题目：关于乳腺癌患者存活率的数据集的分析

作者：李菀彤 （1610120061）

数据集摘要：

|  |  |
| --- | --- |
| 名称 | 英文：Haberman's Survival Data Set  中文：Haberman的生存数据集 |
| 特征简介 | 1.手术时患者的年龄（数字）  2.病人手术年（年-1900年，数字）  3.腋窝淋巴结阳性检出数（数值）  4.存活率（类属性）  --1=患者存活5年或更长时间  --2=病人在5年内死亡 |
| 记录数 | 306条 |
| 分析目标 | 问题：手术时患者的年龄、病人手术年、腋窝淋巴结阳性检出数能否影响乳腺癌患者存活率？  目标：探究手术时患者的年龄、病人手术年、腋窝淋巴结阳性检出数对乳腺癌患者存活率的影响与作用。  方法：数据可视化分析、KNN分类算法和k-means算法。 |

1. 简介

这次大作业我准备针对乳腺癌患者手术后存活率进行分析，此数据集来源于UCI网站（http://archive.ics.uci.edu/ml/datasets/Haberman%27s+Survival），共有306行四列。它是1999年3月4日由Tjen-Sien Lim提供的，共有306条记录、四个属性值，此数据集内容为1958年至1970年间在芝加哥大学比林斯医院进行的关于乳腺癌手术患者存活率的研究的病例。它是一个从属于生活分类的多元数据集，属性值为手术时患者的年龄、病人手术年、腋窝淋巴结阳性检出数和存活率，属性特征为数字。

我准备从三方面做分析。针对第一种分析，我用可视化方法，通过对数据的箱线图、直方图、散点矩阵图和3D散点图四种不同的图形的观察，可以看出存活率与手术时患者的年龄、病人手术年、腋窝淋巴结阳性检出数三个属性共同作用时有关。针对第二种分析，我用KNN分类算法，由于数据集的相关性越强，所有数据属于同一分类，通过得到的数据集相关性数值较大，说明整个数据集属于同一个分类。针对第三种分析，我用k-means算法，说明存活率与单个属性的相关性较低。

1. 数据读取

我使用Pandas库中的read\_csv方法读取乳腺癌患者手术后存活率的数据集，具体代码如下图所示。

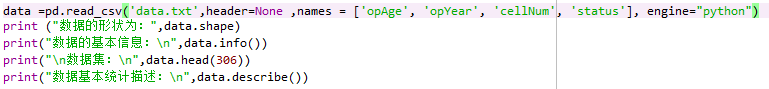


图 1 数据读取、基本描述方法

图1中有五个命令，用到names数组型参数，里面包含'opAge', 'opYear', 'cellNum', 'status'四个属性值，分别代表数据集中手术时患者的年龄、病人手术年、腋窝淋巴结阳性检出数和存活率。再利用data.shape方法读取数据的形状、data.info()方法读取数据的基本信息、data.head()方法读取数据集的具体数据。最后再利用data.describe()方法进行基本数据描述。

