Advances in Applied Mathematics 应用数学进展, 2020, 9(10), 1757-1765
Published Online October 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.910203

Hans汉斯

一种支持向量机预处理方法的研究

韩成志*，李梦婷，郑恩涛，马国春#

杭州师范大学理学院，浙江 杭州
Email: 1224764141@qq.com, limengting@stu.hznu.edu.cn, 851347680@qq.com, #maguochun@163.com

收稿日期：2020年10月7日；录用日期：2020年10月20日；发布日期：2020年10月27日

摘　要

支持向量机(Support Vector Machine, SVM)在处理大规模数据集时，随着样本维度增高，样本数量增多会出现训练时间显著增多的问题。为了解决该问题，文章提出了一种基于主成分分析(Principal Component Analysis, PCA)和K边界近邻法(K Nearest Bound Neighbor, KNBN)的SVM预处理方法；先用PCA对训练数据降维消除训练数据中的冗余信息，然后利用KNBN预选取训练数据中的支持向量来减少训练数据量。数值实验结果表明，与PCA-SVM、KNBN-SVM和无数据预处理的SVM方法相比，采用本文提出的SVM预处理方法既保持了良好的分类预测精度，又缩短了大量训练时间。

关键词

支持向量机，主成分分析法，K边界近邻法，预处理，训练时间

Research on a Preprocessing Method of Support Vector Machine

Chengzhi Han*, Mengting Li, Entao Zheng, Guochun Ma#

College of Science, Hangzhou Normal University, Hangzhou Zhejiang
Email: 1224764141@qq.com, limengting@stu.hznu.edu.cn, 851347680@qq.com, #maguochun@163.com

Received: Oct. 7th, 2020; accepted: Oct. 20th, 2020; published: Oct. 27th, 2020

Abstract

When the large-scale data set with higher dimension and larger number of samples is processed by the Support Vector Machine (SVM), the training time will increase significantly. In order to

*第一作者。
#通讯作者。

文章引用: 韩成志, 李梦婷, 郑恩涛, 马国春. 一种支持向量机预处理方法的研究[J]. 应用数学进展, 2020, 9(10): 1757-1765. DOI: 10.12677/aam.2020.910203

solve this problem, based on Principal Component Analysis (PCA) and K Nearest Bound Neighbor (KNBN), a preprocessing method of SVM is proposed. Firstly, PCA is used to reduce the dimension of the training data to eliminate the redundant information in the training data, and then KNBN is used to preselect the support vectors in the training data to reduce the amount of training data. The numerical experiment results show that SVM preprocessing method proposed in this paper, compared with PCA-SVM, KNBN-SVM and SVM without data preprocessing, can not only keep good classification prediction accuracy, but also save a lot of training time.

Keywords

Support Vector Machine, Principal Component Analysis, K Nearest Bound Neighbor, Preprocessing, Training Time

1. 引言

支持向量机(Support Vector Machine, SVM)最早在 1964 年被 Vapnik 和 Cortes 等人提出的，它是一个有监督学习的二分类模型[1] [2] [3]。在上世纪 90 年代 SVM 得到快速发展并衍生出一系列改进和扩展的算法。它被广泛应用在人像识别[4]和文本分类[5]等模式识别问题中。

SVM 的学习策略是求解能够将训练数据集按照类别正确划分并且使几何间隔最大化的分离超平面，可转化为求解一个凸二次规划问题[3]。一般情况下，随着训练样本规模的扩大，即样本维度升高和样本数量增多，会导致 SVM 的训练时间延长，分类精度可能下降，存储空间也有所增加。如何对 SVM 数据进行预处理来提高 SVM 分类效率成为近年来的一个研究热点。

