文章编号：2095-2163(2020)12-0018-04 中图分类号：TP301.6 文献标志码：A

基于机器学习的医疗数据分析

龙跃飞，李泽滔
（贵州大学 电气工程学院，贵阳 550025）

摘 要：数据分析可以解决数据量大，数据结构复杂等问题，在医疗方面可对大量的医疗数据进行精准分析，本文主要研究监督学习算法中的决策树算法、随机森林算法、K-最近邻算法在乳腺癌数据分析中的应用。通过建立统计分析模型、验证比较得出 K-最近邻算法为这 3 个算法中最合适于乳腺癌数据分析的算法，准确率高达 99.86%。

关键词：数据分析；监督学习算法；K-最近邻；乳腺癌

Medical data analysis based on machine learning

LONG Yuefei, LI Zetao
(College of Electrical Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】 Data analysis can solve the problems of large amount of data and complex data structure, and can accurately analyze a large amount of medical data. This paper mainly studies the application of decision tree algorithm, random forest algorithm and K-nearest neighbor algorithm in breast cancer data analysis. Through the establishment of statistical analysis model, verification and comparison, it is concluded that K- nearest neighbor algorithm is the most suitable algorithm for breast cancer data analysis among the three algorithms, and the accuracy rate is as high as 99.86%.

【Key words】 Data analysis; Supervised learning algorithm; K-Nearest Neighbor; Breast cancer

0 引 言

2014 年《中国乳腺癌现状报告》发表的数据显示，乳腺癌在中国各个省份的肿瘤排名中均处于前 4。预计截至 2023 年中国的乳腺癌病发女性人数达到 23 万人，比 2008 年上升 31%[1]。近年来对于乳腺癌的治疗从个体化治疗慢慢转变成精准治疗，治疗倾向于个体化、精准治疗化以及大数据分析化等[2]。精准治疗的理念已经被逐渐认可，医疗技术也是日益先进，对病人的治疗也逐渐具有针对性，根据实际病情制定个体化方案，在最快时间内到达最好的效果。同时，数据处理技术的提升可以解决一些数据量大、数据结构复杂的疾病数据，帮助医生得到有用的数据从而快速诊断[3]。

此次研究目的是基于部分监督学习算法（决策树、K-最近邻、随机森林）在数据分析中的应用，对这几种算法进行交叉验证比较，找出在对乳腺癌数据分析中误差率最小的算法。

1 监督学习算法

1.1 决策树算法原理

决策树算法原理是从根节点开始，根据不同的变量的特征属性以及输出值选择进行分类的，直至得到最终的叶子节点[4]（决策结果）。

决策树是根据不同的划分属性度量对数据集进行划分，其中划分数据集的常用方法有 ID3 以及 C4.5 两种算法。

在 ID3 算法中每次迭代都会计算每个属性的信息增益，然后按照信息增益最高的属性划分数据，不断重复该过程直到结束[5]；C4.5 算法的优点是产生的分类规则简单易懂，准确率高。但在分类计算过程中需要多次扫描数据集的顺序，重新排序，效率较低，无法对大的数据集进行划分计算。

1.2 K-最近邻算法（KNN）

K-最近邻算法（KNN）作为一种常用分类算法，其算法思想是某个样本在属性空间中的 K 个最近邻的样本的大部分属性属于同一类别，该样本也视为属于这一类别（K 为较小的正整数）。其原理如图 1 所示，有两类已分类的数据（红三角、蓝方块），而绿圆则为待测定样本，若 $K = 3$，与绿圆最近邻的 3 个样本中的红三角有两个，蓝方块有一个，因此绿圆为红三角属性；当 $K = 5$，蓝方块有 3 个，红三角有两个，因此绿圆为蓝方块属性。K-最近邻算法的流程框架如图 2 所示。

作者简介：龙跃飞（1995-），男，硕士研究生，主要研究方向：计算机控制技术；李泽滔（1960-），男，博士，教授，博士生导师，主要研究方向：智能电网、故障诊断、计算机控制技术等。
通讯作者：李泽滔 Email：2235667181@qq.com

收稿日期：2020-09-24

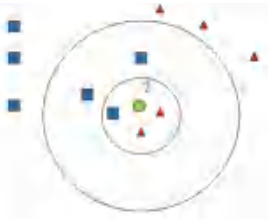

图 1　K 邻近算法原理图

Fig. 1　Schematic diagram of K proximity algorithm

图 2　K-最邻近算法流程框架

Fig. 2　K- nearest neighbor algorithm flow chart

1.3　随机森林分类法原理

随机森林分类是由许多决策树组合在一起的一个分类器，但是随机森林之中的每棵树之间是相互独立的[6]，所以用随机森林分类器对数据进行分析判断时，其中的每一棵决策树也在对该数据进行分析判断，然后分类，决策树所得到的结果中多数的类别水平便是随机森林的分析结果[7]。