通过上述方法，实现对csv文件的读取和基本统计描述，读取结果如下图所示。

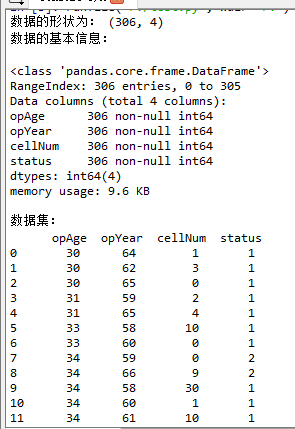


图 2 数据读取结果 1

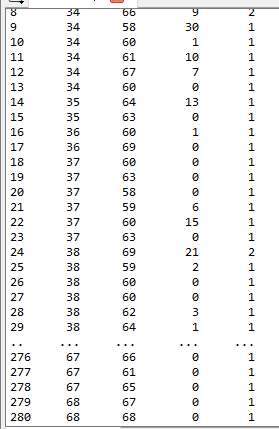


图 3 数据读取结果 2

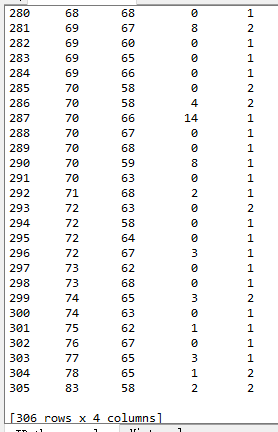


图 4 数据读取结果3

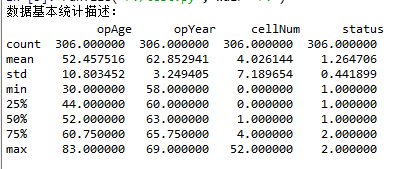


图 5 数据基本统计描述结果

1. 数据预处理

数据读取后，可以确定有306个样本数据，每个数据有四个属性。预处理的目的是把数据转换成便于观察分析、传送或进一步处理的形式，找到其中的空值或者有错误的值进行修改。所以进行的步骤有检查并去除非法字符、检查并处理缺失值。

具体代码如下图所示。

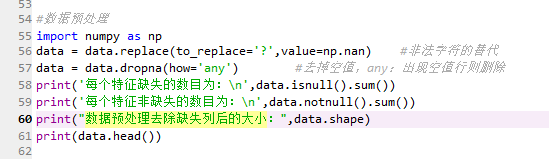


图 6 数据预处理方法

图6中有六个命令，第一个用data.replace()替换非法字符，第二个用data.dropna()去除空值，第三个用data.isnull().sum()读取特征缺失数目,第四个用data.notnull().sum()读取特征非缺失的数目，第五个用data.shape方法读取预处理后列的大小，第六个命令是用data.head()读取前四条数据。

预处理结果如下图。

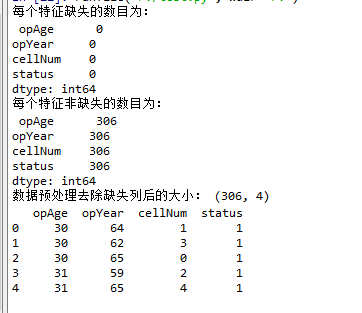


图 7 数据预处理结果

预处理后可知，此数据没有非法字符也没有缺失值，可以继续进行后续分析。

1. 用可视化方法实现数据图表的分析

借助于图形化手段，通过箱线图、直方图、散点矩阵图和3D散点图四种不同的图形，清晰有效地传达与沟通信息，能直观看出数据集内四个不同属性的值的分布。对整个数据集有个初步清晰的印象，便于后续的分析。

具体代码如下图。

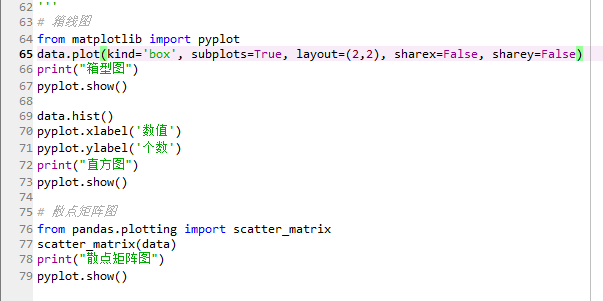


图 8 可视化分析过程1

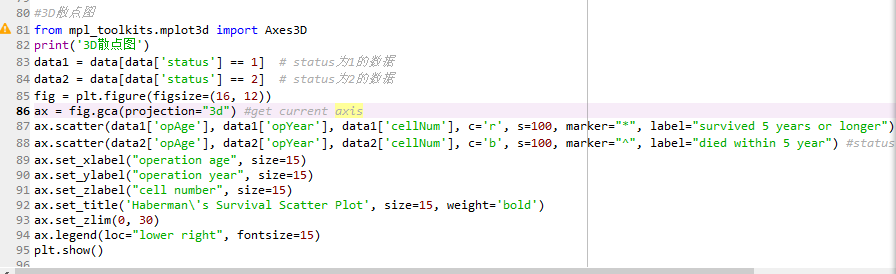


图 9 可视化分析过程 2

上图中分析过程主要分为四个过程用到多个命令，分别用于绘制箱线图、直方图、散点矩阵图和3D散点图。第一个过程的主要命令是data.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)绘制箱线图；第二个过程的主要命令是data.hist()用于绘制直方图；第三个过程的主要命令是scatter\_matrix(data)用于绘制散点矩阵图；第四个过程绘制3D散点图较为复杂，需要先将整个数据data根据‘status’的值1和2分为两个数据data1和data2，再通过ax.scatter()、ax.set\_xlabel()、ax.set\_ylabel()、ax.set\_zlabel()、ax.set\_title()、ax.set\_zlim()方法进行具体绘制。