对 SVM 数据降维消除其中的冗余特征信息，可以起到提高 SVM 分类精度和减少 SVM 计算量的效果。PCA 通过提取最能代表原始数据本质特征的特征因子来实现数据降维。目前有关 PCA 和 SVM 结合方面的研究有很多。2018 年，余金澳[6]等人提出一种面向方位敏感性的 PCA-SVM 分类识别方法，与传统的 SVM 分类方法相比，该方法对 SAR 图像的地面目标具有较高的分类识别率和运行效率。2018 年，Anis Ben Aicha [7]等人利用 PCA 提取最关键、最相关的特征，采用 SVM 作为分类技术，实现正常肿瘤和癌前肿瘤的鉴别，此方法具有较好的敏感性、特异性、精密度和准确性。2019 年，汪雯琦[8]等人提出基于 PCA 和 SVM 分类的跨年龄人脸识别方法，该方法具有速度快和准确度高的优点。2019 年，Wang [9]等人提出基于机器视觉的有色金属报废车辆分离系统的分类算法及运行参数优化研究，其中利用了 PCA-SVM 使得识别精度较高，计算速度足够快。

SVM 的最优分离超平面由只占全体训练数据集一小部分的支持向量来确定[10]，通过预选取有可能成为支持向量的样本，舍弃非支持向量来减少训练样本数量，进而大幅度缩短 SVM 训练时间。从几何方面来看，线性可分情况下支持向量主要分布在两类样本的边界上且彼此之间靠的很近的那些样本[11]，因此可以提取那些异类样本之间离得很近的边界向量作为支持向量候选集，最后将其作为训练样本集进行 SVM 训练，可以减少计算量，提升训练速度。一些学者在支持向量预选取方面进行了研究。2008 年，Zhang [10]提出基于 KNN 法的支持向量预选取方法，计算距离每个样本最近的 k 个样本，若 k 个样本中至少有一个异类样本，则这 k 个样本是边界向量。2009 年，徐红敏[11]等人提出一种支持向量机快速分类算法，在两类样本中选取与每类样本中每个样本最近的异类样本作为边界向量。2013 年，胡志军[12]

等人提出基于距离排序的快速支持向量机分类算法，先计算每类样本与异类样本类中心的距离，再按照距离大小排序选取一定比例的小距离样本作为候选支持向量。2013 年，李庆[13]等人提出 K 边界近邻法支持向量预选取方法，在两类样本中选取与每类样本中每个样本最近的 k 个异类样本作为边界向量。

上述研究是从降低样本的维度方面或是从减少样本数量方面解决问题，比较片面性，不适用于同时维度高和数量多的大规模数据集预处理。为了克服这个问题，本文首先选用 PCA 对 SVM 大规模数据集降维，然后用 KNBN 在降维后的数据集上预选取支持向量得到一个约简集，称该方法为基于 PCA 和 KNBN 的 SVM 预处理方法。该 SVM 预处理方法能够结合 PCA 和 KNBN 的优点，拥有较高的分类预测精度，而且其训练时间大量缩短。

本文接下来第 2 节是主成分分析法的简介，第 3 节是 K 边界近邻法支持向量预选取方法的描述，第 4 节介绍基于 PCA 和 KNBN 的 SVM 预处理方法，第 5 节是数值实验结果及比较。

2. 主成分分析法简介

主成分分析的基本原理是将原来变量重新组合成一组新的线性无关的几个变量，同时根据实际需要从中可以取出几个较少的变量尽可能多地反映原来变量的信息的统计方法，即在原有样本的 M 维空间内，用 M 个标准正交基进行重新映射，然后选取其中最重要的 D 个正交基进行保留，而在这 D 个正交基坐标轴上的坐标值就是原有样本映射到低维后的坐标。设原始样本矩阵为 $X=\begin{bmatrix}x_1 & x_2 & \cdots & x_N\end{bmatrix}^{\mathrm{T}}$，其中 $x_i \in R^M$，M 是样本的维度，$i=1,2,\cdots,N$，矩阵 X 是一个 N 行 M 列矩阵。

