于竞争算法;“≈”表示 NN_MCESP 算法和竞争算法的分类性能之间无显著差异;“-”表示 NN_MCESP 算法在精度上显著劣于竞争算法。

2.4 参数影响分析:近邻参数

近邻参数 K 是近邻法中的重要参数,其取值对最终分类结果产生显著影响。由于 NN_MCESP 算法的最终分类环节同样基于近邻法思想,因此 K 也是该算法的重要参数。为了充分对比不同近邻参数情况下,NN_MCESP 算法和经典 NN 算法的性能差异,分别设置对比实验分析了不同参数对性能的影响。 $K = 1, 3, 5, 7$ 这 4 种近邻参数设置下,NN_MCESP 算法和 NN 算法的分类性能结果,如表 5 所示。

表 2 实验数据集

Table 2 Experimental data sets

序号	数据集	样本数	特征数	类别数
1	Vote	435	16	2
2	Monk-2	432	6	2
3	ionosphere	351	34	2
4	Musk1	476	166	2
5	Magic	19 020	10	2
6	FeaSelData	268	8	2
7	Musk	6 598	166	2
8	parkinsons	195	22	2
9	Iris	150	4	3
10	movement_libras	360	90	15
11	segment	2 310	19	7
12	Vowel	990	10	11

从表 5 结果可以看出,在不同的近邻参数设置下,NN_MCESP 算法在多数数据集上的结果仍优于 NN 算法。随着近邻参数 K 的变化,NN_MCESP 算法的分类准确度在多数情况下没有表现出剧烈的变化,而基于 NN 算法则存在较大差异。为进一步对比验证这一发现,计算两种算法在 4 种情况下平均分类结果的变异系数,结果如图 1 所示。

从图 1 中两种算法在不同近邻参数设置下分类结果的变异系数对比可以发现,NN_MCESP 算法在

表 5 NN_MCESP 算法和 NN 算法在不同近邻参数设置下的分类结果

Table 5 Classification results of NN_MCESP and NN under different nearest neighbor parameters

算法	序号	数据集	$K=1$	$K=3$	$K=5$	$K=7$
NN_MCESP 算法	1	Vote	94.15 ± 0.80	94.09 ± 0.57	94.21 ± 0.64	94.16 ± 0.89
	2	Monk-2	91.84 ± 4.73	96.76 ± 2.17	96.23 ± 2.96	95.48 ± 3.09
	3	ionosphere	87.15 ± 1.22	81.93 ± 4.12	84.34 ± 3.57	86.70 ± 4.03
	4	Musk1	88.02 ± 3.09	87.15 ± 3.26	86.73 ± 2.89	87.01 ± 2.69
	5	Magic	82.82 ± 0.50	83.97 ± 0.50	84.30 ± 0.52	84.25 ± 0.48
	6	FeaSelData	90.07 ± 1.61	89.81 ± 0.68	87.97 ± 1.21	88.03 ± 0.98
	7	Musk	96.91 ± 0.46	97.01 ± 0.58	97.26 ± 0.67	97.77 ± 0.68
	8	parkinsons	94.60 ± 0.83	91.13 ± 1.36	90.56 ± 1.03	89.03 ± 1.83
	9	Iris	96.53 ± 2.44	96.40 ± 2.92	95.87 ± 3.05	95.47 ± 3.35
	10	movement_libras	85.89 ± 4.22	82.50 ± 3.26	80.75 ± 5.11	76.02 ± 4.32
	11	segment	97.09 ± 0.79	96.38 ± 0.82	95.58 ± 1.10	95.10 ± 0.87
	12	Vowel	89.05 ± 0.56	86.63 ± 0.65	85.15 ± 0.31	83.13 ± 1.92
NN 算法	1	Vote	91.42 ± 2.33	91.66 ± 0.60	93.15 ± 0.44	93.11 ± 0.43
	2	Monk-2	75.81 ± 3.86	95.11 ± 1.88	92.82 ± 2.40	89.63 ± 3.15
	3	ionosphere	86.70 ± 4.03	87.10 ± 4.98	87.11 ± 4.97	87.05 ± 5.01
	4	Musk1	87.96 ± 2.69	86.31 ± 3.62	85.73 ± 3.68	85.26 ± 3.16
	5	Magic	81.79 ± 0.51	83.46 ± 0.51	83.77 ± 0.48	83.94 ± 0.49
	6	FeaSelData	89.89 ± 3.53	89.11 ± 0.71	87.15 ± 0.96	86.31 ± 1.05
	7	Musk	94.77 ± 0.58	96.74 ± 0.18	96.61 ± 0.14	96.45 ± 0.11
	8	parkinsons	94.46 ± 1.30	91.08 ± 1.00	90.82 ± 1.33	89.69 ± 1.24
	9	Iris	93.46 ± 3.66	94.40 ± 4.22	94.33 ± 3.82	95.20 ± 3.50
	10	movement_libras	84.88 ± 4.27	80.82 ± 4.85	78.03 ± 4.41	72.08 ± 5.00
	11	segment	96.31 ± 1.12	95.43 ± 0.97	94.59 ± 0.82	93.94 ± 1.15
	12	Vowel	89.37 ± 0.28	87.11 ± 0.52	83.82 ± 0.65	78.98 ± 0.83