可视化分析结果如下图所示。

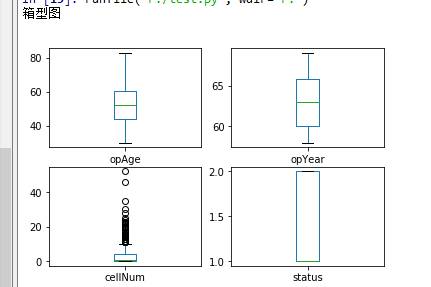


图 10 可视化分析-箱型图

由箱型图可以看出做手术的年龄最小值小于40岁，最大值超过80岁，均值在50岁左右，下四分位数在40岁以上，说明手术时大部分患者的年龄都超过40岁；手术年份距1900年的年份最小值低于60年，最大值高于65年，但是均值处于60-65之间，下四分位数也在60年以上，说明手术年份大部分都在1960年之后；腋窝淋巴结阳性检出数分布大部分集中在20以下，说明患者手术后腋窝淋巴结阳性检出数正常情况应该控制在20以下；存活率1和2皆有，还需要等待后续具体分析。

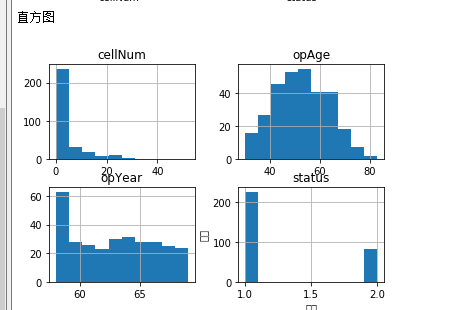


图 11 可视化分析-直方图

由直方图可以看出腋窝淋巴结阳性检出数在0-5范围内极多，有超过200位，说明患者手术后腋窝淋巴结阳性检出数正常情况应该极少；做手术的年龄在50岁时人数最多，大部分集中在40到60岁；手术年份距1900年为60年以下时数量最多，后续年份的手术数量相似，没有大的异常点，；存活率1的个数大于2的个数的两倍，说明存活率较高。

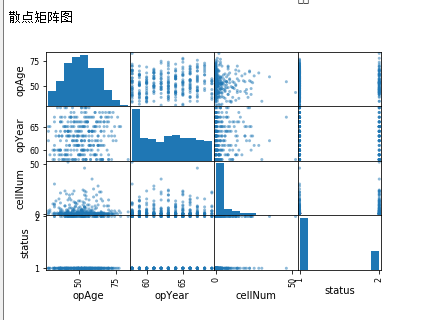


图 12 可视化分析-散点矩阵图

散点矩阵图可观察到的内容与直方图相似，大致可以看出患者做手术时的年龄大概在50左右的居多，做手术的年份除了数据开始年份数量较多其他年份都数量大小都相似，大部分患者术后腋窝淋巴结阳性检出数都极少，术后患者存活率大于死亡率。