下面是 PCA 降维的详细步骤：

1) 原始样本中心化，形成矩阵 Y。Y 中每个元素为：

$$y_{ij}=x_{ij}-\overline{x}_j \in Y, \tag{1}$$

其中 $\overline{x}_j$ 是样本各维分量的平均值，$j=1,2,\cdots,M$，矩阵 Y 是 N 行 M 列矩阵。

2) 计算协方差矩阵 R：

$$R=\frac{1}{N-1}Y^{\mathrm{T}}Y, \tag{2}$$

其中 Y^{T} 是 Y 的转置矩阵，矩阵 R 是 M 行 M 列矩阵。

3) 求协方差矩阵 R 的特征值和特征向量，并将特征值按大小进行排序得：

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M \geq 0, \tag{3}$$

排序之后的特征值对应的特征向量为：

$$e_1, e_2, \cdots, e_M, \tag{4}$$

其中 $e_j \in R^M$，$j=1,2,\cdots,M$。

4) 计算特征值累计贡献率 η，

$$\eta=\frac{\sum_{i=1}^{D}\lambda_i}{\sum_{i=1}^{M}\lambda_i}\times 100\%, \tag{5}$$

其中 $0 \leq \eta \leq 100\%$，$0 \leq D \leq M$，D 为正整数。选取使得特征值累计贡献率大于 η 时的最小整数 D，这时将前 D 个较大特征值对应的特征向量作为标准正交基矩阵 E，

$$E=\begin{bmatrix}e_1 & e_2 & \cdots & e_D\end{bmatrix}, \tag{6}$$

矩阵 E 是 M 行 D 列矩阵。原始样本在标准正交基下的投影，即降维过后的样本矩阵 X^* 为：

$$X^* = XE, \tag{7}$$

矩阵 X^* 是 N 行 D 列矩阵，原始样本便从 M 维降至 D 维。

3. K 边界近邻法支持向量预选取方法简介

K 边界近邻法支持向量预选取的基本思想是通过选取距离每个样本最近的 k 个异类样本来构造支持向量候选集。已知训练样本集分为两类，正类样本 T_1 和负类样本 T_2，

$$T_1 = \left\{x_1^+, x_2^+, \cdots, x_{N_1}^+\right\}, \tag{8}$$

$$T_2 = \left\{x_1^-, x_2^-, \cdots, x_{N_2}^-\right\}, \tag{9}$$

其中 $x_i^+, x_j^- \in R^M$，$i = 1,2,\cdots,N_1$，$j = 1,2,\cdots,N_2$，N_1 是正类样本的个数，N_2 是负类样本的个数，M 是样本的维度。

3.1. 样本距离

当两类样本线性可分时，两类样本之间距离可用欧氏距离来表示：

$$d\left(x_i^+, x_j^-\right) = \left\|x_i^+ - x_j^-\right\|_2, \tag{10}$$

当两类样本非线性可分时，通过映射函数 $\phi(\cdot)$ 将原输入空间映射到高维的特征空间中，样本在高维空间中变得线性可分，这时的样本距离被称为非线性距离：

$$d\left(x_i^+, x_j^-\right) = \sqrt{K\left(x_i^+, x_i^+\right) - 2K\left(x_i^+, x_j^-\right) + K\left(x_j^-, x_j^-\right)}, \tag{11}$$

其中 $K(\cdot,\cdot)$ 是核函数。高斯核函数是一种常用的核函数：

$$K\left(x_i^+, x_j^-\right) = \exp\left(-\left\|x_i^+ - x_j^-\right\|^2 / 2\sigma^2\right), \tag{12}$$

其中 σ 是一个常数，采用高斯核函数时，非线性距离变为：

$$d\left(x_i^+, x_j^-\right) = \sqrt{2 - 2K\left(x_i^+, x_j^-\right)}。 \tag{13}$$

