## 前言

乳腺癌是严重危害妇女健康的恶性肿瘤，其发病率已居女性恶性肿瘤的首位。乳腺癌术后复发转移，是导致最终治疗失败的主要原因。[[1](#_ENREF_1)] 乳腺癌也是现在是女性最常见的恶性肿瘤,居女性恶性肿瘤之首,且发病率正逐年递增。乳腺癌的防治今天越来越受到人们的重视。中医治疗乳腺癌(乳岩)有上千年的历史。[[2](#_ENREF_2)]

中医药物治疗乳腺癌具有非常广泛的适应症和独特的优势。[[3](#_ENREF_3)]从整体出发，调整机体气压、阴阳、肺腑功能的平衡，根据不同的临床病症进行辨证论治。确定“先症而治”的方向：即在后续症状未出现之前，需要截断恶化病情的那些后续症候。发现中医症状间的关联关系和诸多症状之间的规律性，并且依据规则分析病因、预测病情的发展以及为未来临床提供有效借鉴。

随着数据挖掘技术的发展，研究者们越来越习惯于应用数据挖掘的方法来处理实际问题，通过统计方法、机器学习技术来挖掘数据中蕴含的规律。对于乳腺癌的研究，也已经有不少学者尝试应用数据挖掘方法来进行分析。

数据挖掘是应用一系列技术从大型数据库中提取人们感兴趣的信息和知识,这些知识或信息是隐含的、事先未知而潜在有用的,可表示为概念、规则、规律、模式等形式。

徐霏通过对43个症状通过聚类分析得出5组症状聚类，得出肝郁脾虚证、肝肾阴虚证、肝胃不和证、水瘀互结证等体质，根据体质对症治疗[[4](#_ENREF_4)]，取得了很好的效果。

景丽俊使用先聚类后关联数据挖掘的方法进行研究，得出了核心方中药材的分布规律，该研究样本无论大小皆适用[[5](#_ENREF_5)]。

本文的研究思路同样通过数据挖掘的方法研究中医症状之间的关联，分析病因，最终达到能够预测病情的目的。在本次研究中我们以乳腺癌TNM分期为主要对象，并把研究工作细分为4个部分：

1. .基于Apriori方法提取与TNM分期相关的关联规则；
2. .基于FP-TREE提取与TNM相关的频繁项集；
3. .基于决策树提取TNM分期规则并进行分类预测；
4. .基于对TNM分期和转移部位、病程阶段等数据的可视化进行定性分析；

以上各部分工作对乳腺癌TNM分期及相关因素进行了多角度的分析，获取了关联规则知识，并实现了多个针对乳腺癌TNM分期的分类器。具体实现方法将在下文中依次介绍。

## 研究方法与实证分析

### 决策树归纳

#### 为什么选择决策树

挖掘各中医症素与乳腺癌TNM分期之间的关系，是分类问题。决策树是一种用于分类的算法，而且决策树表达直观，易于理解，易于提取规则。而且用决策树训练生成决策树模型后，可通过对中医症素对乳腺癌实现预测，具有较高的正确率。

#### 决策树方法简介

决策树是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。

### ID3决策树

ID3算法是一种贪心算法，用来构造决策树。ID3算法起源于概念学习系统（CLS）。在每个节点选取还尚未被用来划分的具有最高信息增益的属性作为划分标准，然后继续这个过程，直到生成的决策树能完美分类训练样例。

1. 信息增益

ID3算法使用信息增益作为属性选择度量。

对D中的元组分类所需要的期望信息：

是D中任意元组属于类的非零概率，并用估计。

代表用属性A将D划分为v个分区或子集{D1，D2，…,Dv}，项充当第j个分区的权重。是基于按A属性划分对D的元组分类所需要的期望信息。需要的期望信息越小，分区的纯度越高。