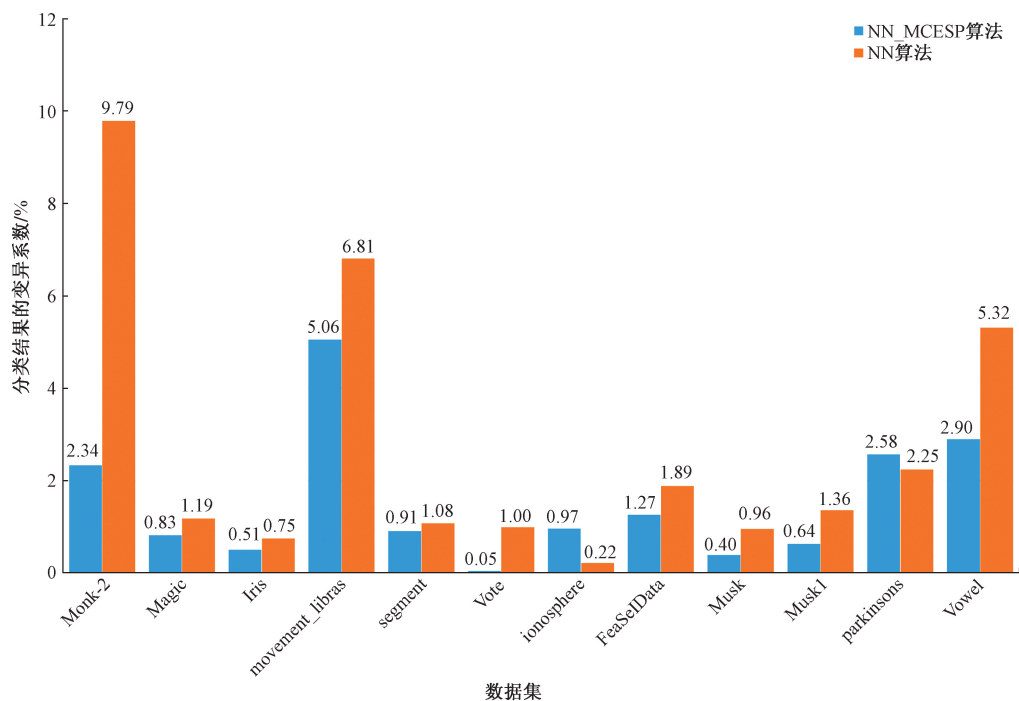


图 1 NN_MCESP 算法和 NN 算法在不同近邻参数设置下分类结果的变异系数

Fig. 1 Coefficient of variation of the classification results of NN_MCESP and NN under different nearest neighbor parameters

