Advances in Applied Mathematics 应用数学进展, 2021, 10(12), 4292-4303
Published Online December 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1012456

Hans汉斯

SVM和RLR算法的对比分析

韩蓓丽

浙江师范大学数学与计算机科学学院，浙江 金华

收稿日期：2021年11月16日；录用日期：2021年12月13日；发布日期：2021年12月20日

摘　要

在机器学习领域中，支持向量机(SVM)和逻辑回归(RLR)作为两种有监督的分类算法，在不同场合下有着不同的分类效果。本文旨在通过数据分析对二者进行比较。数值实验结果显示：随着数据样本维度的增加，SVM的预测准确率、稳定性、计算时间及计算资源占用情况比RLR更好；对存在离群值的样本数据分类时，SVM在稳定性和分类效果方面表现更佳；在高维小样本中，RLR预测准确率比SVM更高，表现更佳。

关键词

SVM，RLR，维度，离群值，稳定性，计算复杂度

Comparative Analysis of SVM and RLR Algorithms

Beili Han

School of Mathematics and Computer Science, Zhejiang Normal University, Jinhua Zhejiang

Received: Nov. 16th, 2021; accepted: Dec. 13th, 2021; published: Dec. 20th, 2021

Abstract

In the field of machine learning, support vector machine (SVM) and logistic regression (RLR), as two supervised classification algorithms, have different classification effects in different situations. This paper aims to compare them through data analysis. The numerical experiment results show that, with the increase of data sample dimension, SVM has better prediction accuracy, stability, calculation time and calculation resource occupation than RLR. When classifying the sample data with outliers, SVM performs better in terms of stability and classification effect. In high-dimensional small samples, RLR has higher prediction accuracy and better performance than SVM.

文章引用: 韩蓓丽. SVM 和 RLR 算法的对比分析[J]. 应用数学进展, 2021, 10(12): 4292-4303.
DOI: 10.12677/aam.2021.1012456

Keywords

SVM, RLR, Dimension, Outliers, Stability, Computational Complexity

Copyright © 2021 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

机器学习中的分类算法属于有监督的学习算法。在理论上主要的区别是，支持向量机(Support Vector Machine, SVM [1])采用 hinge 损失函数，学习分类决策边界；正则逻辑回归(Regularization Logistic Regression, RLR [2])采用对数损失函数，学习某一类的条件概率。在应用方面的主要区别是 SVM 硬间隔分类器表现更佳，RLR 软间隔分类器表现更佳。对于 SVM 和 RLR 两种分类算法，目前已有一些理论研究和应用实践。

对于 SVM 分类算法，在 2007 年 Steinwart 和 Scovel 给出带有 hinge 损失的 SVM 分类算法。2009 年和 2011 年向道红相继在基于高斯和凸损失分类[3]研究中，给出 hinge 凸损失函数于再生核希尔伯特空间中 SVM 分类器的学习率，得出该学习率与空间维度和样本量有关[4]。在 2010 年 Jeongyounahn [5]研究了 SVM 在高维小样本数据中产生的大量数据堆积问题，并在最大数据堆积方向、SVM、DWD、RLR 上进行了平均误分率的应用研究。而在 2011 年，Yufeng Liu [6]等人在硬或软分类器的应用探究中，给出 SVM 硬分类器倾向于运用靠近分界线的数据点产生分类决策边界而不产生概率估计。对于 RLR 分类算法，在 2010 年向道红对于无零点的 logistic 损失函数相关分类算法做出误差分析[7]，通过投影算子考虑无界情况，给出 RLR 分类器的学习速率，得到该学习率与空间维度和样本量有关。同年在数据堆积应用研究中，给出 RLR 分类算法的平均误分率随着维度变化的走势。而对于两者的直接对比，目前尚未有成熟的应用研究，本文将对两种分类算法进行比较研究，得到在不同情况下两种分类算法的表现能力。

本文分为三部分，第一部分介绍 hinge 损失函数和 SVM 分类算法原理。第二部分介绍 logistic 损失函数和 RLR 分类算法原理。第三部分进行仿真实验，通过 PCA 降维，首先在 699 个不含有离群值的 2~9 维样本数据中，得到预测准确率走势图，比较两种分类算法的分类效果；其次在 1060 个含有离群值的数据中进行同样的实验，比较两种算法的鲁棒性；然后在阿尔兹海默症高维小样本数据集中，对两种算法的稳定性和分类效果进行对比；最后通过实验运行时间和占有内存，就算法复杂度做比对。