信息增益定义为原来的信息需求（仅基于类比例）与新的信息需求（对A划分后）之间的差。即

选择具有最高信息增益的属性A作为结点N的分裂属性。

1. 数据预处理

只考察肝气郁结证型系数，热毒蕴结证型系数，冲任失调证型系数，气血两虚证型系数，脾胃虚弱证型系数，肝肾阴虚证型系数等6个属性，分别以属性A,B,C,D,E,F表示。“TNM分期”用属性X表示。对有缺失值的行删除，然后对A,B,C,D,E,F6个属性的数据分别取每个属性的中位数。对每个属性,高于自身中位数的置1，低于中位数的置0。选取第101-700这700条数据作为训练集，其他作为测试集。

1. ID3算法

输入：训练数据集D，特征集A

输出：决策树T

创建一个结点N

（1）如果D中的所有实例属于同一类 ，则返回N作为叶结点，标记为类。

（2）如果A=∅，返回N作为叶结点，标记为D中的多数类；//多数表决

（3）按计算A中个特征的信息增益，选择信息增益最大的特征；

（4）用标记结点N；

（5）A=A-

（6）对的每一可能值，依 将D分割为若干子集；

（7）如果是空集，加一个树叶到结点N，标记为D中的多数类；否则：以（训练集*,* 特征集A）为输入重复（1）-（7）步。

1. ID3算法优缺点

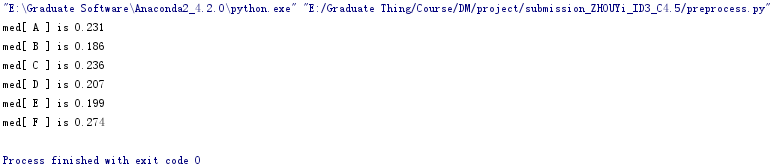
优点：最基本的决策树算法，易于理解。

缺点：（1）不能处理连续型数据；

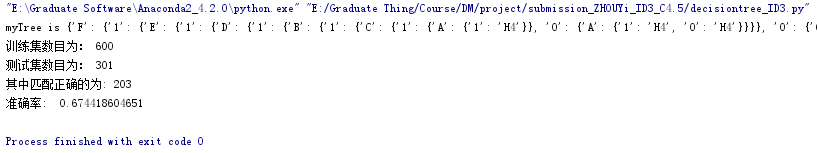
（2）信息增益度量偏向具有许多输出的测试，它倾向于选择具有大量值的属性。（所以必须删除数据集中的id，否则每个id成为一个分区。）

1. 测试结果

数据预处理所得各属性中位数：



ID3算法运行结果截图：



### C4.5决策树

ID3的后继C4.5使用增益率作为属性选择度量。

C4.5用信息增益率来选择属性，克服了ID3用信息增益选择属性时偏向选择取值多的属性的不足；

1. 增益率

通过ID3算法计算得到Gain(A)。

选择具有最大增益率的属性作为分裂属性。

1. 数据预处理

同2.3.2.1 (a)