绝大多数的数据集上的变异系数均低于 NN 算法 (10/12), 表现更优。对于这一结果可以总结, 传统的 NN 算法在原始的样本空间进行距离度量, 在处

理各类样本分布稀疏甚至存在高度混叠的问题时, 其性能必然更容易受到近邻参数值选取的影响。而 NN_MCESP 算法由于考虑了每类样本的生成聚

类特性,将它们投影到了相对更具有致密性分布的空间中,因此即便在不同的近邻参数设置下,仍能表现出最佳的稳定性。

2.5 参数影响分析:最小成分累积贡献度阈值参数

NN_MCESP 算法基于主成分分析的最小成分提取本征向量,在其构成的生成子空间进行距离度量实现近邻分类。其中入选子空间构成向量的最小成分数由本征值累积贡献度 T 决定,这是 NN_MCESP 算法的另一个重要参数。为分析不同最小成分累积贡献度参数对 NN_MCESP 算法的影响,分

别以 $T=5、10、15、20$ 参数设置开展对比实验,分类性能结果如表 6 所示。

由表 6 结果可知,在不同的最小成分累积贡献度参数设置下,NN_MCESP 算法在多数数据集上的分类结果并没有表现出显著性差异(10/12)。这表明,即使只考虑非常小的阈值获取低维最小成分本征向量,将数据压缩到贡献度占比较小的子空间内,NN_MCESP 算法仍能获得较高的分类精度,这主要归功于各类最小成分本征向量子空间内的类聚集效应。

表 6 NN_MCESP 算法在不同最小分量累积贡献度阈值参数设置下的分类结果
Table 6 Classification results of NN_MCESP under different threshold parameters of the minimum components cumulative contribution

序号	数据集	$T=5$	$T=10$	$T=15$	$T=20$
1	Vote	94.15 ± 0.80	94.60 ± 0.72	95.01 ± 0.52	95.74 ± 0.75
2	Monk-2	91.84 ± 4.73	92.03 ± 2.18	95.65 ± 3.20	95.93 ± 3.72
3	ionosphere	87.15 ± 1.22	87.22 ± 1.31	87.53 ± 1.20	87.70 ± 1.04
4	Musk1	88.02 ± 3.09	86.85 ± 3.03	86.61 ± 2.68	86.09 ± 2.30
5	Magic	82.82 ± 0.50	83.02 ± 0.46	83.21 ± 0.45	83.75 ± 0.64
6	FeaSelData	90.07 ± 1.61	90.23 ± 0.66	90.65 ± 0.80	90.82 ± 0.61
7	Musk	96.91 ± 0.46	96.76 ± 0.28	97.15 ± 0.38	96.76 ± 0.18
8	parkinsons	94.60 ± 0.83	94.13 ± 1.38	94.88 ± 1.44	95.28 ± 2.08
9	Iris	96.53 ± 2.44	96.13 ± 2.56	96.73 ± 2.86	96.80 ± 2.68
10	movement_libras	85.89 ± 4.22	85.44 ± 3.14	86.31 ± 3.07	86.64 ± 3.07
11	segment	97.09 ± 0.79	96.60 ± 0.72	97.13 ± 0.91	97.22 ± 0.89
12	Vowel	89.05 ± 0.56	91.22 ± 0.52	95.75 ± 0.57	96.95 ± 0.39

2.6 最小成分本征向量相似度分析

针对各类样本分别构造最小成分本征向量子空间,其相似度对算法分类性能的影响较为显著。若每类样本集对应最小分量本征向量相似度较高,则在对应生成子空间的区分度就会受到影响,无论数据被投影到任意类的生成子空间中,变换后的数据仍可能不具备很好的可分性,不利于随后的近邻分类。因此选取典型二分类数据集,分别计算各类数据进行主成分分析后得到本征向量的余弦相似度,以评价在其生成子空间进行数据变换后的可分性,结果如图 2 所示。

余弦相似度的计算公式为

$$\text{similarity} = |\cos\theta| = \frac{|\mathbf{v}_1 \mathbf{v}_2|}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|} \quad (3)$$