2. SVM 分类算法原理

2.1. SVM

早在 1995 年，Vapnik [8]就提出了支持向量机，相比于其他的分类算法，它有着比较高的泛化能力。一组给定的样本集 $V=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，$x$ 可以取连续型数值也可以取离散型数值，对应的 y_i 取值为±1。SVM 分类算法的目标在于基于训练样本集找到样本空间的划分超平面，把两种不同的类别分开。

定义超平面为：

$$w^{\mathrm{T}}x+b=0, \tag{1.1}$$

其中，上式(1.1)中 $w=(w_1,w_2,\cdots,w_d)^{\mathrm{T}}$ 代表法向量，决定了超平面的方向，b 代表截距，决定了超平面和原点之间的距离。这两项共同决定了超平面的位置，故可记为超平面 (w,b)。任意一个样本点到超平面的

距离为：

$$\gamma = \frac{\left|w^{\mathrm{T}}x + b\right|}{\|w\|}.$$

在样本空间中，若超平面(w,b)满足下列条件：

$$\begin{cases} w^{\mathrm{T}}x_i + b \geq +1, y_i = +1 \\ w^{\mathrm{T}}x_i + b \leq -1, y_i = -1 \end{cases} \tag{1.2}$$

则表明超平面(w,b)可正确分类样本。

见图 1 中，画圈的三个样本满足(1.2)式，因此被称为支持向量(support vector)。分类之后，两个分属不同类别的支持向量到超平面的距离之和称为间隔：

$$\gamma = \frac{2}{\|w\|}.$$

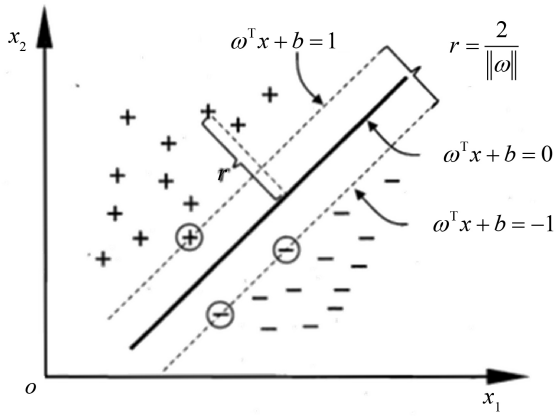

Figure 1. Support vector and interval
图 1. 支持向量与间隔

最优超平面既要正确分类，又要保证两类样本间的间隔达到最大，为最大间隔的分类超平面，也称为为软间隔，不需要所有样本都划分正确。故寻找最优分类超平面可归结为求解如下优化算法：

$$\begin{aligned} &\max_{w,b} \frac{2}{\|w\|} \\ &\text{s.t. } y_i\left(w^{\mathrm{T}}x_i + b\right) \geq 1, i = 1,2,\cdots,n \end{aligned} \tag{1.3}$$

线性可分支持向量机算法可转化为如下优化问题：

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 \\ &\text{s.t. } y_i\left(w^{\mathrm{T}}x_i + b\right) \geq 1, i = 1,2,\cdots,n \end{aligned} \tag{1.4}$$

2.2. hinge 损失函数

$$\phi_h\left(y\left(w\cdot x + b\right)\right) = \left[1 - y\left(w\cdot x + b\right)\right]_+ \tag{1.5}$$

上式为 hinge 损失函数(hinge loss function [2])，如图 2 所示，下标“+”表示取正值的函数，

$$[z]_+ = \begin{cases} z, z > 0 \\ 0, z \le 0 \end{cases}$$

这意味着当样本点被正确分类时，损失为 0，否则损失为$1-y_i(w\cdot x_i+b)$。

见图 2 可以看到 hinge 损失相对于 logistic 损失的特点是有零点，在 1 处不可导。

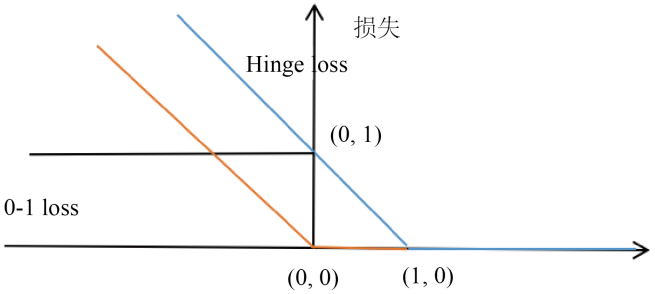

Figure 2. Hinge loss function
图 2. hinge 损失函数

2.3. SVM 分类算法

令 $f(x_i)=w\cdot x+b$ 为通过训练数据学习出的线性分类器，且为了克服过拟合，加入正则项。假设所有函数均定义在再生核希尔伯特空间中，$f\in\mathcal{H}_k$，k 代表核函数。