3.2. KNBN 支持向量预选取方法

下面是 KNBN 方法的具体步骤：

1) 从正类样本中选择一个样本，求其与所有负类样本之间的距离，保留最近的 k 个负类样本，将他们放入边界向量集当中。

2) 返回步骤(1)，直至遍历所有的正类样本截止。

3) 将所有负类样本按照步骤(1)和步骤(2)操作，保留离每个负类样本最近的 k 个正类样本，将他们也放入支持向量候选集当中。

4) 把上面得到的边界向量集当中的相同样本删去，进行唯一化处理，最终得到支持向量候选集。

如图 1 所示，其中有两类样本，当 $k = 4$ 时使用 KNBN 支持向量预选取方法得到边界向量集，适当选取 k 值，边界向量集一定能包含所有的支持向量，这样便构造出一个支持向量候选集。

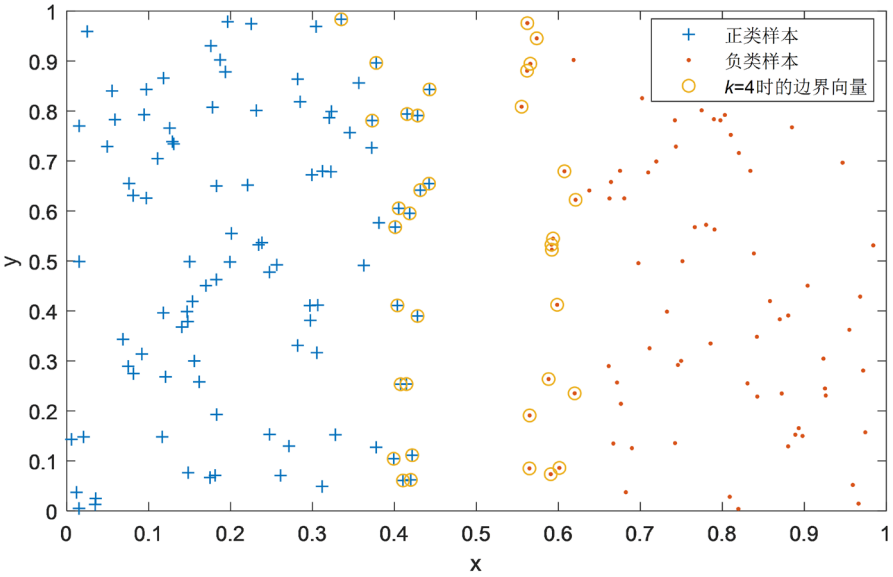

Figure 1. Diagram of KNBN support vector preselection
图 1. KNBN 支持向量预选取示意图

4. 基于 PCA 和 KNBN 的 SVM 预处理方法

4.1. 基于 PCA 和 KNBN 的 SVM 预处理方法介绍

根据 PCA 降维的特征和 KNBN 支持向量预选取的特性，本文提出基于 PCA 和 KNBN 的 SVM 预处理方法。为了尽可能保存数据的原有结构信息，所以该数据预处理方法先将训练数据集进行 PCA 降维处理，再把降维过后的训练数据集进行 KNBN 支持向量预选取，最终得到一个 SVM 大规模数据集的约简集。本文所提算法的流程图如图 2 所示。

下面是基于 PCA 和 KNBN 的 SVM 预处理算法的具体描述。

算法：基于 PCA 和 KNBN 的 SVM 预处理算法。

输入：原始训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i\in R^M$，$y_i\in\{-1,1\}$，$i=1,2,\cdots,N$，PCA 的特征值累计贡献率 η，KNBN 的参数值 k。

输出：原始训练数据集经 PCA 和 KNBN 处理过后的约简集 S。

Step 1. 将原始训练数据集 T 进行 PCA 降维。

1) 将 T 去掉类别特征，形成训练样本矩阵 X。

2) 根据 X 构建中心化矩阵 Y，Y 中的每个元素如式(1)所示。

3) 按照式(2)构建协方差矩阵 R。

4) 求协方差矩阵 R 的特征值及对应的特征项量，并对特征值按照从大到小的顺序进行排序。

5) 按照公式(5)计算特征值累计贡献率 η，确定 D 值，将这 D 个最大特征值对应的特征向量构成新基矩阵 E，将训练样本矩阵 X 乘以 E，即可得到降维后训练样本样本矩阵 X^*，添上正负类别属性后形成降维后的训练数据集 T^*。

