Computer Science and Application 计算机科学与应用, 2023, 13(7), 1420-1427
Published Online July 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.137140

Hans汉斯

基于机器学习的新冠疑似人员预测

朱韦宇，樊纪山*，江　阳，户彩凤，孙巧榆

江苏海洋大学电子工程学院，江苏　连云港

收稿日期：2023年6月18日；录用日期：2023年7月17日；发布日期：2023年7月26日

摘　要

自2020年以来，新冠疫情的迅速扩散，对全世界的社会经济和公共卫生造成了十分严重的影响。如何快速准确地诊断和预测潜在的新冠病例是当前亟需解决的问题。机器学习技术具有处理大量数据、提高预测准确率等优势，在疫情防控中有广泛应用。本文基于机器学习算法，对疑似新冠人员的预测进行研究。通过随机函数生成有效病例数据，并利用SVM (Support Vector Machine)分类模型对其进行训练，经过多次实验，我们发现该模型能够准确地预测疾病，而且具有较高的可靠性。

关键词

机器学习，新冠，预测，SVM分类模型

Machine Learning-Based Prediction of Suspected Coronavirus Personnel

Weiyu Zhu, Jishan Fan*, Yang Jiang, Caifeng Hu, Qiaoyu Sun

School of Electronic Engineering, Jiangsu Ocean University, Lianyungang Jiangsu

Received: Jun. 18th, 2023; accepted: Jul. 17th, 2023; published: Jul. 26th, 2023

Abstract

Since 2020, the epidemic situation in COVID-19 has spread rapidly, which has had a very serious impact on social economy and public health all over the world. How to diagnose and predict potential COVID-19 cases quickly and accurately is an urgent problem at present. Machine learning technology has the advantages of processing large amounts of data and improving prediction accuracy, and has been widely used in epidemic prevention and control. Based on the machine learning algorithm, this paper studies the prediction of people suspected of COVID-19. The valid

*通讯作者。

文章引用: 朱韦宇, 樊纪山, 江阳, 户彩凤, 孙巧榆. 基于机器学习的新冠疑似人员预测[J]. 计算机科学与应用, 2023, 13(7): 1420-1427. DOI: 10.12677/csa.2023.137140

case data is generated by random functions and trained and tested using the SVM (Support Vector Machine) classification model, and after many experiments, we found that the model can accurately predict the disease and has high reliability.

Keywords

Machine Learning, COVID-19, Prediction, SVM Classification Model

1. 引言

新冠病毒以惊人的速度传播，给世界各国的公共卫生和经济带来了严重威胁[1]。在疫情爆发初期，由于对病毒的认识不足，导致许多潜在病例未能及时被发现和隔离，给疫情的控制带来了极大的困难。随着疫情的发展和变化，针对新冠疫情的防控策略[2]也在不断地更新和改进。针对疫情而言，做好疑似新冠人员的筛查和诊断是抗疫的重要手段。但传统方法对疑似患者的筛查和诊断都会受到疫情爆发等不同程度的影响，效率或多或少都会下降，因此亟需新的方法来提高疑似病例的筛查和诊断的效率。

机器学习是一门以大量的数据为基础的前沿科学，其被广泛地运用于各个不同的领域[3] [4]，尤其是在医学和其他研究领域。Vaidya 和王晓丽就曾利用机器学习来研究和预测糖尿病和冠心病的相关信息[5] [6]，而且还被证明能够帮助更好地识别和治疗疾病。而在疫情防控中，机器学习技术也具有很大的应用前景，例如任建强[7]等基于机器学习的疫情三步预测模型 TSPMGML 预测未来确诊人数和实际感染规模。机器学习算法能够处理大量的数据，从而提高数据分析和预测的准确性，同时还能够自动地调整模型参数，以适应不同的数据集和任务类型。

因此，利用机器学习算法进行疫情预测，可以通过收集大量的疫情数据，发现潜在的疑似病例，并提供科学依据来制定疫情防控策略。本文基于机器学习的算法，采用分类模型中的支持向量机(SVM)分类模型，同时以十折交叉验证法作为验证方法，对疑似新冠的患者实施预测研究。我们利用随机函数生成数据，通过控制相关参数保证数据的合理性，利用分类模型进行训练和测试。