则带有 hinge 损失的 SVM 分类器为：

$$f_z = \arg\min_{f\in\mathcal{H}_k}\left\{\frac{1}{m}\sum_{i=1}^{m}\left(1-y_i f(x_i)\right)_+ + \lambda\|f\|_k^2\right\} \tag{1.6}$$

f_z 表示决策函数；λ 表示正则化参数。

3. RLR 分类算法原理

3.1. 逻辑回归

逻辑回归虽然名为回归，其实是一种分类学习方法，主要用于二分类问题。该分类模型用条件概率分布 $P(Y|X)$ 表示，

$$P(Y=1|x)=\frac{\exp(w\cdot x+b)}{1+\exp(w\cdot x+b)},\quad P(Y=0|x)=\frac{1}{1+\exp(w\cdot x+b)}$$

其中，$w=\left(w^{(1)},w^{(2)},\cdots,w^{(n)},b\right)^{\mathrm{T}}$ 称为权值向量，b 为偏置，$x=\left(x^{(1)},x^{(2)},\cdots,x^{(n)},1\right)^{\mathrm{T}}$ 为输入的随机变量。当 w 和 x 的内积取值接近正无穷则概率值为 1，反之为 0，为线性函数模型。w 是需要通过训练数据学习的参数。

因乳腺癌数据在低维空间线性不可分，则需要对原始数据进行多项式变换，加入高次项，来解决决策边界非常复杂的情形，使之非线性可分。

令 $h(x;\theta)=b+w_1x_1+w_2x_2+w_3x_1^2+w_4x_2^2+w_5x_1x_2+w_6x_1x_2^2+\cdots$

$$h(x;\theta)=f\left(\theta^{\mathrm{T}}x\right),$$

Logistic 分布函数(同 sigmoid 函数)如图 3，形式为：

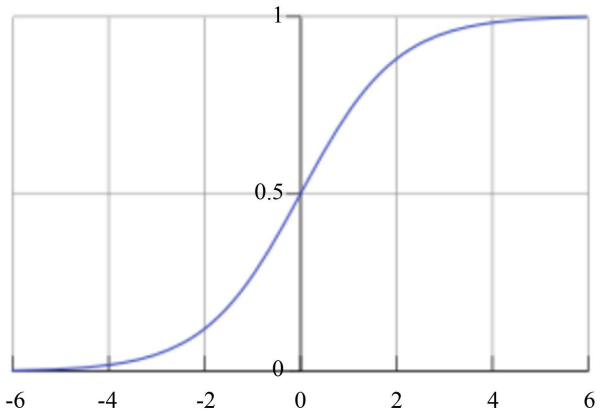

Figure 3. Sigmoid function image
图 3. Sigmoid 函数图像

$$f(z)=\frac{1}{1+\mathrm{e}^{-z}}$$

通过函数 $f(z)$ 对 $\theta^{\mathrm{T}}x$ 进行变换，将取值挤压到[0, 1]区间内：

$$f\left(\theta^{\mathrm{T}}x\right)=\frac{1}{1+\mathrm{e}^{-\theta^{\mathrm{T}}x}}.$$

θ^{T} 表示参数集。

$$h(x;\theta)=f\left(\theta^{\mathrm{T}}x\right),$$

当 $\theta^{\mathrm{T}}x\geq 0$ 时，$h(x;\theta)\geq 0.5$，则样本为正例，输出 1；反之，输出为 0。

3.2. logistic 损失函数

$$\phi_l\left(-yf(x)\right)=\log\left(1+\exp\left(-yf(x)\right)\right)$$

如图 4 可以看到 logistic 损失相对于 hinge 损失的特点是没有零点，在 1 处可微。

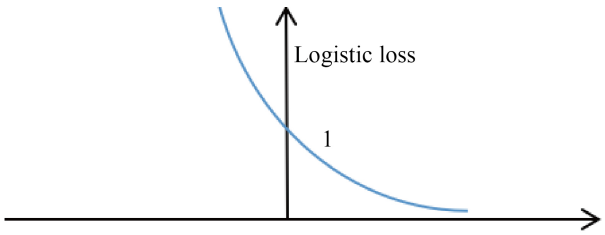

Figure 4. Logistic loss
图 4. logistic 损失

3.3. RLR 分类算法

由于 LR 分类算法存在过拟合，因此在逻辑回归中，引入正则化项。通常使用 L_1 正则化。加入 L_1 正则项容易得到稀疏解(The sparse solution [4])。也可使用 L_2 正则项。引入正则化参数 $\lambda>0$，假设所有函数均定义在再生核希尔伯特空间中，$f\in\mathcal{H}_k$，k 代表核函数。