Step 2. 在降维后的训练数据集 T^* 进行 KNBN 支持向量预选取。

1) 将数据集 T^* 分为正类数据集 T_1 和负类数据集 T_2。

2) 从 T_1 中选取一个样本，求其与所有负类样本间的距离，保留与其最近的 k 个负类样本作为在正类样本约束下的负类边界向量，直至遍历 T_1，形成负类边界向量集 S_2。

3) 同样，从 T_2 中选取一个样本，求其与所有正类样本间的距离，保留与其最近的 k 个正类样本作为在负类样本约束下的正类边界向量，直至遍历 T_2，形成正类边界向量集 S_1。

4) 将 S_1 和 S_2 合并，$S^* = S_1 \bigcup S_2$，并对 S^* 进行唯一化处理，即删除相同的样本，可得到 T^* 的支持向量候选集为 S，即得到原始训练数据集 T 的约简集 S。

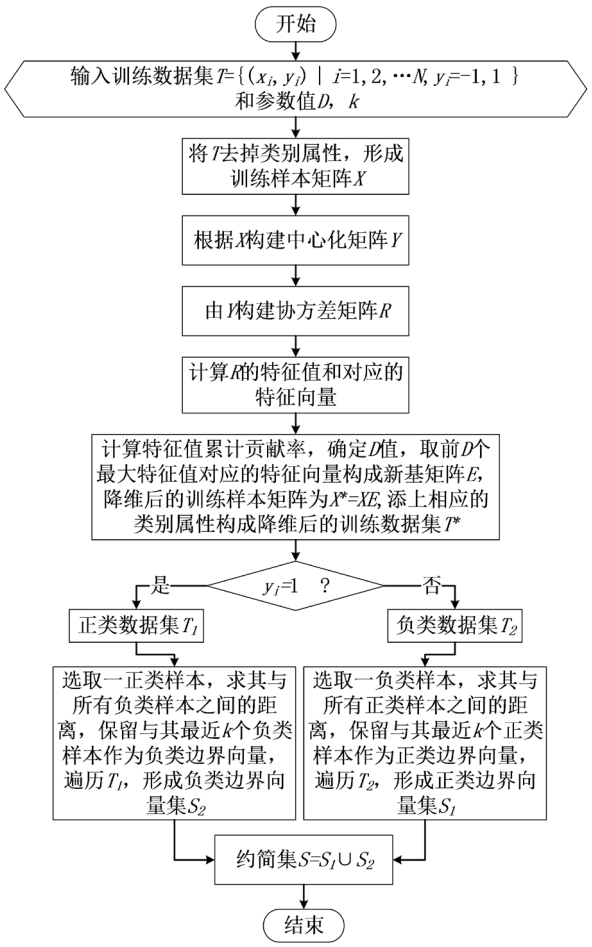

Figure 2. Flow chart of SVM preprocessing method based on PCA and KNBN
图 2. 基于 PCA 和 KNBN 的 SVM 预处理方法流程图

4.2. 算法复杂度分析

设原始训练样本数量为 N，样本维度为 M，$M < N$。经 PCA 降维后样本维度为 D，$D < M$，样本数量不变，其中正类样本数量为 N_1，负类样本数量为 N_2，$N_1 + N_2 = N$；降维后的训练样本经 KNBN 支持向量预选取之后得到的约简集中样本数量为 l，一般情况下，$l \ll N$。基于 PCA 和 KNBN 的 SVM 预处理算法复杂度是由 PCA，KNBN 和约简集进行 SVM 训练的时间复杂度共同决定的。