根据实验研究结果分析，此模型能够有效地识别分析疑似新冠患者，在实际中能根据分析结果实施有效措施控制疫情。

2. 支持向量机(SVM)的基本原理

2.1. 分类器的选择

在机器学习中，常用的分类模型有决策树、朴素贝叶斯、支持向量机(SVM)和神经网络等[8]。本文中，使用分类模型中的支持向量机(SVM)作为分类模型。

支持向量机(Support Vector Machine, SVM)成为一项具备强大自适应作用的机制，可以实现高效的数据挖掘和预测，它可以在保证精度的前提下，尽可能地减少实际数据的偏离，从而进一步提高建模的精度。支撑向量机(SVM)成为一项具备自适应作用的机器学习方法，在统计分类、回归等领域得到了普遍的应用。它的一项重要优势在于，可以实现对数据的自动调整，从而减少实际的输入偏离。

SVM，它的核心概念在于通过将复杂的二分类算法，如高维特征空间和一个超平面，有效地划分出

各种类型的信息，从而实现对复杂网络的有效分类，同时 SVM 具有良好的泛化能力和可靠性，适用于处理高维稀疏数据集，其原理图如图 1 所示。

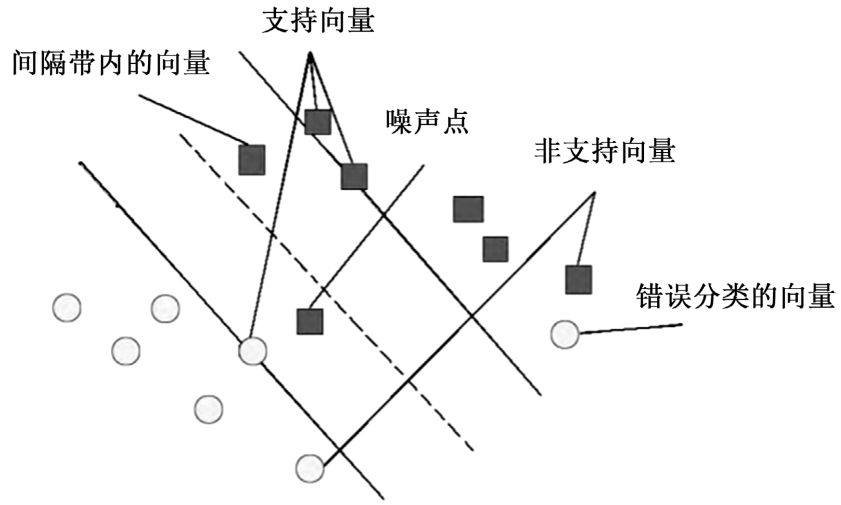

Figure 1. SVM principle
图 1. SVM 原理

SVM 拥有深厚的统计学知识，而且已经取得了许多令人瞩目的成果。它能够高效地处理高维度的数据，从而克服维度的挑战。SVM 的核心思想之一便是利用模型的某个子集作为决策的边界，即所谓的“支持向量”，从而获得了更高的效率和准确性。

数据分类是机器学习中一种重要的任务，它旨在通过二元分析，将给定的数据点划分为两个不同的类别，以便更好地识别出新的数据点，并将它们归入相应的类别中。假设有若干个给定的数据点，它们都属于二元分类中的两个类别。我们的分类目的是针对任何一个新的数据点，确定它应该属于哪个类别。

2.2. SVM 核函数

SVM 算法的性能取决于核函数的选择，特别是对于那些复杂的线性数据，这种选择非常重要，因为它们可以提高算法的效率和准确性。

由于本文选择数据集中样本量一般，特征量较少，因此选择使用高斯核函数(RBF Kernel)，RBF 核函数[9]可以将原始特征映射到广义的高维空间，从而使其变得更加可分离。同时，此核函数可以轻松地处理不可分离和非线性问题，并具有良好的鲁棒性。由于 SVM 使用了二次规划等数学优化方法进行训练，因此向量机还可以推广到较大规模的数据集中应用，而且 RBF 核函数的计算速度也比其他核函数快，下面为高斯核函数表达式：

$$k\left(x,x_i\right)=\exp\left(-\frac{\left\|x-x_i\right\|^2}{\delta^2}\right)$$