1. C4.5算法

输入：训练数据集D，特征集A

输出：决策树T

创建一个结点N

（1）如果D中的所有实例属于同一类 ，则返回N作为叶结点，标记为类。

（2）如果A=∅，返回N作为叶结点，标记为D中的多数类；//多数表决

（3）按

计算A中个特征的信息增益，选择信息增益最大的特征；

（4）如果信息增益最大的特征的GainRate(A)<0.025,则返回N作为叶结点，标记为D中的多数类；

（5）用标记结点N；

（6）A=A-；

（7）对的每一可能值，依,将D分割为若干子集；

（8）如果Di是空集，加一个树叶到结点N，标记为D中的多数类；否则：以（训练集D*i,* 特征集A）为输入重复（1）-（8）步。

1. C4.5算法优缺点

优点：1）C4.5用信息增益率来选择属性，克服了ID3用信息增益选择属性时偏向选择取值多的属性的不足；

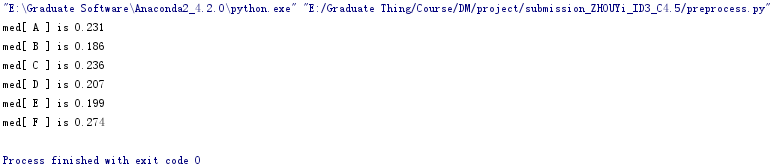
2）既可以处理连续型变量，又可以处理离散型变量。

缺点：1）C4.5只适合于能够驻留于内存的数据集，当训练集大得无法在内存容纳时程序无法运行。

除此之外，本程序采用了预剪枝方法：当信息增益最大的特征gif的GainRate(A)<0.025时则不再扩展子树，虽然起到了一定的剪枝作用，但由于欠拟合导致准确率下降。改进方法：可尝试采用后剪枝方法。