则带有 logistic 损失的 RLR 分类器为：

$$f_z = \arg\min_{f\in\mathcal{H}_k}\left\{\frac{1}{m}\sum_{i=1}^{m}\log\left(1+\exp\left(-y_i f\left(x_i\right)\right)\right)+\lambda\left\|f\right\|_k^2\right\}$$

4. 仿真实验

4.1. 数据统计描述

实验环境 python3.0，sklearn。

采用南斯拉夫卢布尔雅那大学医疗中心肿瘤研究所提供的威斯康辛乳腺癌数据集，用 python3 中的 strategy 参数操作，将缺失值用该列的均值替换，得到可以完整可进行实验的数据。部分数据如表 1 所示：

Table 1. Partial data of breast cancer dataset
表 1. 乳腺癌数据集部分数据

Mean_radius	Mean_texture	Mean_perimeter	Mean_area	Mean_smoothness
17.99	10.38	122.8	1001	0.1184
20.57	17.77	132.9	1326	0.08474
19.69	21.25	130	1203	0.1096
11.42	20.38	77.58	386.1	0.1425
20.29	14.34	135.1	1297	0.1003
12.45	15.7	82.57	477.1	0.1278
18.25	19.98	119.6	1040	0.09463
13.71	20.83	90.2	577.9	0.1189
13	21.82	87.5	519.8	0.1273

见图 5 数据中各属性的二分类数据特征用面积图展示，两两之间的相关性用散点图展示。可以看出，mean_radius 和 mean_perimeter 之间有明显的线性相关关系。

数据集共有 1060 行数据，30 个属性，分 2 大类。由于一部分属性具有相关性，运用 PCA (主成分分析，principal component analysis)进行特征提取，最大程度保留原始信息。从原先的 30 个维度，分别降低到 2~9 个互不相关的维度。最后一列数据为输出类，出于与原数据区分的目的，将 0 更改为 2 类和 1 更改为 4 类。9 维数据统计描述如表 2。

部分数据统计后，画出箱式数据分布图，见图 6，其中圆圈表示离群值。

4.2. 在乳腺癌数据集上的分类效果比较

图 6 中可观察得到，该数据集中含有离群值，为了比较在离群值的影响下，SVM 和 RLR 两种分类算法的鲁棒性和分类效果，进行了如下实验。

1) 删除离群值数据集

将离群值删除后，得到 699 个样本数据，进行十折交叉验证，选取 2 组作为测试集，剩余 8 组作为训练集，代入两种分类算法，计算得到预测准确率。为消除选取测试集带来的偶然性，每次选取不同于上一次的 2 组，一共进行 45 次实验。