2.3. 分类约束规则

高斯核函数是支持向量机(SVM)中广泛使用的核函数之一，使用高斯核函数来进行分类是一种有效的方法，它能够将复杂的非线性问题转换为高维特征空间，从而更好地描述和预测数据的复杂性。为了在使用高斯核函数进行分类时获得最佳结果，有一些限制条件需要满足。

1) 非负性：对于任意的样本 x_i 和 x_j，高斯核函数的取值始终为非负数，即 $K\left(x_i,x_j\right)\geq 0$；

2) 对称性：对于任意的样本 x_j，高斯核函数的取值在 i,j 顺序和 j,i 顺序下应该相等，即 $K\left(x_i,x_j\right)=K\left(x_j,x_i\right)$；

3) 可正定性：对于任意的样本 $x_1,x_2,\cdots,x_n$ 和任意的系数 $\alpha_1,\alpha_2,\cdots,\alpha_n$，如果对于所有 $i,j\in[1,n]$，有 $\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K\left(x_i,x_j\right)\geq 0$ 则称高斯核函数是可正定的。这个条件证明了高斯核函数可以定义新的内积，在高维空间上使用支持向量机算法进行分类。

通过以上约束条件，能够确保高斯核函数作为一个有效合法的核函数，在支持向量机中能正确有效地处理非线性分类问题。

回归支持向量机算法中，采用 ε-不敏感损失函数的算法，要寻找回归函数 $f\left(x,a\right)=w\cdot x+b$ 中的参数 $\overline{w}$、$\overline{b}$，问题转化为最小化下式：

$$\min_{w,\xi}\varnothing\left(w,\xi\right)=\frac{1}{2}\|w\|^2+c\sum_i\left(\xi_i^*+\xi_i\right)$$

约束为：

$$\begin{cases}y_i-\left(w\cdot x_i\right)-b\leq\varepsilon+\xi_i^*,i=1,2,\cdots,k\\\left(w\cdot x_i\right)+b-y_i\leq\varepsilon+\xi_i,i=1,2,\cdots,k\\\xi_i^*\geq 0,i=1,2,\cdots,k\\\xi_i\geq 0,i=1,2,\cdots,k\end{cases}$$

其中 ξ_i^*、ξ_i 为松弛变量。

3. 基于 SVM 的新冠疑似人员预测

3.1. 数据来源

由于新冠各项检测指标数据的隐私性而无法直接搜索到实际数据集，我们使用了随机函数生成的数据作为本论文的数据来源。具体来说，用 rand 函数[10]生成了一系列随机数据[11]。每个数据集包含了一定数量的样本和相应的属性，并通过调整生成参数来控制数据的噪声程度和分布形状。

通过随机抽样的方式，收集体温、年龄、咳嗽、肌肉酸痛程度、血液常规、PCT、CRP、核酸检测、风险地区等多个指标的数据，构建出一个完整的原始特征库，并依据严格的数据采集标准进行筛选。

3.2. 数据预处理

3.2.1. 数据归一化和离散化

当一个人的体温超出正常范围时，通过归一化，可以更容易地确定一个人的体温偏离正常值的程度。归一化可以使不同特征的数据统一到一个相似的尺度上，避免数据之间的数量级差异对分析结果造成影响，方便进行比较、可视化和处理。

例如，在研究不同年龄段人群的体温情况时，通过对人体体温进行归一化可以消除年龄因素的影响，从而更好地比较不同年龄段的数据，此处对体温、年龄、血常规数据进行归一化处理。

为了减少数据中噪声和异常值的影响，方便模型构建和算法优化，提高数据的可比性，在处理人员年龄和体温这两个连续指标时对其进行离散化，将数值分成一定数量或宽度的区间。通过离散化，可以将连续数据转化为分类数据，方便对数据进行分析、规约和挖掘，同时在一定程度上避免了极值或异常值对数据统计特征和算法准确率的影响。如图 2 和图 3 所示。

