Advances in Applied Mathematics 应用数学进展, 2021, 10(1), 305-310
Published Online January 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.101035

Hans汉斯

基于三种分类算法的CTG数据应用研究

刘青松*，瞿颖秋，张晓良

重庆理工大学理学院，重庆
Email: *liuqingsong950729@163.com

收稿日期：2020年12月25日；录用日期：2021年1月19日；发布日期：2021年1月28日

摘　要

本文采用邻近算法、决策树、支持向量机三种分类方法对胎心宫缩监数据(CTG)进行分类分析，得出每种方法的分类结果，并就每种方法的准确率、误判率进行判别。通过研究表明，决策树可以对实际数据进行很好的分类。

关键词

邻近算法，决策树，SVM，分类

Research on CTG Data Application Based on Three Classification Algorithms

Qingsong Liu*, Yingqiu Qu, Xiaoliang Zhang

College of Science, Chongqing University of Technology, Chongqing
Email: *liuqingsong950729@163.com

Received: Dec. 25th, 2020; accepted: Jan. 19th, 2021; published: Jan. 28th, 2021

Abstract

In this paper, three classification methods, namely proximity algorithm, decision tree and support vector machine, are used to classify and analyze the fetal heart contractions (CTG) data. The classification results of each method were obtained, and the accuracy and misjudgment rates of each method were identified. The results show that decision tree can classify the actual data well.

*通讯作者。

文章引用: 刘青松, 瞿颖秋, 张晓良. 基于三种分类算法的 CTG 数据应用研究[J]. 应用数学进展, 2021, 10(1): 305-310. DOI: 10.12677/aam.2021.101035

Keywords

Proximity Algorithm, Decision Tree, The SVM, Classification

1. 绪论

随着互联网计算机技术的飞速发展，人们逐步陷入了一个“数据丰富，处理苦难”的尴尬地步。如何从海量数据中挖掘出有用的信息，并进行分类分析得出想要的信息已经越来越成为专家学者关注的课题。因此，基于机器学习的数据挖掘方法已经成为当前研究的热门话题。常用的几种方法有 KNN 算法、决策树、支持向量机，它们大量应用到生活、教育、医疗等各个方面。医疗资源短缺是现阶段医疗领域中突出的问题，疾病被正确诊断的有效率是人们所关心的问题。但是一名合格的医生培养需要大量的时间、金钱的投入，因此，利用机器诊断疾病成为专家学者研究的方向。目前，已经有很多疾病，例如冠心病、癌症、哮喘等，有了系统的研究，但是基于胎心宫缩监(cardiotocography, CTG)的研究相对落后。

Haritopoulos M 等[1]提出了处理 CTG 信号最常见的步骤，并研究了过去十年的胎心率分析的特征提取方法，一些特征分类方法的也被提出，以供进一步的研究。Chamidah N [2]通过使用 K-Means 方法对 UCI 中的 CTG 数据进行选择，将从原始特征中提取出来的 7 个抽象特征利用 SVM 模型进行分类，达到了较高的准确率。聂磊[3]利用机器学习的若干办法研究了胎儿窘迫临床数据，从数据中提取计算机能够识别的病理特征，针对病理特征建立模型并进行准确分类，得到了比较好的分类效果。刘志康[4]将胎儿监护的 CTG 信号数字化，针对不同强度的宫缩信号进行自动分类，在 CTG 胎儿智能监护方面提供了理论支持和技术基础。

本文所感兴趣的是使用邻近算法、决策树、支持向量机三种分类方法对胎心宫缩监数据(CTG)分别进行分类分析。

2. 相关理论

2.1. 邻近分类

邻近分类是一种非参数学习的方法，它是把具有相似属性的事物划分为一类，即“物以类聚，人以群分”，其主要步骤在于计算距离以及 k 的数值确定。这里的距离不仅仅包含传统的距离，还包含两个案例之间的相似性。常用的距离为欧氏距离：

$$\operatorname{dist}(p,q)=\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}$$

其中，p、q 是两个案例，p_n,q_n 是案例 p、q 的第 n 个特征的值。k 值的选取往往在 3~10 之间。一种做法是设置 k^2 等于训练集中的事件数目，另一种做法是各种测试数据测试多个 k 值，选择一个分类效果最好的 k 值。

2.2. 决策树分类

决策树是基于特征的值把数据分解成很多较小的子集，这些子集的元素都是相似的，因此，决策树

的核心思想就是分而治之。决策树的难点在于如何进行分割，熵常被用来度量纯度，以此来确定分割的标准，用信息增益来决定根据那个特征进行分割，信息增益越高，表明分割以后的分区越平衡。