1) 该预处理方法 Step 1 的 PCA 降维过程中，去掉类别特征形成训练样本矩阵的运算次数可以忽略不计，构建中心化矩阵时需要计算 $2MN$ 次，计算协方差矩阵时需要运算 M^2N 次，协方差矩阵的特征值分解时需要 M^3 次运算[14]，构建新基矩阵需要将特征值按大小进行排序，排序[15]时需要 $M\log M$ 次运算，计算累计贡献率 η 和确定 D 值需要 $(M+3D-2)$ 次运算，构造降维后的新矩阵需要进行 $(2DMN-DN)$ 次运算。综合上述过程，PCA 的总运算次数为：

$$M^2N+2DMN+2MN-DN+M^3+M\log M+M+3D-2, \tag{14}$$

其中 $D<M$，则 PCA 的时间复杂度为 $O\left(M^2N\right)$。

2) 该预处理方法 Step 2 的 KNBN 支持向量预选取过程中，正负类数据集的分开需要进行 N 次运算，构造边界向量集这一过程中，对一个正类样本求其与所有负类样本之间距离，通常情况下需使用非线性距离，其中核函数采用高斯核函数，则需要 $(3D+9)N_2$ 次运算，再将求得的 N_2 个距离进行排序时需要 $N_2\log N_2$ 次运算，直至遍历所有正类样本后截止，则构造负类样本边界向量集需要运算 $\left(N_1N_2\log N_2+3DN_2+9N_2\right)$ 次，按照上述过程在构造正类样本边界向量集时则需要运算 $\left(N_1N_2\log N_1+3DN_1+9N_1\right)$次，下面唯一化处理过程的运算次数可忽略不计。KNBN 过程总运算次数为：

$$N_1N_2\log N_1N_2+(3D+9)(N_1+N_2), \tag{15}$$

其中 $N_1+N_2=N$，$N_1N_2\le N^2/4$，则 KNBN 的时间复杂度为 $O\left(N^2\log N\right)$。

3) 一般情况下，标准 SVM 的时间复杂度[16]是 $O\left(N^3\right)$。原始训练样本经 PCA-KNBN 预处理后的得到的约简集有 l 个样本，所以约简集进行 SVM 训练的时间复杂度为 $O\left(l^3\right)$。

综上所得，基于 PCA 和 KNBN 的 SVM 预处理算法复杂度为 $O\left(N^2\log N+M^2N+l^3\right)$，所以在某种情况下，如原始训练数据集经过 PCA-KNBN 处理过后的约简集规模很小时，则 SVM 训练的时间是可以缩短的。

5. 与现有方法的数值实验

PCA-KNBN-SVM 表示基于 PCA 和 KNBN 预处理的标准 SVM，PCA-SVM [8]表示只进行 PCA 降维预处理的标准 SVM，KNBN-SVM [13]表示只进行 KNBN 支持向量预选取预处理的标准 SVM。为了验证基于 PCA 和 KNBN 的 SVM 预处理算法的有效性，将 PCA-KNBN-SVM 与 PCA-SVM，KNBN-SVM 和无数据预处理的标准 SVM 进行比较。实验采用 Matlab R2018a，在 2.3 GHz，Pentium，Dual CPU，4 GB 内存的硬件平台上进行。SVM 训练选用 Libsvm-3.24 函数包，其中核函数采用高斯核函数，取 $\sigma=1.3$。

5.1. 数据介绍

实验采用 UCI 数据库中的 Polish companies bankruptcy data 数据集[17]。该数据集有 5 个适合二分类的数据样本，分别是 1 year，2 year，3 year，4 year 和 5 year 数据，每个数据都有 64 个特征属性和 1 个类别属性。每个数据集的训练样本数量和测试样本数量具体情况如表 1 所示。

Table 1. Polish company bankruptcy data set
表 1. 波兰公司破产数据集

数据集	训练样本/个	测试样本/个
1 year	2800	4200
2 year	3500	6500
3 year	4000	6500
4 year	3000	6700
5 year	3000	2900