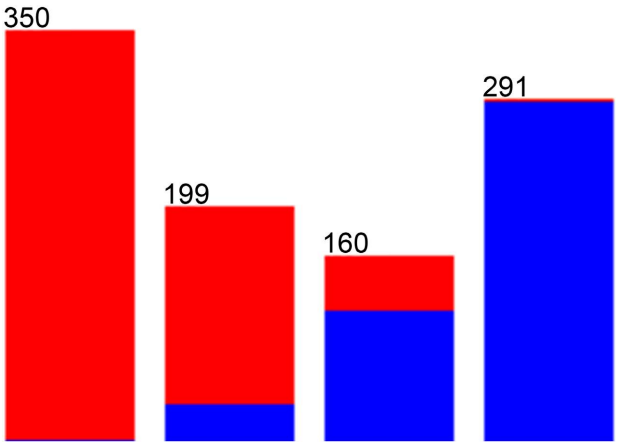

Figure 2. Normalized and discretized body temperature data
图 2. 归一化和离散化后的体温数据图

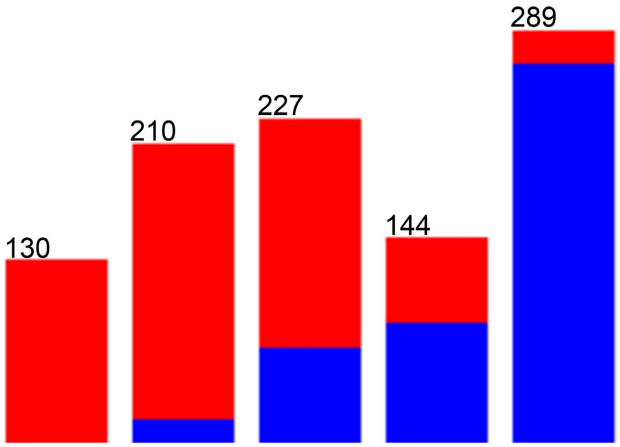

Figure 3. Age data graph after normalization and discretization
图 3. 归一化和离散化后的年龄数据图

3.2.2. 数据权重分配

当处理疑似新冠的指标时，设置每个指标的权重是一件非常重要的事情，这有助于确定准确性和优先级。考虑到目前的疫情情况，可以将指标的权重设置如下：

发烧：30%。由于新冠病毒最常见的症状是发热，所以这个指标应该具有相当高的权重。在对发热的评估中，通常会使用体温计来测量身体温度。一般认为，正常体温为 36.5~37.0 摄氏度。如果患者的体温超过了此范围，那么他们可能已经感染了新冠病毒。

年龄：权重 15%。虽然年龄并不是明显的症状，但年龄与感染新冠病毒产生的威胁有很大关系。较年长者更容易受到感染的影响，患上严重病例的可能性也更大。

咳嗽：7%。咳嗽也是新冠病毒感染的常见症状。咳嗽可以是干咳或带痰，这与不同程度的上呼吸道炎症有关。但并非所有患者都有咳嗽，其相对重要性相对较小。

肌肉酸痛：权重 5%。和咳嗽类似，并非所有患者都肌肉酸痛。

血常规：权重 10%。血常规可以提供关于白细胞计数和血小板数量的信息。在判断感染或血液问题是否与新冠病毒有关时，血常规尤其重要。新冠肺炎患者的血常规特点是白细胞计数正常或下降不同程度，同时，细胞计数及细胞比也会有不同程度下降；血沉及 c 反应蛋白则会升高。

Pct：权重 10%。Pct 是一种与细菌感染相关的蛋白质(正常参考值是小于 0.5 μg/L)，但在 COVID-19 的评估中，它被认为是 COVID-19 感染指标之一。

Crp：权重 8%。Crp 是一种反应性蛋白质，它在机体受到感染或炎症时会增加(CRP 在血液中的正常值范围是在 5~10 mg/L 之间，CRP 如果超过 5 就考虑有疾病的情况引起)。因此，它是 COVID-19 感染评估的另一个关键指标。