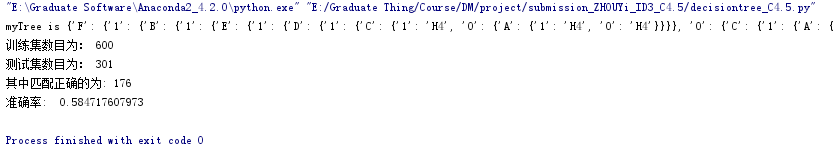
1. 测试结果

数据预处理所得各属性中位数：

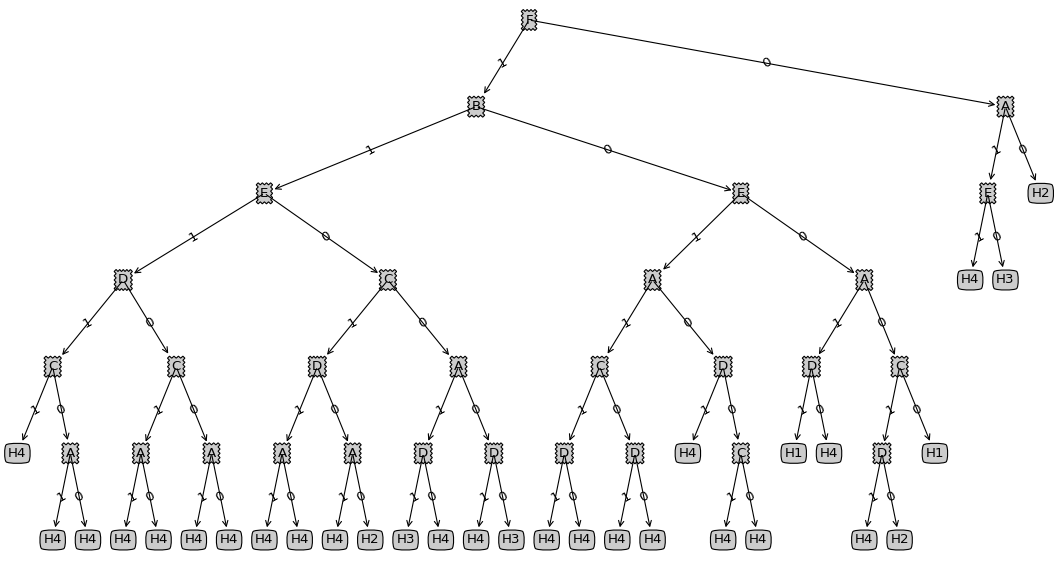


C4.5运行效果截图：

（对树进行了预剪枝）



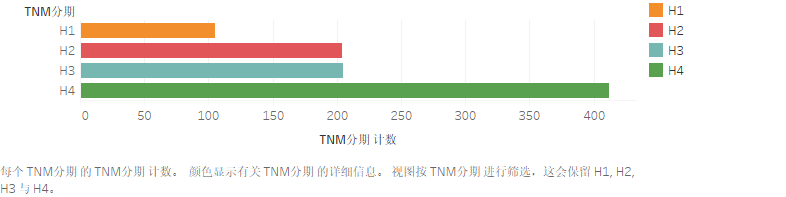
预剪枝后的决策树为：



### 可视化分析

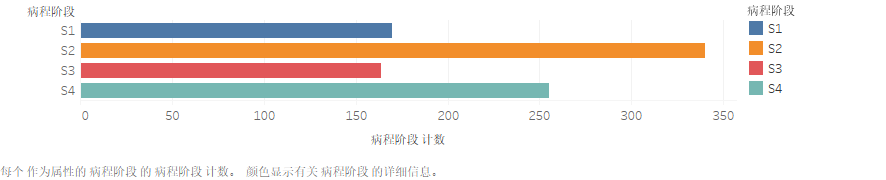
#### 数据特征

（1）TNN分期计数



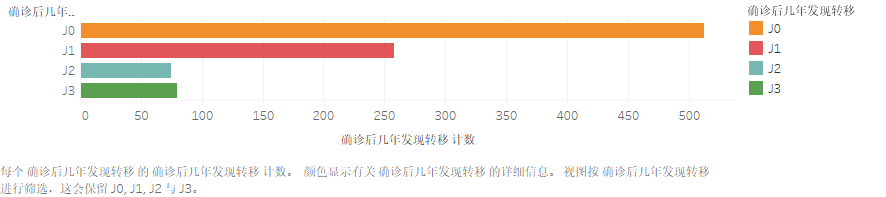
样本数据中处于TNN分期第四分期的人最多，基本是第二、第三分期人数的两倍大小。

（2）病程阶段的计数



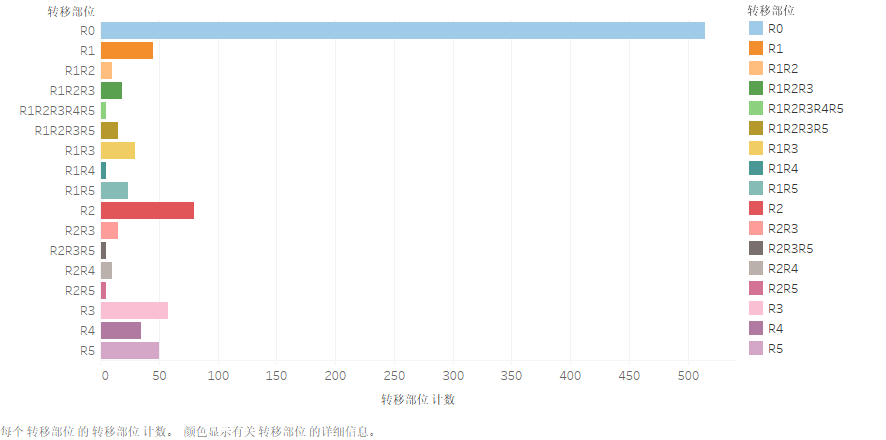
样本数据中处于S2围化阶段的人最多，其次是S4巩固期阶段的人数，S1围手术期和S3围放疗期的人数相当。

（3）确诊几年发现转移的计数



样本数据中无转移的病人数目最多，随着确诊时间的增长，转移数有下降的趋势。

（4）转移部位的计数



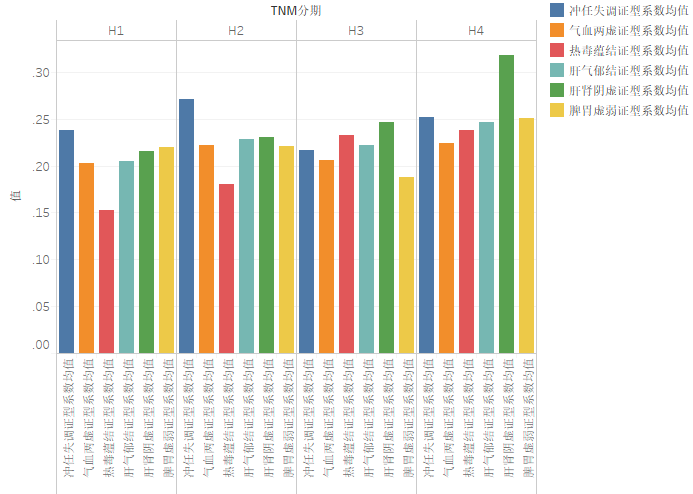
样本数据中，R0无转移的情况最多，其次是转移到R2肺部，R3脑部，R5其它部位，R1骨，R4肝，并且转移部位不单一，有扩散趋势，有转移多部位情况。