5.2. 数值实验

本文分别采用无预处理的标准 SVM，PCA-SVM，KNBN-SVM 和 PCA-KNBN-SVM 对 5 个数据集进行训练和测试，其中 KNBN 算法中使用的样本距离为非线性距离式(12)所示，其中核函数参数值 $\sigma=1.3$。PCA 降维过程中，特征值累计贡献率设置为 $\eta=99.5\%$，由此确定的 5 个数据集中的 PCA 参数值 D 如表 2 所示。

KNBN 支持向量预选取的参数值[13]设置为$k=4$。

Table 2. Numerical experiment parameter value list
表 2. 数值实验参数值列表

数据集	1 year	2 year	3 year	4 year	5 year
D	5	5	7	5	4

实验结果记录的是各算法对应各数据集的训练时间和预测准确度。其中训练时间包括数据预处理时间和 SVM 训练的时间。数值实验结果如表 3 所示，其中数值结果是 100 次独立数值实验结果的平均值。

Table 3. List of numerical experiment results
表 3. 数值实验结果列表

数据集	算法	训练时间/s	分类准确度/%
1 year	SVM	6.7661	96.2143
	PCA-SVM	3.3170	96.2381
	KNBN-SVM	3.6625	96.0810
	PCA-KNBN-SVM	1.3604	96.1542
2 year	SVM	11.3482	96.0831
	PCA-SVM	5.3114	96.1538
	KNBN-SVM	6.0901	96.0308
	PCA-KNBN-SVM	1.9551	94.0554
3 year	SVM	9.0499	95.2400
	PCA-SVM	4.3659	95.2877
	KNBN-SVM	4.9939	95.1385
	PCA-KNBN-SVM	1.4137	95.2154
4 year	SVM	6.6030	94.7493
	PCA-SVM	3.2719	94.8358
	KNBN-SVM	4.1297	94.7164
	PCA-KNBN-SVM	1.2211	94.7910
5 year	SVM	3.9666	93.1241
	PCA-SVM	2.0605	93.1724
	KNBN-SVM	2.9333	92.8345
	PCA-KNBN-SVM	0.8076	93.1310

从表 3 中可以观察出，各算法在训练时间和分类准确度方面具有差异性，下面对它们进行分析比较。

在训练时间方面，使用 PCA-SVM 算法对每个数据集进行实验，其训练时间比无数据预处理的标准 SVM 要低很多，平均低了 51%左右；KNBN-SVM 算法的训练时间相对于无数据预处理的标准 SVM 也下降了很多，平均下降 41%左右；本文提出的 PCA-KNBN-SVM 算法训练时间相对于无数据预处理的标准 SVM 平均下降了 86%左右，相对于 PCA-SVM 平均下降了 63%左右，相对于 KNBN-SVM 平均下降了 69%左右。

在分类准确度方面，PCA-SVM 算法要比无数据预处理的标准 SVM 稍高一些，平均高了 0.06%左右；

KNBN-SVM 算法要比无数据预处理的标准 SVM 偏低一些，平均低了 0.13%左右；PCA-KNBN-SVM 算法相对于 PCA-KNBN 平均下降了 0.15%左右，相对于 KNBN-SVM 平均升高了 0.18%左右，而相对于无数据预处理的标准 SVM 在 1 year，2 year 和 3 year 数据中有些许降低，但在 4 year 和 5 year 数据有些许升高。

综合来看，PCA-KNBN-SVM 算法相对于无预处理的 SVM，PCA-SVM 和 KNBN-SVM，其训练时间缩短效果很明显；在分类准确度方面相对于 PCA-SVM 有所下降，相对于 KNBN-SVM 却有所提高，而相对于无预处理的 SVM 在某些数据中分类准确度稍高或稍低一些，总而言之，PCA-KNBN-SVM 算法相对于其他算法在训练时间方面缩短效果很明显，而在分类准确度方面变化不大。