核酸检测：5%。核酸检测是诊断新冠病毒感染的重要方法之一。核酸检测是一种特定的实验室技术，可以检测新冠病毒的特异性基因序列，对疑似病例进行明确诊断。然而，由于核酸检测结果通常需要等待 2~3 天，这段时间内症状可能会进一步扩散，严重影响患者及时就诊和治疗的机会。本文的目的在于快速判断各项症状，因为核酸检测的及时性存在不足，而且其结果也可能存在假阴性情况，有时需要进行多次核酸检测才能获得可靠结果。考虑到获取准确结果的确实需要时间，因此我们将其权重设为较小。但是，为了确保及时性和准确性，我们可以借助多种方法进行判断和诊断。

风险地区：权重 10%，疫情严重情况对于判断患者是否有疑似感染的重要性不可忽视，如果来自疫情严重的地区的患者出现症状，则需要更加重视。如表 1 所示。

Table 1. Standard test system result data
表 1. 标准试验系统结果数据

训练对象	权重
体温	30%
年龄	15%
咳嗽	7%
肌肉酸痛	5%
血常规	10%
Pct	10%
Crp	8%
核酸检测	5%
风险地区	10%

3.2.3. 构造乘法函数

可以使用以下公式来构造乘法函数：

$$f(x) = w_1^{x_1} * w_2^{x_2} * w_3^{x_3} * w_4^{x_4} * w_5^{x_5} * w_6^{x_6} * w_7^{x_7} * w_8^{x_8} * w_9^{x_9}$$

其中，w_1 对应发热的权重(0.30)，w_2 对应年龄的权重(0.15)，w_3 对应血常规的权重(0.10)，w_4 对应 pct 的权重(0.10)，w_5 对应 crp 的权重(0.08)，w_6 对应咳嗽的权重(0.07)，w_7 对应肌肉酸痛的权重(0.05)，w_8 对应核酸检测的权重(0.05)，w_9 对应风险地区的权重(0.1)。

而各自的 x 则代表对应指标的特征提取值。例如，如果一个患者有发热、血常规正常、pct 值偏高，核酸检测为阳性，其他指标为零，则其特征向量可以表示为(1, 0, 0, 1, 0, 0, 0, 1, 0)，这样计算出的函数值可以作为该患者确诊为新冠病例的可能性的参考。值越大，则指示该患者越有可能是新冠病例。

3.3. 模型评估和结果分析

采用已经经过数据预处理的 1000 条患者的各项指标数据，使用十折交叉验证[12]的方法来进行训练

和测试，将数据集分为 10 份，每次使用 9 份数据进行训练模型，剩余的 1 份数据用于测试模型，然后重复 10 次以取得平均精度，同时使用混淆矩阵来评估模型的性能，其中每一行代表实际类别，每一列代表预测类别。

采用 weka 机器学习软件，采用 SVM 模型进行结果分析，测试结果如表 2 所示。

Table 2. Accuracy and error rate based on SVM model
表 2. 基于 SVM 模型的准确率和错误率

Summary	quantity	percentage
Correctly Classified Instances	917	91.7%
Incorrectly Classified Instances	83	8.3%

采用准确率和召回率作为两个主要的评价指标，准确率(Precision)是指分类器预测为正类的样本中有多少是真正的正类。其计算公式为：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

其中，TP 是预测为正类且正确的数目，FP 是预测为正类但错误的数目。

召回率(Recall)是指所有实际正例中被分类器正确识别为正例的样本数占比。其计算公式为：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

其中，TP 是预测为正类且正确的数目，FN 是预测为负类但错误的数目。测试结果如表 3 所示。

Table 3. Detailed accuracy by category
表 3. 按类别划分的详细准确性

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	Class
0.922	0.087	0.891	0.922	0.906	0.832	1
0.913	0.078	0.938	0.913	0.925	0.832	0