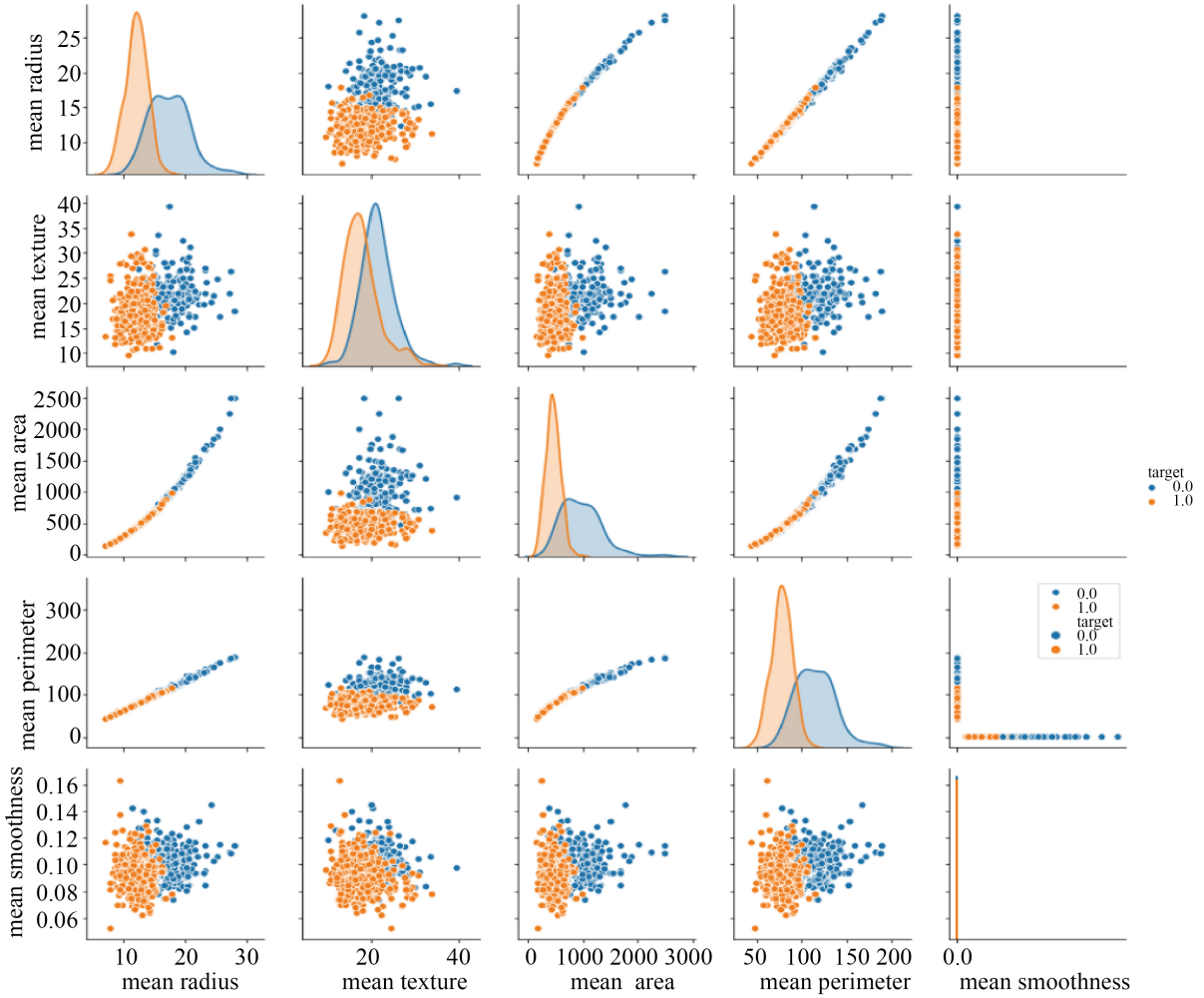

Figure 5. Scatter plot of partial attribute correlation
图 5. 部分属性相关性散点图

Table 2. Statistical description of PCA dimensionality reduction to 9-dimensional data
表 2. PCA 降维到 9 维数据统计描述

(a)

	C_D	C_T	U_C_Si	U_C_Sh	M_A
count	1.060000e+03	1060	1060	1060	1060
mean	1.012428e+06	64.22	63.37	63.42	63.15
std	5.077446e+05	103.14	103.63	103.60	103.76
min	6.163400e+04	0.00	0.00	0.00	0.00
25%	8.975498e+05	3.00	1.00	1.00	1.00
50%	8.978145e+05	6.00	5.00	5.00	4.00
75%	1.200180e+06	95.25	95.25	95.25	95.25
max	1.345435e+07	360.00	360.00	360.00	360.00

(b)

	S_E_C_S	B_C	N_N	M	Class
count	1060	1060	1060	1060	1060
mean	63.42	63.57	63.19	62.35	2.77
std	103.59	103.51	103.74	104.21	0.97
min	0.00	0.00	0.00	0.00	0.00
25%	2.00	2.00	1.00	1.00	2.00
50%	4.00	5.00	4.00	1.00	2.00
75%	95.25	95.25	95.25	95.25	4.00
max	360.00	360.00	360.00	360.00	4.00

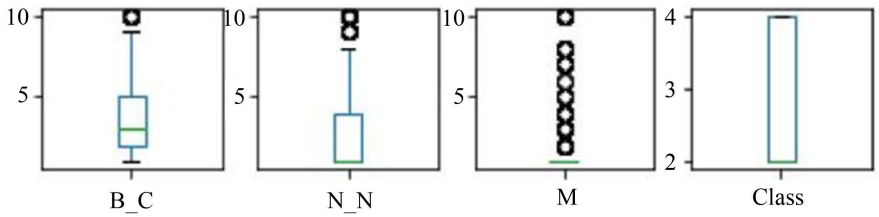

Figure 6. Nine-dimensional data distribution diagram
图 6. 九维部分数据分布图

在 2~9 维数据集中，分别训练数据。可以得到各分类算法预测准确率的箱型图。仅展示 6 维各分类算法预测准确率箱型图。

见图 7，可看出带有 l_1 正则项的 RLR 分类算法准确率平均值为 0.955，带有 l_2 正则项的 RLR 分类算法准确率平均值为 0.961，SVM 分类算法准确率平均值为 0.97。