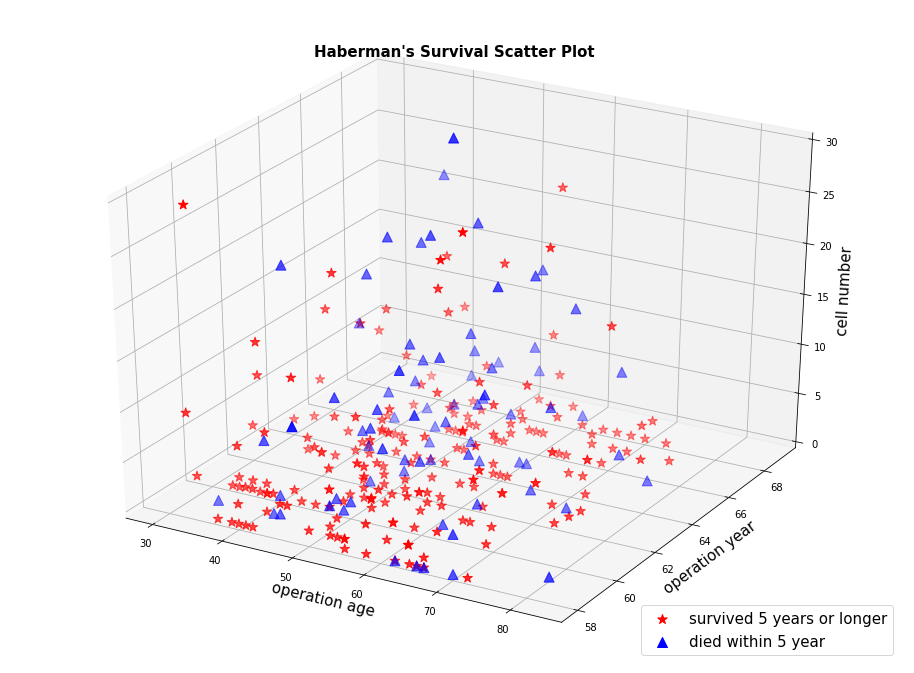


图 13 可视化分析-3D散点图

上图中红色五角星是指手术后存活的乳腺癌患者，蓝色三角形是手术后死亡的乳腺癌患者，通过观察红色五角星和蓝色三角形的位置和坐标关系，大概可以分析乳腺癌患者存活和死亡情况即存活率与手术时患者的年龄、病人手术年、腋窝淋巴结阳性检出数的关系。由图中可以看出，腋窝淋巴结阳性检出数越小，存活人数越多，即存活率越大；腋窝淋巴结阳性检出数越大，存活人数越少，即存活率越小。

综上所述，通过可视化分析的箱线图、直方图、散点矩阵图和3D散点图四个图形，可以看出患者做手术时年龄大部分集中在40到60岁、每年做手术的患者数量都大致相同、腋窝淋巴结阳性检出数较少，存活率与腋窝淋巴结阳性检出数有关。

1. 用分类方法实现数据相关性的分析

通过分类方法中的KNN算法，分析样本数据集里的每条样本的相关性，猜测此样本集的存活情况是否和样本集提供数据相同。预测结果与实际数据相同的概率越大，说明数据集的相关性越强，所有数据属于同一分类。

这个分类分析方法运用了KNN（K最邻）算法，它是著名的模式识别统计学方法，如果一个实例在特征空间中的K个最相似（即特征空间中最近邻）的实例中的大多数属于某一个类别，则该实例也属于这个类别，所选择的邻居都是已经正确分类的实例。

它具体过程有：准备数据、划分数据集、采用欧氏距离作为距离评价准则、10折交叉验证（一次和十次验证）、测试。

具体代码如下图所示。

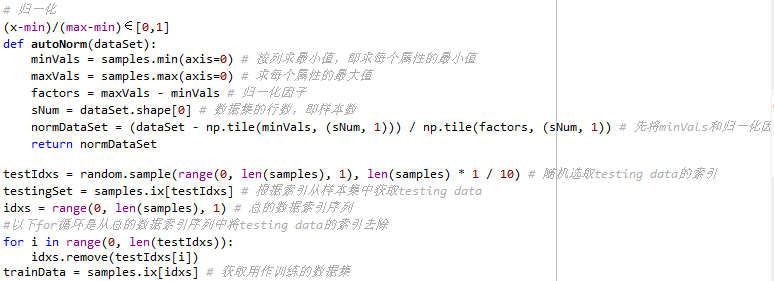


图 14 KNN算法过程1

上图是KNN算法的具体过程，先将数据进行归一化，定义autoNorm()函数，此函数返回一个数据集；再进行划分数据集，运用random.sample()方法随机选取索引，运用samples.ix[]获取测试数据，用for循环idxs.remove()去除索引，最后运用samples.ix[]获取用作训练的数据集。