此外得到混淆矩阵，预测阴性且正确的个数为 515，预测阳性且正确的个数为 402，经过测试，如表 4 所示，可见模型准确性和有效性较高。

Table 4. Confusion matrix
表 4. 混淆矩阵

a	*b*	classified as
402	34	*a* = 1
49	515	*b* = 0

4. 结论

首先，为了减少噪声，我们需要使用随机函数对数据进行合理化生成。通过控制参数，可以有效地降低噪音水平，并确保所生成的数据具有较高的质量。对数据进行离散化和归一化处理，增强数据的可比较性和可靠性；然后采用 SVM 模型进行数据的分类和预测，实验数据表明，基于 SVM 的预测模型，

准确率较高，具有较好的预测能力。然后使用 SVM 模型对数据进行分类和预测。实验数据表明，基于 SVM 的预测模型具有高准确率和良好的预测能力。

此外使用机器学习技术进行疑似新冠人员预测是具有潜力的。虽然疫情对我们的影响不再像之前一样严重，但是基于机器学习对疑似新冠人员的快速筛查和诊断的方法是可行的。

机器学习可以通过自动处理数据、挑选适当特征和算法来提高预测结果的精度和效率，但是，数据收集和处理、特征选择、算法选择和模型评估和改进等诸多方面的挑战也需要克服。因此，在未来的研究中，需要进一步探索机器学习在疑似新冠人员预测中的应用，并持续优化和改进机器学习模型的性能。

致　谢

首先，要感谢我们的导师樊纪山老师给予宝贵的指导和支持，让我们认真学习科研方法和思维方式，取得了此次研究的成果，安心完成这篇论文。其次我要感谢学校提供了相关的学习环境，使我们能够顺利完成这篇论文。最后，也要感谢所有曾经帮助过的人，他们的帮助和支持不仅对我的学习工作有着重要影响，同时也让我在生活中不感孤单。

参考文献

[1] 韩昇洙, 韩子砚. 新冠肺炎疫情如何影响全球经济[J]. 国际金融, 2020(7): 3-6. https://doi.org/10.16474/j.cnki.1673-8489.2020.07.001

[2] 吴金龙, 丁小兵, 刘志钢. 上海市轨道交通系统防疫策略研究——以新型冠状病毒肺炎疫情为背景[J]. 城市交通, 2020, 18(3): 46-50. https://doi.org/10.13813/j.cn11-5141/u.2020.0303

[3] Daniel, F. (2020) 7 Applications of Machine Learning in Pharma and Medicine. TechEmergence.

[4] 詹骐源. 机器学习的发展史及应用前景[J]. 科技传播, 2018, 10(21): 138-139. https://doi.org/10.16607/j.cnki.1674-6708.2018.21.069

[5] Vaishya, V. and Vishwamitra, L.K. (2021) Diabetes Detection Using Neighborhood Component Analysis with SVM and Random Forest Classifier. *Design Engineering*, No. 6, 7534-7551.

[6] 王晓丽, 施天行, 彭德荣, 等. 两种机器学习算法构建老年冠心病患病风险评估模型的效能比较研究[J]. 中华全科医学, 2021, 19(4): 523-527.

[7] 任建强, 崔亚鹏, 倪顺江. 基于机器学习的新冠肺炎疫情趋势预测方法[J/OL]. 清华大学学报(自然科学版): 1-9. https://doi.org/10.16511/j.cnki.qhdxxb.2023.22.006, 2023-05-15.

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

[9] 王建国, 赵鹏飞, 张文兴, 秦波, 刘文婧. 多尺度高斯核支持向量机算法[J]. 机床与液压, 2020, 48(20): 5-8.

[10] 王鹏, 董平轩, 冉玉梅. 《概率论与数理统计》教学中随机数的 Matlab 实现[J]. 电脑知识与技术, 2023, 19(2): 138-140+161. https://doi.org/10.14004/j.cnki.ckt.2023.0106

[11] 张晓军, 曹惠茹. Matlab 中的随机函数[J]. 电脑编程技巧与维护, 2010(14): 115-116+127. https://doi.org/10.16184/j.cnki.comprg.2010.14.002

[12] 杨敬娜. 基于十折法的最小二乘支持向量机参数选取方法[J]. 机械工程师, 2011(12): 28-29.