2.3. 支持向量机

20世纪90年代初，Vapnik等人[5]基于统计学习理论提出了一种新的机器学习方法——支持向量机。为了确保学习机器的实际风险达到最小，以结构风险最小化准则为理论基础，合适的函数子集和判别函数可以用该准则选择，从而保证了得到误差很小的分类器。核函数的确定是支持向量机的核心，对于不同的支持向量，它们满足 Mercer 条件的核函数也不同，不同类型的算法就由此产生。

3. 实例分析

本文选择胎心宫缩监数据集作为数据源[6]，它来源于 Frank and Asuncion (2010)。该数据集包含了2126个观测值以及23个变量，其中，前20个变量为定量变量，后3个变量分类变量；前22个变量为自变量，后一个变量为因变量，它是一个分类变量，也就是本文所感兴趣的变量，将它的三个水平数值化：1代表正常，2代表疑似，3代表病态，它们分别有1655，295，176个观测数据。本文将使用精确度、十折交叉验证的平均误差率和 Kappa 统计量的值来评估分类方法的好坏。

3.1. 邻近分类法

在读取数据之后，将 NSP 变量重新编码为 normal，uncertain，disease。初步分析原始数据，发现有77.8%的胎儿是正常的，有13.9%的胎儿疑似不正常，只有8.3%的胎儿被认为是病态的。

将这些特征进行 min-max 标准化后，选取数据的80%作为数据的训练集，20%的数据作为测试集，也即是有1701个训练的样本，425个测试的数据。本文选取的 k 的值为6，并利用测试数据集合进行预测，得到的结果汇入一个交叉表中，见下图1：

w_test_labels	w_test_pred normal	uncertain	disease	Row Total
normal	281	54	4	339
	0.829	0.159	0.012	0.798
	0.904	0.900	0.074	
	0.661	0.127	0.009	
uncertain	3	2	0	5
	0.600	0.400	0.000	0.012
	0.010	0.033	0.000	
	0.007	0.005	0.000	
disease	27	4	50	81
	0.333	0.049	0.617	0.191
	0.087	0.067	0.926	
	0.064	0.009	0.118	
Column Total	311	60	54	425
	0.732	0.141	0.127	

Figure 1. The neighbor classification result
图 1. 邻近分类结果图

在图1中，第一列代表原始测试集数据之中每个胎儿的分类情况，第一行代表预测测试集的每个胎儿的分类情况。有4个胎儿是正常的，但是被判定为病态，有27个是病态的被判定为正常的，这个方向上的错误可能会产生极其严重的后果。主对角线代表预测结果与实际结果一致的情况，有281个被判定为正常，2个被判定为疑似，50个被判定为病态，正确的判断率为82.9%。通过十折交叉验证后，得到其平均误差率为0.3791054。此外，计算其 Kappa 统计量的值为：

$$\frac{0.664+0.09+0.115-0.732\times0.198-0.146\times0.012-0.122\times0.191}{1-0.732\times0.198-0.146\times0.012-0.122\times0.191}=0.842$$

3.2. 决策树分类

同样的，这里选取 80%的数据作为训练集，20%的数据作为测试集，树的大小为 13，如图 2。

```
Class specified by attribute `outcome'

Read 1701 cases (23 attributes) from undefined.data

Decision tree:

CLASS > 7:
:...CLASS <= 9: disease (95/1)
:   CLASS > 9: uncertain (197/1)
CLASS <= 7:
:...CLASS <= 4: normal (920/1)
    CLASS > 4:
    :...CLASS <= 5: uncertain (68/3)
        CLASS > 5:
        :...Mean > 146:
            :...Min <= 69: uncertain (10)
            :   Min > 69: normal (11)
            Mean <= 146:
            :...DP <= 0.000793651: normal (336/2)
                DP > 0.000793651:
                :...MSTV > 3.8: uncertain (3)
                    MSTV <= 3.8:
                    :...Nmax > 12: uncertain (2)
                        Nmax <= 12:
                        :...CLASS <= 6: normal (38/1)
                            CLASS > 6:
                            :...FM > 0.1893204: uncertain (2)
                                FM <= 0.1893204:
                                :...MSTV <= 1.5: uncertain (10/3)
                                    MSTV > 1.5: normal (9)
```

Figure 2. Decision tree
图 2. 决策树

图 2 仅仅显示了决策树的前几个分支，括号中的数字表示符合该决策准则的案例数量以及根据该决策不正确分类的案例的数量。如果 CLASS 大于 7 的条件下，如果小于等于 9，那么就判断为患病，(95/1)表示有 95 个案例符合该决策树的条件，有一个被错误地归类为不患病；如果大于 9，则判定为疑似，且有 197 个案例符合此条件，只有一个被错误判断。

```
Evaluation on training data (1701 cases):

        Decision Tree
      ----------------
      Size      Errors

        13   12( 0.7%)   <<


       (a)   (b)   (c)    <-classified as
      ----  ----  ----
      1310     6          (a): class normal
         4   285     1    (b): class uncertain
               1    94    (c): class disease
```