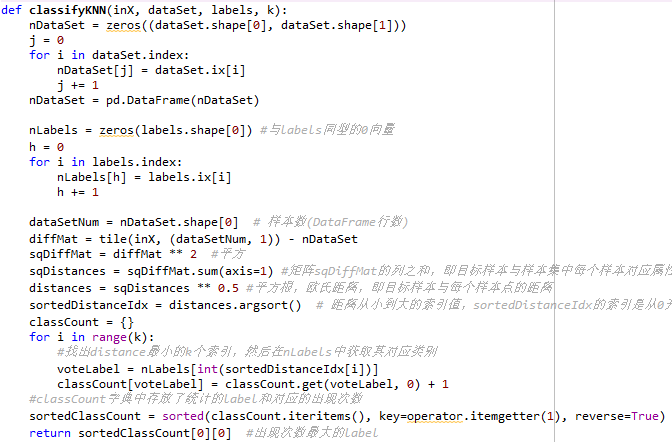


图 15 KNN算法过程2

上图是KNN算法即classifyKNN函数的具体实现，最后返回出现次数最大的标签。

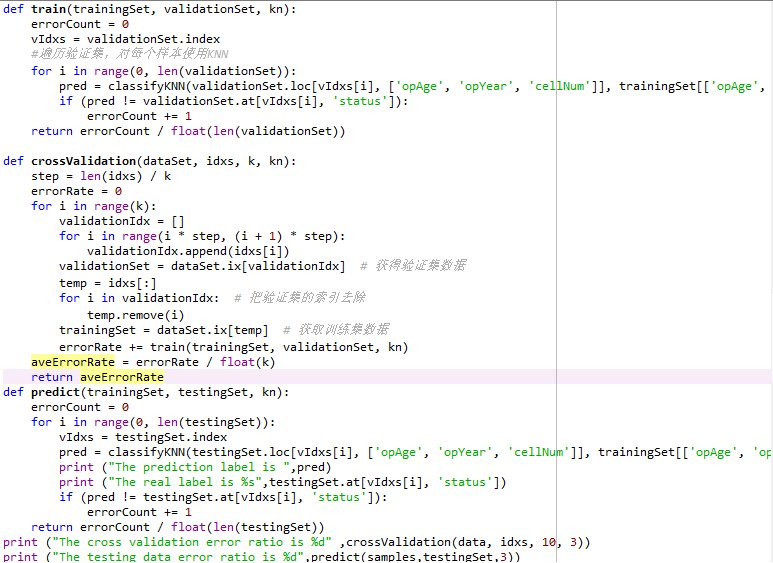


图 16 KNN算法过程3

上图是一次验证train、十次验证crossValidation、测试predict三个函数的具体实现。

KNN算法分析结果如下图所示。

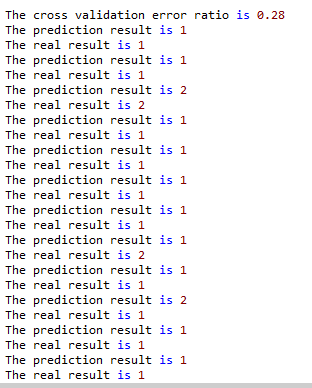


图 17 KNN算法结果1

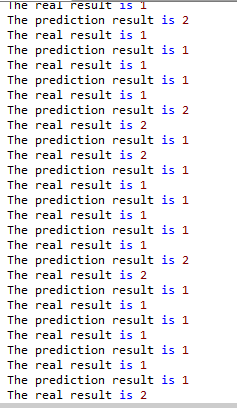


图 18 KNN算法结果2

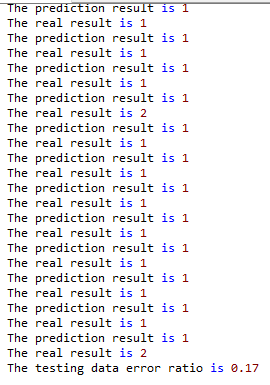


图 19 KNN算法结果3

由于交叉验证错误率是：0.28，测试数据错误率是：0.17，说明预测结果与实际数据相似度很高。说明整个数据集属于同一个分类，乳腺癌已手术患者存活时间(寿命)与患者做手术时的年龄、患者做手术的年份、阳性腋窝淋巴结的数目有很大关系。

1. 用聚类模型方法实现不同属性与存活率的关系分析

这次分析的目的是判断乳腺癌已手术患者存活率分别与患者做手术时的年龄、患者做手术的年份、阳性腋窝淋巴结的数目的相关性，分析只有一种属性影响的情况下，乳腺癌已手术患者存活率。