式(3)中: \mathbf{v}_1 和 \mathbf{v}_2 为两类样本集对应的本征向量。

由图 2 所示的分析结果可知,总体而言,除 Musk 数据集外,其他所选二分类数据集上,两类样本对应的较小成分本征向量的相似度保持在较低的水平,这就可以保证原始样本数据在投影到其组

成的子空间后保持较好的聚集度,从而获得较好的可分性,这有助于提高算法的分类性能。

3 结论

基于最小成分本征向量生成子空间投影的近邻分类算法有助于构造适用于分类任务的距离度量方式,提升分类精度。提出的 NN_MCESP 算法,旨在充分考虑各类样本集类聚集属性,结合经典的主成分分析和近邻法,有效地提取各类样本最小成分组成本征向量以构建投影子空间,并基于子空间内的近邻投票准则实现分类,通过在多个标准分类数据集上的分类判别实验并与多种分类算法比较,验证了所提出的 NN_MCESP 算法的有效性和稳定性,得到如下结论。

(1) NN_MCESP 算法采用最小成分本征向量子空间内的距离度量方式,分类性能显著优于经典分类算法。

(2) NN_MCESP 算法考虑了每类样本的生成聚集特性,在不同近邻参数下的稳定性较经典近邻算

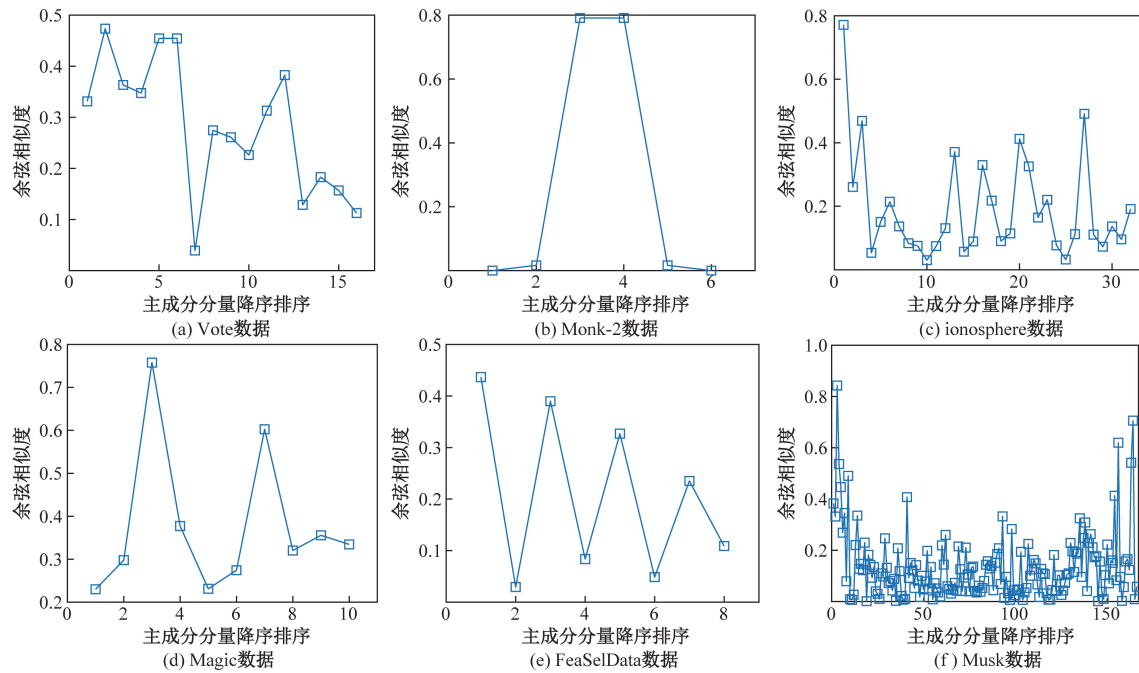


图2 各类数据进行主成分分析得到本征向量的余弦相似度

Fig. 2 Cosine similarity of eigenvectors obtained through PCA on binary classification datasets