（5）病素属性

肝气郁结证型系数、热毒蕴结证型系数、冲任失调证型系数、气血两虚证型系数、脾胃虚弱证型系数、肝肾阴虚证型系数

#### 统计分析

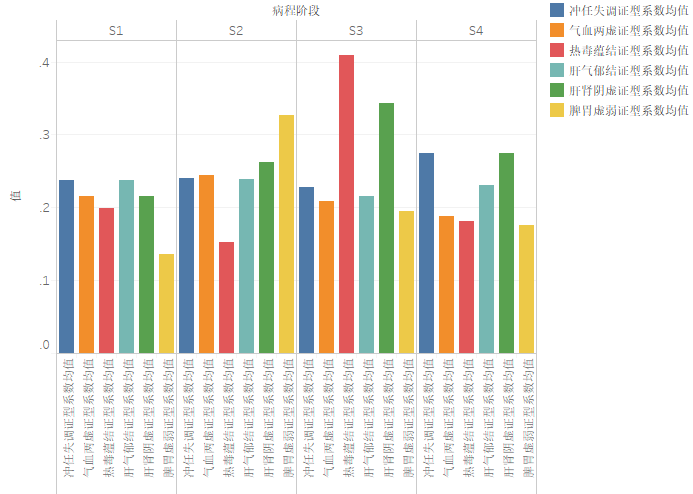
（1）TNN分期和各病素关系



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TNM分期 | 冲任失调证型系数均值 | 气血两虚证型系数均值 | 热毒蕴结证型系数均值 | 肝气郁结证型系数均值 | 肝肾阴虚证型系数均值 | 脾胃虚弱证型系数均值 |
| H1 | 0.2381 | 0.2033 | 0.1523 | 0.2050 | 0.2159 | 0.2199 |
| H2 | 0.2712 | 0.2218 | 0.1805 | 0.2280 | 0.2310 | 0.2212 |
| H3 | 0.2171 | 0.2056 | 0.2330 | 0.2219 | 0.2461 | 0.1875 |
| H4 | 0.2524 | 0.2247 | 0.2378 | 0.2462 | 0.3182 | 0.2510 |

分析样本数据可以发现：H1期，冲任失调证系数较高，热毒蕴结症型系数较低；H2期冲任失调证系数依然教高，其余系数都有上升；H3期热毒蕴结证型系数和肝肾阴虚证型系数上升 ，其余系数下降；H4期所有系数都有上升，肝肾阴虚证型系数在所有分期里是最高的。