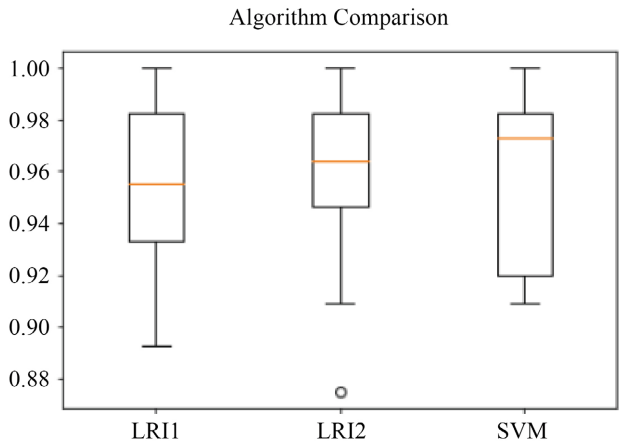

Figure 7. Box plot of the test accuracy of each six-dimensional classification algorithm
图 7. 六维各分类算法测试准确率箱线图

将 8 次训练得到的准确率平均值，以折线图展示如图 8：

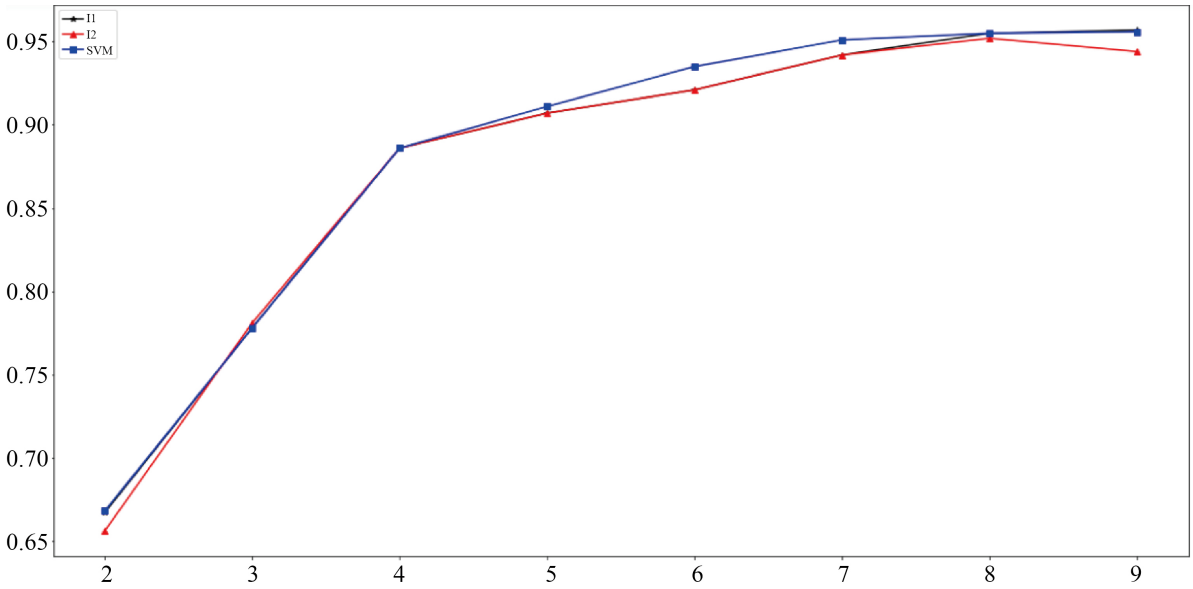

Figure 8. The prediction accuracy of each classification algorithm increases with the dimensionality line chart
图 8. 各分类算法预测准确率随维度增加折线图

从图 8 中可以看出，在低维空间中两种分类算法的预测准确率几乎一致。随着维度的增加，分类算法的预测准确率均逐渐升高，且 SVM 算法呈现比较好的分类效果。

1) 含有离群值数据集

运用带有离群值的 1060 个样本进行同样的实验，可以得到两种分类算法预测准确率随维度增加的折线图。

通过对比图 8 和图 9，可以得到 SVM 算法在离群值的影响下，依旧保持很好的分类效果，而 RLR 则受到一定程度的影响，因此，在乳腺癌数据集中，SVM 分类算法更具有鲁棒性。