6. 结语

本文为了解决 SVM 在遇到大规模的数据集时出现训练时间增多和分类精度可能下降的问题，提出了一种基于 PCA 和 KNBN 的 SVM 预处理方法。通过选取 UCI 数据库中的 Polish companies bankruptcy data 数据集进行数值实验，从数据实验结果观察得出在该预处理方法下的 SVM 相对于无预处理的 SVM，只进行 PCA 预处理的 SVM 和只进行 KNBN 预处理的 SVM 有大幅度缩减训练时间的效果，在分类准确度方面相对于其他算法变化不大。结果表明基于 PCA 和 KNBN 的 SVM 预处理方法是可以在保持良好分类精度下提高训练速度的一种 SVM 预处理好方法。

参考文献

[1] Vapnik, V. (1998) Statistical Learning Theory. Vol. 3, Chapter 10-11, Wiley, New York, 401-492.

[2] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 121-139, 298-300.

[3] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 第 7 章, 95-135.

[4] Qin, J. and He, Z.S. (2005) A SVM Face Recognition Method Based on Gabor-Featured Key Points. *Proceedings of* 2005 *International Conference on Machine Learning and Cybernetics*, **8**, 5144-5149. https://doi.org/10.1109/ICMLC.2005.1527850

[5] Sun, A., Lim, E.P. and Ng, W.K. (2002) Web Classification Using Support Vector Machine. *Proceedings of the* 4*th International Workshop on Web Information and Data Management*, McLean, Virginia, November 2002, 96-99. https://doi.org/10.1145/584931.584952

[6] 余金澳, 吴彦鸿. 一种面向方位敏感性的 PCA-SVM 分类识别方法[J]. 无线电工程, 2018, 48(2): 83-87.

[7] Aicha, A.B. (2018) Noninvasive Detection of Potentially Precancerous Lesions of Vocal Fold Based on Glottal Wave Signal and SVM Approaches. *Procedia Computer Science*, **126**, 586-595. https://doi.org/10.1016/j.procs.2018.07.293

[8] 汪雯琦, 高广阔. 基于 PCA 和 SVM 分类的跨年龄人脸识别[J]. 计算机时代, 2019(7): 1-4+8.

[9] Wang, C., Hu, Z.L., Pang, Q. and Hua, L. (2019) Research on the Classification Algorithm and Operation Parameters Optimization of the System for Separating Non-Ferrous Metals from End-of-Life Vehicles Based on Machine Vision. *Waste Management*, **100**, 10-17. https://doi.org/10.1016/j.wasman.2019.08.043

[10] Zhang, L., *et al.* (2008) Support Vectors Pre-Extracting for Support Vector Machine Based on K Nearest Neighbour Method. *Proceedings of IEEE International Conference on Information and Automation*, Zhangjiajie, 20-23 June 2008, 1353-1358.

[11] 徐红敏, 王若鹏, 张怀念. 支持向量机的快速分类算法[J]. 北京石油化工学院学报, 2009, 17(4): 55-58.

[12] 胡志军, 王鸿斌, 张惠斌. 基于距离排序的快速支持向量机分类算法[J]. 计算机应用与软件, 2013, 30(4): 85-87+100.

[13] 李庆, 胡捍英. 支持向量预选取的 K 边界近邻法[J]. 电路与系统学报, 2013, 18(2): 91-96.

[14] 万静, 吴凡, 何云斌, 李松. 新的降维标准下的高维数据聚类算法[J]. 计算机科学与探索, 2020, 14(1): 96-107.

[15] 陆微微, 刘晶. 一种提高 K-近邻算法效率的新算法[J]. 计算机工程与应用, 2008, 44(4): 163-165+178.

[16] Tsang, I.W., Kwok, J.T. and Cheung, P.-M. (2005) Core Vector Machines: Fast SVM Training on Very Large Data Sets. *The Journal of Machine Learning Research*, **6**, 363-392.

[17] Tomczak, S. Polish Companies Bankruptcy Data. Data Set. http://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data, 2020-09-25.