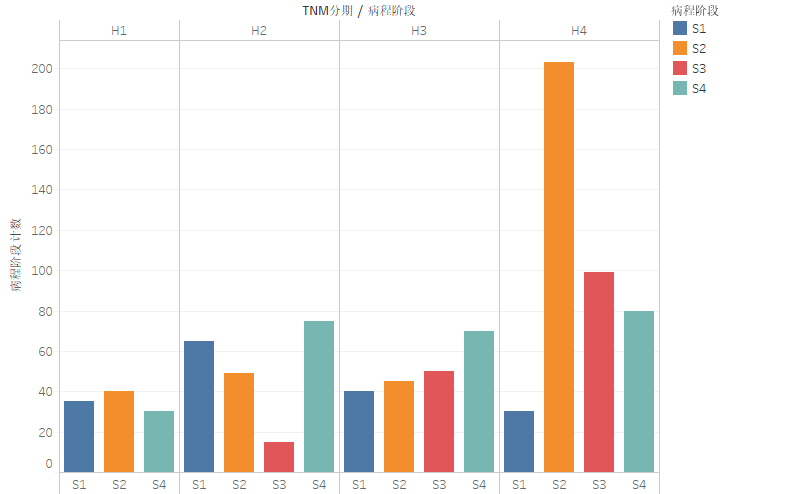
（2）不同病程阶段的病素分布



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 病程阶段 | 冲任失调证型系数均值 | 气血两虚证型系数均值 | 热毒蕴结证型系数均值 | 肝气郁结证型系数均值 | 肝肾阴虚证型系数均值 | 脾胃虚弱证型系数均值 |
| S1 | 0.2378 | 0.2158 | 0.1992 | 0.2376 | 0.2148 | 0.1359 |
| S2 | 0.2407 | 0.2443 | 0.1524 | 0.2380 | 0.2630 | 0.3272 |
| S3 | 0.2282 | 0.2083 | 0.4091 | 0.2162 | 0.3428 | 0.1948 |
| S4 | 0.2741 | 0.1884 | 0.1809 | 0.2306 | 0.2749 | 0.1762 |

从样本数据得知S1围手术期各病素比较低，脾胃虚弱症型系数最低；S2围化疗阶段，脾胃虚弱症型系数是所有阶段最高的；S3围放疗期，热毒蕴结证型系数和肝肾阴虚证型系数上升；S4巩固期冲任失调证型系数、肝肾阴虚证型系数，肝气郁结证型系数高于另外病素系数。对于病程阶段的治疗有一定的指导作用。

（3）TNN分期和病程阶段关系



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 病程阶段 | | | |
| TNM分期 | S1围手术期 | S2围化疗期 | S3围放疗期 | S4巩固期 |
| H1 | 35 | 40 | 0 | 30 |
| H2 | 65 | 49 | 15 | 75 |
| H3 | 40 | 45 | 50 | 70 |
| H4 | 30 | 203 | 99 | 80 |

H1：S1围手术期、S2围化疗期、S4巩固期的病人数差别不大，S3围放疗期的病人基本没有。

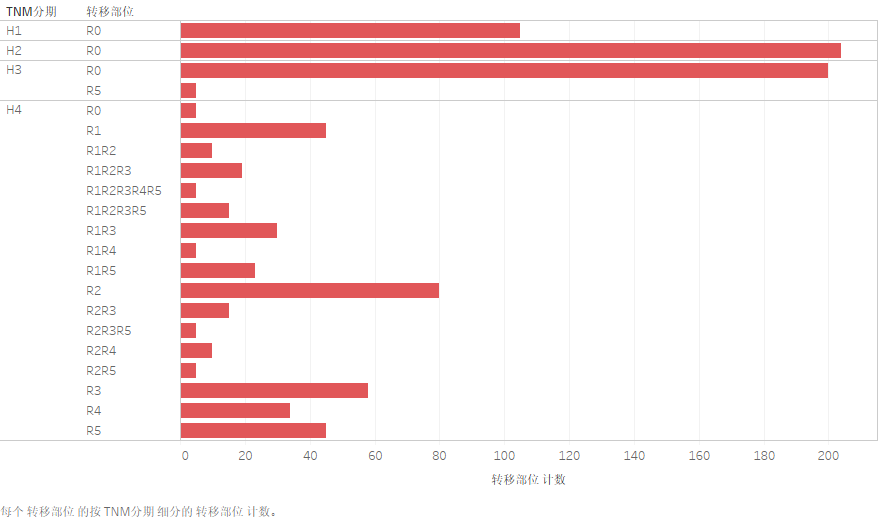
H2：处于S1围手术期和S4巩固期的病人数较多，S2围化疗期的病人数次之，S3围放疗期的病人数则较少