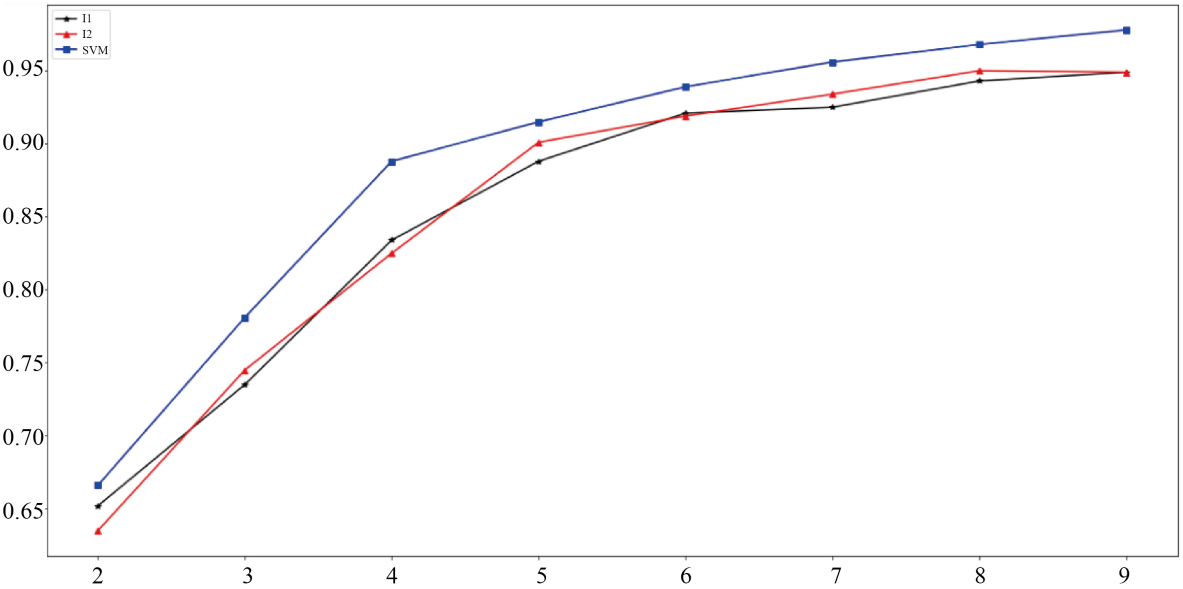

Figure 9. The prediction accuracy of each classification algorithm increases with the dimensionality line chart
图 9. 各分类算法预测准确率随维度增加折线图

4.3. 高维小样本数据集上的分类效果比较

由于乳腺癌数据集不属于高维小样本数据，因此采用来自天津大学医学部发布的阿尔兹海默症数据集，有 111 个样本数据，其中每个数据有 11,340 个特征变量，部分数据展示如表 3：

Table 3. Partial data of Alzheimer's disease data set
表 3. 阿尔兹海默症数据集部分数据

Left_corticospinal	Right_corticospinal	Left_IFOF	Right_IFOF	Left_uncinate	Right_uncinate
69.3	69	72	71.3	78	56
72	73.5	58.4	74.2	73.2	63
75.4	66.7	81	77.6	58.9	78.6
70	72.4	73.2	58.2	68.5	72
78.1	75.3	65.4	65.5	60	67

由于维度过大，只展示部分维度数据统计描述，见表 4：

Table 4. Statistical description of some data of Alzheimer's disease
表 4. 阿尔兹海默症部分数据统计描述

	0	1	2	3	4
count	111	111	111	111	111
mean	5717.12	557.98	198.63	5329.86	9658.26
std	1300.30	330.63	166.52	3743.59	5089.43
min	3402.60	0.00	0.00	0.00	0.00
25%	4769.10	313.82	97.15	2611.55	6544.85
50%	5531.90	513.01	132.62	4913.20	7762.70
75%	6406.25	761.32	226.36	6961.60	12001.5
max	10147.00	1799.30	924.61	21019.0	28525.0

对这一数据集进行训练，同样采用十折交叉验证，选择 2 组作为测试集，得到分类准确率，以箱线图展示：

图 10 中可以看出，带 l_1 惩罚的 RLR 分类算法的预测准确率浮动最大，最低为 0.8，最高为 0.93，平均值为 0.901；带 l_2 惩罚的 RLR 分类算法的预测准确率浮动较大，最低为 0.865，最高为 0.962，平均值为 0.903；SVM 分类算法的预测准确率一直在 0.89 附近。出现该结果的原因是，SVM 中采用 hinge 损失，会出现样本堆积，使得在样本量非常少的情况下，参与学习分类边界的数据点更加少，因此，在高维小样本数据中，SVM 表现会很差，而 RLR 呈现更加优秀的预测效果。