随机森林分类算法的步骤：

(1)计算待测数据与样本数据的距离；

待测样本：$X = (x_1, x_2, \ldots, x_n)$；

训练样本；$Y = (y_1, y_2, \ldots, y_n)$；

则样本距离公式(1)为：

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \qquad (1)$$

(2)根据距离从小到大排序；

(3)选取前 K 个训练样本；

(4)算出前 K 个样本的类别的出现频率；

(5)出现频率最高的样本类别就是待测样本的类别。

2　几种算法在数据分析中的应用

2.1　决策树算法在数据分析中的应用

2.1.1　决策树建模

class 作为二输出的因变量(结果为良性 Benign 和恶性肿瘤 Malignant)，其它的 9 个变量作为自变量，利用 R 语言中的决策树程辑包 rpart 函数对乳腺癌数据集里的 699 组数据进行决策树分类建模，得到的决策树模型如图 3 所示。

从图 3 可以直观的看出决策树算法的分类过程，先后通过对每个自变量进行分析比较进行决策，经过层层决策最终得到最终结果[8]。

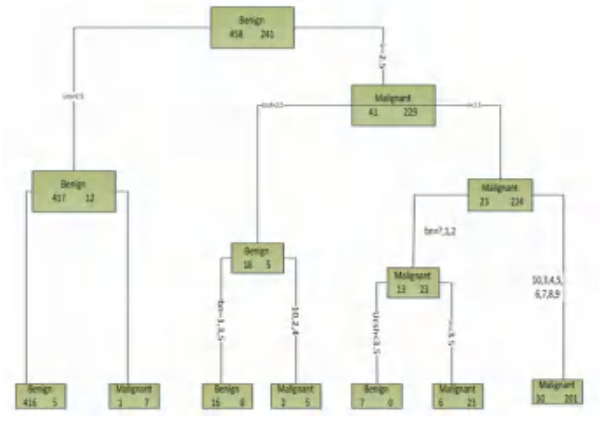

图 3　决策树分类图

Fig. 3　Classification diagram of decision tree

其中 n = 699 表示的是分析的数据集中有 699 组数据；node 为每个节点所对应的序号；split 则是上一节点决策分裂到目前节点的分类标准，如图 4 中是节点 2)所示 ucsi<2.5 作为其上一节点进行分类的标准，小于 2.5 分类为 Benign，大于 2.5 分类为 malignant；n 表示该节点所包含的数据量，节点 2)中有 429 组数据；loss 表示的是当人们在该节点按照少数服从多数对数据进行分类则分类结果中分类出错的数据量，节点 2)有 12 组数据在分类时是不对的；这里的因变量是二分类变量，而 yval 表示经过上一节点进行分类后该节点中的多数因变量输出，节点 2)表示多数因变量输出值为 Benign；(yprob)表示的是在该节点中二分类的因变量值所占比例，其中 Benign 占比约 97.2%，malignant 占比约 2.8%；* denotes terminal node 表示该节点已经终结。其它每个节点的分类细节步骤也都是按照上述进行决策分类。

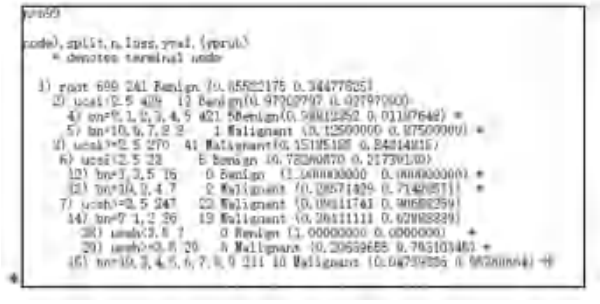

图 4　决策树分类细节图

Fig. 4　Classification Details of Decision Tree

2.1.2　决策树误判率

通过 R 语言建模运行得出了决策树算法的混淆矩阵以及误判率结果，如图 5 所示。从图 5 中的混淆矩阵可以看出决策树模型在对 699 组乳腺癌训

练数据集进行预测,其中有 24 个输出结果出现误判,误判率为 3.43%,误判率是在把原始数据变成训练数据集后再把训练数据集当作测试对象进行测试,判断正确率为 96.57%。

```
> m.p=predict(m,mydata,type='class')
> table (m.p,mydata$class)

m.p          Benign Malignant
  Benign        439         5
  Malignant      19       236
> sum(m.p!=mydata$class) /nrow(mydata)
[1] 0.03433476
```