法更佳。

(3) 低相似度各类样本最小成分本征向量的入选有助于提升 NN_MCESP 算法的性能。

参 考 文 献

- [1] 梁淑蓉, 陈基漓, 谢晓兰. 基于权重搜索树改进 K 近邻的高维分类算法[J]. 科学技术与工程, 2021, 21(7): 2760-2766.
Liang Shurong, Chen Jili, Xie Xiaolan. Improved K -nearest neighbor algorithm based on weight search tree for high-dimensional classification[J]. Science Technology and Engineering, 2021, 21(7): 2760-2766.
- [2] 邓秀勤, 郑丽苹, 张逸群, 等. 基于新的距离度量的异构属性数据子空间聚类[J]. 郑州大学学报(工学版), 2023, 44(2): 53-60.
Deng Xiuqin, Zheng Liping, Zhang Yiqun, et al. Subspace clustering of heterogeneous-attribute data based on a new distance metric [J]. Journal of Zhengzhou University (Engineering Science), 2023, 44(2): 53-60.
- [3] Gou J, Ma H, Ou W, et al. A generalized mean distance-based K -nearest neighbor classifier[J]. Expert System with Application, 2019, 115: 356-372.
- [4] Le N D, Nguyen N T H. A metric learning-based method for biomedical entity linking[J]. Frontiers in Research Metrics and Analytics, 2023, 8: 1247094.
- [5] Li D, Tian Y. Global and local metric learning via eigenvectors [J]. Knowledge-Based System, 2017, 116: 152-162.
- [6] Nguyen B, Morell C, De Baets B. Distance metric learning with the universum[J]. Pattern Recognition Letter, 2017, 100: 37-43.
- [7] Nguyen B, Morell C, De Baets B. An approach to supervised distance metric learning based on difference of convex functions programming[J]. Pattern Recognition, 2018, 81: 562-574.
- [8] 张杰, 曲洪权, 柳长安, 等. 基于双子空间 PCA 降维的脑力负荷分类[J]. 科学技术与工程, 2024, 24(11): 4433-4438.
Zhang Jie, Qu Hongquan, Liu Chang'an, et al. Research on classification of mental workload based on dimension reduction of PCA in two subspaces[J]. Science Technology and Engineering, 2024, 24(11): 4433-4438.
- [9] 李晓娟, 张芳媛, 喻玲. 基于主成分分析-BP 神经网络的风电备件需求预测[J]. 科学技术与工程, 2024, 24(1): 281-288.
Li Xiaojuan, Zhang Fangyuan, Yu Ling. Research on wind power spare parts demand forecasting based on PCA-BP neural network[J]. Science Technology and Engineering, 2024, 24(1): 281-288.
- [10] 李蒙蒙, 尚志刚, 李志辉. 结合投影与近邻操作的支持向量快速筛选方法[J]. 郑州大学学报(工学版), 2017, 38(3): 49-53.
Li Mengmeng, Shang Zhigang, Li Zhihui. Fast method to filter support vectors combined with operation of projection and nearest neighbors' selection[J]. Journal of Zhengzhou University (Engineering Science), 2017, 38(3): 49-53.
- [11] Sibli W, Kuntz P, Meyer F. A review on dimensionality reduction for multi-label classification[J]. IEEE Trans. on Knowledge and Data Engineering, 2021, 33(3): 839-857.
- [12] Ruan Y, Xiao Y, Hao Z, et al. A convex multi-class model via distance metric learning based class-to-instance confidence [J]. Knowledge-Based Systems, 2022, 258: 109791.
- [13] Luo P, Lin L, Liu X. Learning compositional shape models of multiple distance metrics by information projection [J]. IEEE Transactions on Neural Networks and Learning System, 2016, 27: 1417-1428.
- [14] Nguyen B, Morell C, De Baets B. Supervised distance metric learning through maximization of the Jeffrey divergence [J]. Pattern Recognition, 2017, 64: 215-225.