4.4. 分类算法复杂性比较

采用威斯康辛乳腺癌数据集，可以得到测试运行时间和占用内存，如表 5 中所示。

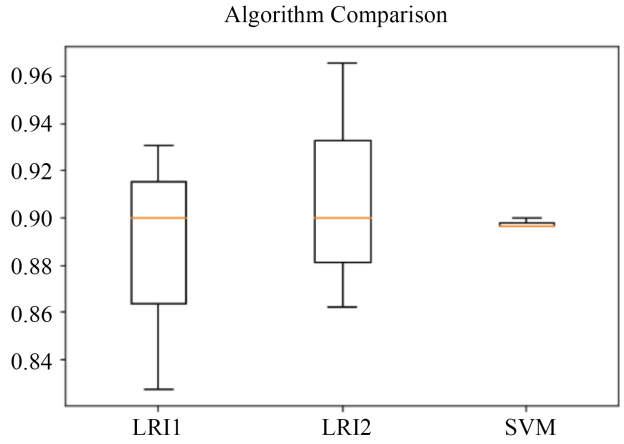

Figure 10. Box plot of prediction accuracy of each classification algorithm
图 10. 各分类算法预测准确率箱线图

表 5 中可以看出，在同等情况下，SVM 的运行时间长但占用内存较小，RLR 的运行时间短，但占有内存高。在低维大样本情况下，SVM 的表现更好，在高维小样本情况下，RLR 的表现更好。这与文献[9]中所呈现的结果有相似的表现。

Table 5. Algorithm complexity
表 5. 算法复杂性

算法	运行时间	占用内存	平均学习率
l_1-RLR	0.0019872188568115234	3.0517578125e−05MB	0.764
l_2-RLR	0.0019872188568115234	3.0517578125e−05MB	0.758
SVM	0.007014751434326172	2.288818359375e−05MB	0.805

5. 结论

在机器学习领域中，支持向量机和逻辑回归作为两种有监督的分类算法，各有千秋，其本质区别是损失函数的不同，SVM 采用 hinge 损失函数，RLR 采用对数损失函数，使得两者在不同的场合下有着不同的分类效果。以探究两者的区别和应用场景为出发点，进行仿真实验。在 2~9 维、699 个不含有离群值乳腺癌样本数据中，得到结论为随着维度的升高 SVM 的预测准确率逐渐高于 RLR；加入离群值数据，得到 1060 个样本，在有离群值的影响下，SVM 呈现稳定的性能，因此 SVM 分类算法更具鲁棒性；在 11,340 维、111 个样本的高维小样本阿尔兹海默症数据集中，进行交叉验证，发现 RLR 的稳定性更高，分类效果更好；在威斯康辛乳腺癌数据实践中，SVM 算法计算复杂度比较高，运行时间更长，占有内存较小，而 RLR 运行时间更短，占有内存较大。

参考文献

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[2] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[3] Xiang, D.H. (2011) Classification with Gaussians and Convex Loss II: Improving Error Bounds by Noise Conditions. *Science China Mathematics*, **54**, 165-171. https://doi.org/10.1007/s11425-010-4043-2

[4] Xiang, D.H. and Zhou, D.X. (2009) Classification with Gaussians and Convex Loss. *Journal of Machine Learning Research*, **10**, 1447-1468.

[5] Ahn, J. (2010) The Maximal Data Piling Direction for Discrimination. *Biometrika*, **97**, 254-259. https://doi.org/10.1093/biomet/asp084

[6] Liu, Y.F., Zhang, H.H. and Wu, Y.C. (2011) Hard or Soft Classification? Large-Margin Unified Machines. *Journal of the American Statistical Association*, **106**, 166-177. https://doi.org/10.1198/jasa.2011.tm10319

[7] Xiang, D.-H. (2010) Logistic Classification with Varying Gaussians. *Computers and Mathematics with Applications*, **61**, 397-407. https://doi.org/10.1016/j.camwa.2010.11.016

[8] 袁梅宇. 机器学习基础原理、算法与实践[M]. 北京: 清华大学出版社, 2018.

[9] 徐新红. 基于高维小样本数据和类别不平衡数据的反距离加权支持向量机[D]: [硕士学位论文]. 济南: 山东大学, 2020.