图 5 决策树混淆矩阵和误判率图

Fig. 5 Confusion matrix and misjudgment rate diagram of decision tree

2.1.3 决策树预测

建立模型是为了在未知的情况下对检测对象进行判断,随意编了一组新的数据进行预测,得到的 R 运行结果如图 6 所示。

```
> new.data
  ct ucsi ucsh ma secs bn bc nn mi class
1  3    1    1  1    1 10  5  2  7    NA
>  predict(a,new.data,type='class')
         1
Malignant
Levels:Benign  Malignant
```

图 6 决策树新数据预测图

Fig. 6 New data prediction diagram of decision tree

从结果中可以看出,输入新数据决策树模型很快就可以判断出结果,图 6 中新数据中类别 class 位置,判断输出结果为 Malignant,决策树模型只要检测到了检测对象的重要检测指标就可以快速的对检测对象做出判断。

2.2 K-最近邻算法在数据分析中的应用

2.2.1 k-最近邻建模

利用 K-近邻学习算法进行数据分析,所选取的 K 值一般为较小的奇数,将乳腺癌数据的二分类水平 class 定义为因变量,其它 9 个检测指标作为自变量进行建模,同时把对象数据三分之二划分为训练数据,剩下的划分为测试数据,通过 R 的运行结果如图 7 所示。

```
> summary (mydata.kknn)

Call :
kknn(formula = class., train=mydata.learn, test=mydata.valid)

Response:"nominal"
        fit       prob.Benign prob.Malignant
1    Benign 1.0000000000    0.000000000
2    Benign 0.9979469915    0.002053009
3    Benign 1.0000000000    0.000000000
4    Benign 1.0000000000    0.000000000
5    Benign 1.0000000000    0.000000000
6 Malignant 0.0000000000    1.000000000
7    Benign 1.0000000000    0.000000000
8 Malignant 0.3164907107    0.683509289
9    Benign 1.0000000000    0.000000000
10 Malignant 0.0000000000    1.000000000
```

图 7 K-近邻分类结果图

Fig. 7 K- nearest neighbor classification results

2.2.2 k-最近邻误判率

求 K-近邻算法的混淆矩阵和误判率,运行结果如图 8 所示。从运行结果中可以发现利用 K-近邻对所选的 699 组乳腺癌数据进行分类时只有一个错误结果,正确率高达 99.86%,可见 K-近邻在此次的乳腺癌数据分析中的分析正确率之高。

```
> table(predict(a,mydata),mydata$class)

           Benign Malignant
  Benign      457         0
  Malignant     1       241
>
```

图 8 K-近邻混淆矩阵和误判率图

Fig. 8 K- nearest neighbor confusion matrix and misjudgment rate diagram

2.2.3 k-最近邻预测

K-近邻算法的预测与决策树算法的预测类似,在原有的数据基础之上添加一组未知因变量水平的数据,通过建立的 K-近邻模型对其进行预测[9],得到的运行结果如图 9 所示,预测结果与决策树算法的预测结果相同。

```
> predict(a,new.data)
[1] Malignant
Levels: Benign Malignant
>
```

图 9 K-近邻新数据预测图

Fig. 9 K- nearest neighbor new data prediction chart

2.3 随机森林分类算法在数据分析中的应用

2.3.1 计算选择随机森林 mtry 值

由于随机森林的抽样构建模型是有放回的抽样,所以没被抽到的数据便成了现成的测试数据,模型内的平均误差值输出如图 10 所示。则误差值对应最小的数值作为 mtry 的值,图 10 结果的 mtry 为 2。

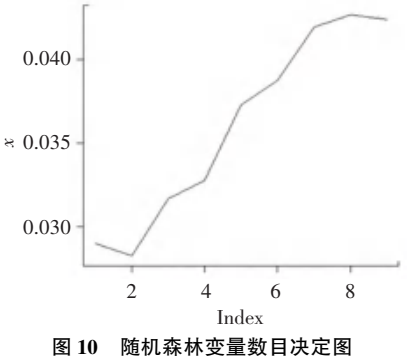

图 10 随机森林变量数目决定图

Fig. 10 Decision diagram of random forest variable number

2.3.2 通过计算选择随机森林 ntree 值

随机森林在构建树的过程中都是由许多树构成的,而决策树又是如同 2.3.1 中一样是由放回抽样构建成的,ntree 的选择也是和 2.3.1 类似在模型中选取误差最小的,结果如图 11 所示。

(下转第 23 页)