H3：S4巩固期病人数最多，其次是S3围放疗期的病人数，再其次是S2围化疗期的病人数，最后是S1围手术期的病人数

H4：处于S2围化疗期的病人数是最多的，达到203人，其次是S3围放疗期的病人数，再其次是S4巩固期病人数，最后是S1围手术期的病人数。

随着TNN期数的增加，化疗、放疗人数明显增加，尤其是癌症后期。初期可以选择手术治疗，后期选择化疗、放疗治疗。

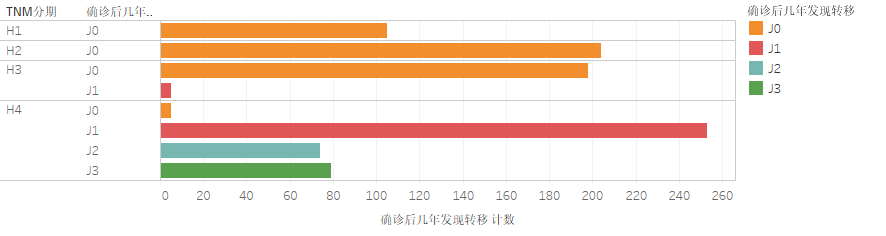
（4）TNN分期和转移部位的关系

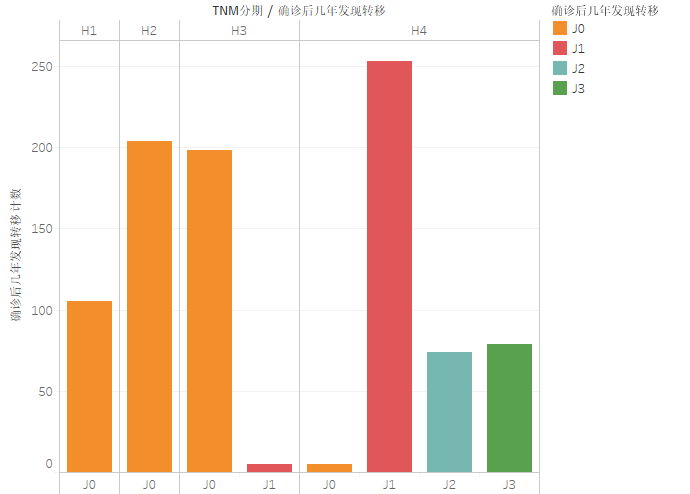




分析样本数据可以发现TNN初期H1、H2基本无转移，但在H3期 转移到其它部位，到了末期H4期基本是扩散到多部位（骨、肺、脑、肝，其它部位），基本上早发现早治疗，到了末期癌症扩散会很严重。