1. means算法是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。

具体代码如下所示。

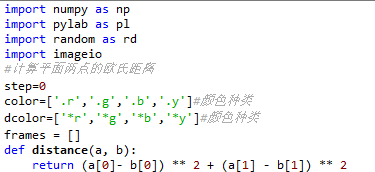


图 20 K-means算法过程1

上图定义了distance函数用来计算并返回平面两点的距离。

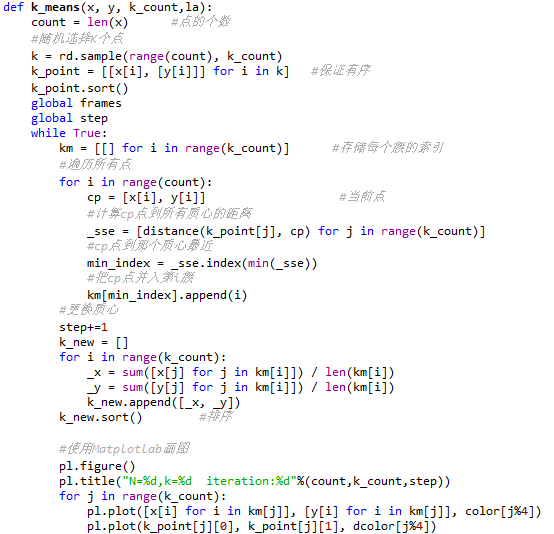


图 21 K-means算法过程2

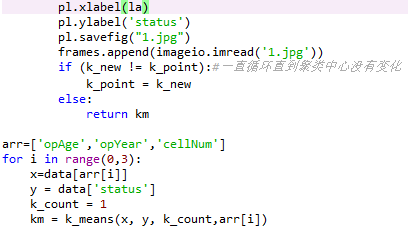


图 22 K-means算法过程 3

上图定义了k-means函数实现k-means算法，主要步骤通过random.sample()随机选取k个点，通过sort()方法进行排序，再通过for循环遍历所有点，再用plot函数绘图，通过for循环改变x轴的参数分别为'opAge','opYear','cellNum'时进行画图。

结果如下图所示。

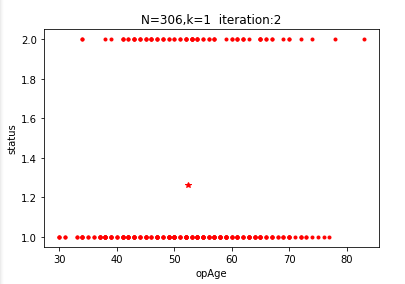


图 23 k-means结果 1

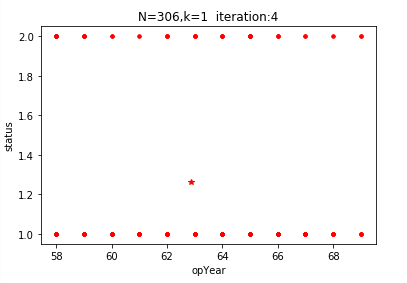


图 24 k-means结果 2

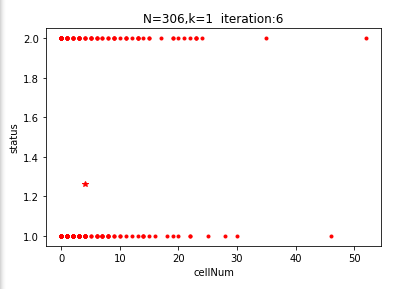


图 25 k-means结果 3

由上图可知，单个属性对存活率的影响不如三个属性共同作用对存活率的影响大。只由一个属性无法准确预测

1. 结论与展望

Haberman的生存数据集里的四个属性具有相关性，每条数据量都可以视作同一个分类，患者的生存状态与手术时患者的年龄、病人手术年、腋窝淋巴结阳性检出数有极大关联，所以患者的生存状态可由其他三个属性值共同推断，但是单个属性对存活率的影响不大。

因此，在医学上，医生在一定程度上可以提前预测患者手术后的存活率，对患者的病情发展和恢复情况有更好的了解。

此处数据分析的难点是KNN算法和K-means算法是，我对这两个算法极为陌生，在几乎全新的算法学习过程和用于分析的过程中，难度最大、收获最多。

此次分析的局限性在于整个数据集的内容距今已经有十多年，医学的迅猛发展让这个数据放在现在也许并没有那么具有参考价值。而关于这个局限性，可以通过查询近几年的数据进行完善，分析是否也具有同样的特性。还可以询问从事医学相关的人员，如今医学对乳腺癌患者的治疗水平和存活率情况。