Figure 3. Confusion matrix of decision tree
图 3. 决策树的混淆矩阵

从图 3 中得出，模型对 1701 个训练子集分类，有 12 个被错误分类，错误率为 0.7%。共有 6 个正常地被错误判定为疑似，4 个疑似被错误判定为正常，1 个疑似被错误判定为患病，1 个患病地被错误判定为疑似。利用训练集的数据模型来预测测试集的分类结果，如图 4 所示。

在 425 个测试结果中，我们的模型正确的预测了 328 个为正常，4 个为疑似，81 个为病态，模型的准确率为 97.18%。该模型错误地把 11 个正常为判定为疑似、1 个疑似的判定为正常，但是这并没有产生十分严重地错误。总的来说，该模型比较好的对数据进行了分类。通过十折交叉验证后，得到其平均误差率为 0.0146011。此外，计算其 Kappa 统计量的值为

$$\frac{0.772+0.009+0.191-0.798\times0.774-0.012\times0.035-0.191\times0.191}{1-0.798\times0.774-0.012\times0.035-0.191\times0.191}=0.919$$

w_test_j$NSP	w_pred_j normal	uncertain	disease	Row Total
normal	328	11	0	339
	0.772	0.026	0.000	
uncertain	1	4	0	5
	0.002	0.009	0.000	
disease	0	0	81	81
	0.000	0.000	0.191	
Column Total	329	15	81	425

Figure 4. Decision tree classification
图 4. 决策树分类情况

3.3. SVM 分类法

这里的训练集和测试集与决策树的训练集和测试集一样，使用的核函数为高斯径向核函数。将得到的测试数据的预测值与其真实值进行对比，得到如下图 5 结果：

w_test_svm$NSP	w_pred_svm normal	uncertain	disease	Row Total
normal	338	1	0	339
	0.795	0.002	0.000	
uncertain	1	4	0	5
	0.002	0.009	0.000	
disease	33	0	48	81
	0.078	0.000	0.113	
Column Total	372	5	48	425

Figure 5. SVM classification results
图 5. SVM 分类结果

从图 5 看出，425 个预测结果中，有 338 个被正确判定为正常，4 个被正确判定为疑似，有 48 个被正确判定为病态，正确判定率为 91.29%。有 33 个病态的被判定为正常的，有 1 个疑似的被判定为正常，1 个正常的被判定为疑似。通过十折交叉验证后，得到其平均误差率为 0.01552904。计算其 Kappa 统计量的值为

$$\frac{0.795+0.009+0.108-0.798\times0.88-0.012\times0.012-0.191\times0.108}{1-0.798\times0.88-0.012\times0.012-0.191\times0.108}=0.680$$

4. 结论

通过以上三种方法对 CGT 数据的研究表明，我们可以得到下表 1：

Table 1. The evaluation of the three classification methods
表 1. 三个分类方法的好坏评判标准

聚类方法	准确率	平均误差率	Kappa 统计量
邻近算法	82.9%	0.3791	0.842

Continued

决策树	97.18%	0.0141	0.919
支持向量机	91.29%	0.0155	0.680

从表1可以看出，准确率最高的是决策树分类的方法，为97.18%，其次是支持向量机，准确率为91.29%，最低的是邻近分类法，仅为 82.9%。经过十折交叉验证后，得到邻近算法的平均误差率为 0.3791，决策树的为 0.0141，支持向量机的为 0.0155。它们的 Kappa 统计量的值分别为：0.842，0.919，0.680。因此总的来说，决策树的准确率最高，平均误差率最低，kappa 统计量的值表现出很好的一致性，在实际中我们应推荐使用决策树的分类方法。

参考文献

[1] Haritopoulos, M. and Nandi, A.I.A.K. (2016) Survey on Cardiotocography Feature Extraction Algorithms for Foetal Welfare Assessment. Springer, Berlin Heidelberg, 1193-1198. https://doi.org/10.1007/978-3-319-32703-7_231

[2] Chamidah, N. and Wasito, I. (2015) Fetal State Classification from Cardiotocography Based on Feature Extraction Using Hybrid K-Means and Support Vector Machine. *IEEE*, Piscataway, 37-41. https://doi.org/10.1109/ICACSIS.2015.7415166

[3] 聂磊. 若干机器学习算法的研究及在胎心监护中的应用[D]: [硕士学位论文]. 长春: 长春工业大学, 2020.

[4] 刘志康. 面向胎儿监护的 CTG 信号数字化和智能分析算法研究[D]: [硕士学位论文]. 杭州: 杭州电子科技大学, 2020.

[5] 林晓佳. 基于改进 Adaboost M1 算法医学图像分类系统的研究[J]. 聊城大学学报(自然科学版), 2015, 28(4): 29-32.

[6] 吴喜之. 复杂数据统计方法——基于 R 的应用(第三版) [M]. 北京: 中国人民大学出版社, 2015.