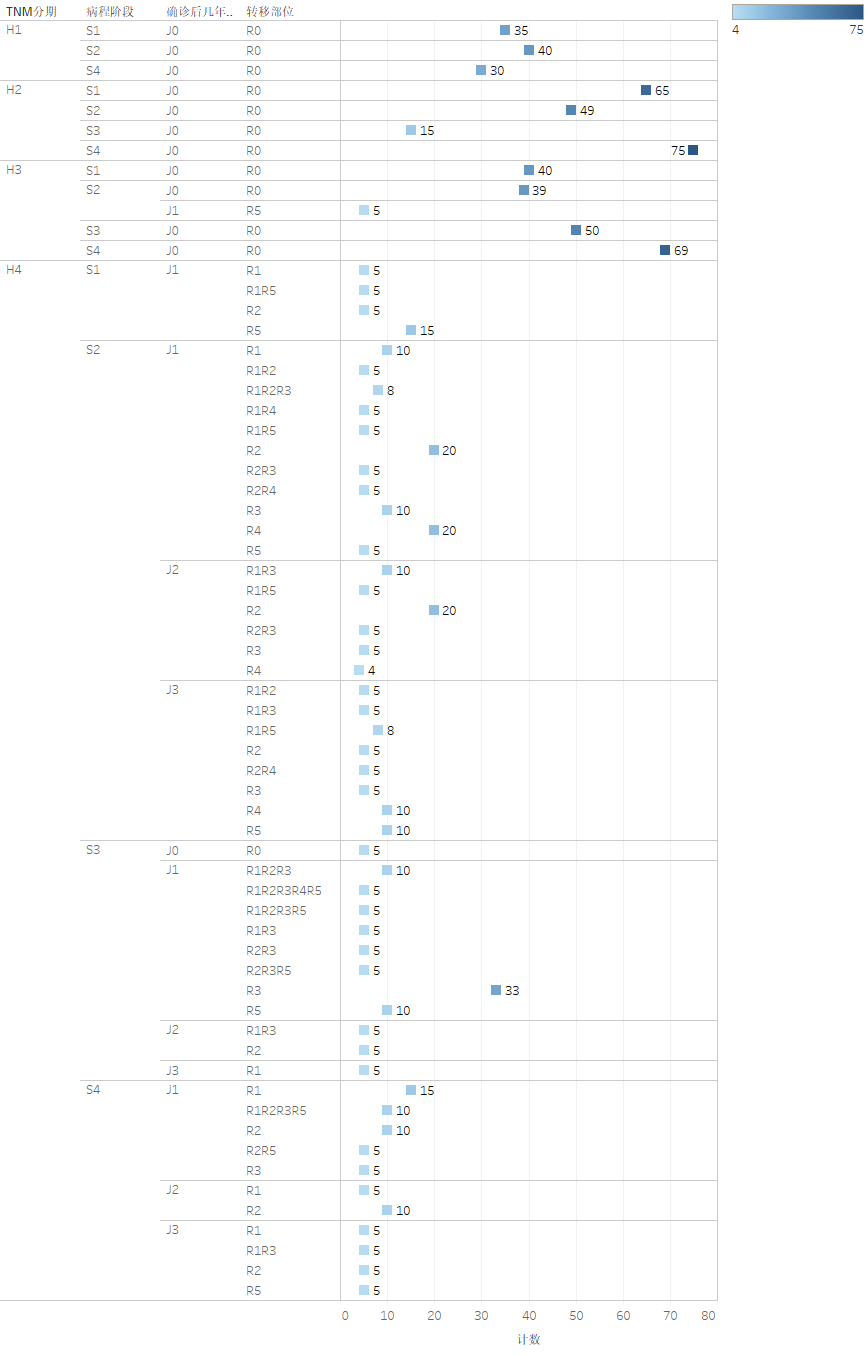
（5）确诊几年发现转移时间和TNN分期关系





分析样本数据可以发现，前期确诊转移的机会小，后期确诊转移时间较快，病情会难以控制。

（6）综合分析



综合分析样本数据，在癌症初期确诊转移的风险不大，初期一般选择手术治疗，末期确诊时转移情况加剧，癌症会扩散转移，而不是单个部位的转移，此时采取围疗、化疗居多。

## 总结和展望

通过以上的研究和实验，我们挖掘出与TNM分期相关的关联规则，并实现了对TNM分期的预测功能。经过这次研究，我们对乳腺癌、中医症素、治疗分期之间的关联有了较为深入的认识，同时我们实现了多个针对TNM分期的分类器，并且能在数据支持下提供不错的分类效果。本次研究中讨论并实现的分类器能够应对实际医疗场景的需要，具有很强的实际意义。接下来我们将进一步改进我们的方法，对关联规则提取、分类器构建进行优化，对其中的参数进行更多的调试，以期获得更好的预测效果，解决实际中TNM分期的分类问题。

同时，通过这一系列的提出问题和解决问题的过程不仅锻炼了我们在数据挖掘动手实践方面的能力更有效的锻炼了我们组成员的严谨的思维。相信以后面临更复杂的问题，我们也能够提出合理高效的解决方案。