数据整理以及关联规则挖掘

Method

1. 数据准备

数据准备阶段，我们要仔细审查整个数据库的数据。我们需要统计数据的分布情况，看样本是否来自随机采样。我们从男性与女性数目，每个数据条目危险因素数目上进行统计。使用分布拟合和假设检验的方法，对数据的随机性进行检验。对危险因素初筛表中的标准进行验证，看看经验初筛表对脑卒中危险因素的划分情况。在初筛表数据的统计中，我们将脑卒中发病先作为一个隐藏因素，先将高中低危人群划分完毕，在其中统计每个人群中脑卒中发病情况。

1. 关联规则挖掘

我们的工作首先对初筛表中，9个危险因素（dfTia、dfHypertension、dfAF、dfSmoking、dfLDL(将未知视为异常)、dfGlycuresis、dfSportsLack、dfOverweight、dfStrokeFamily）与脑卒中发病的关系进行关联规则挖掘。对9个关联规则进行全排列，共有511种组合，我们对每个组合进行全局扫描，统计其支持度、置信度、提升度等参数。在关联规则中，先导（antecedent）为危险因素的组合，后继(consequent)为是否为脑卒中患者。

1. 关联规则参数选择

在关联规则挖掘中，我们选择支持度（support），置信度（confidence），提升度（lift）。

规则的先导为I={I1,I2,..Im}，是危险因素项的集合，后继为X，表示是否为脑卒中患者。

其中支持度为数据库中，同时包含I，X的百分比，即概率。支持度描述这个规则的广泛性，即是否普遍。

置信度为，在数据库包含I的所有事务中，包含X的百分比。置信度描述这个规则是否有趣（interesting）。但是我们可以发现，如果数据库中包含X的事务远远多于I，那么可能置信度会比较高，但是这种规则并不是我们想要的有趣的规则。因此我们引入提升度（lift）.

提升度定义为：，即用置信度比上X的概率，消除如果X比I普遍很多的情况。提升度描述了规则之间的关联，如果大于1，则表示正相关，提升度越大相关性越大。

Result

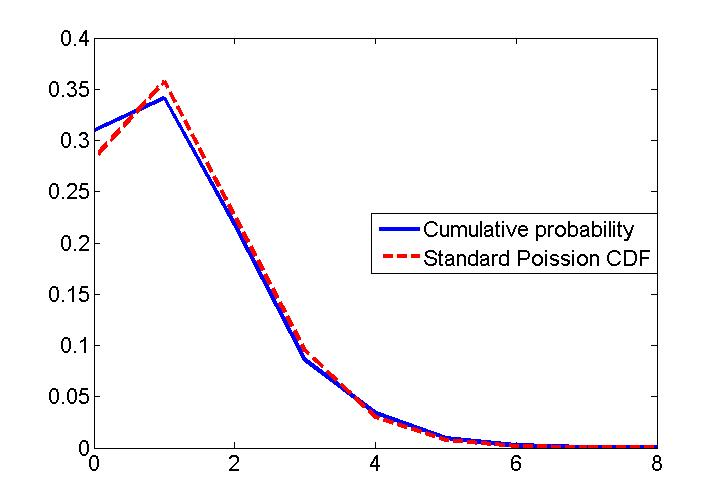
1. 数据库大概

脑卒中筛查数据库中，共有数据862,244条，没有重复数据条目。其中脑卒中发病人数16862人，占总人数的1.96%。数据中，男性人数397765，占总人数的46.13%，女性占总人数的53.87%。中国实际人口比例数据（第六次人口普查）中，男性占51.27%。由于男性人数在人群中服从二项分布，当人群数目足够大时，近似服从正态分布。通过计算可知，调查表中男性数据近似服从总人口的分布，可知抽样的男性女性也是随机的。

按照初筛结果中，对每个条目的危险因素数目进行统计，得到结果如图：

危险因素数目近似服从泊松分布。泊松分布描述单位时间内随机事件发生次数，这与危险因素贴合，可知样本独立随机。我们首先对这个分布进行分布拟合，其服从lambda=1.2629的泊松分布。

各个危险因素在人群中的比例，下图是对危险因素进行泊松分布的假设检验，其中蓝线为真是病例数据的累计概率分布，红线为（lambda=1.2629）的理想泊松分布。



从这个危险因素数目在人群中的分布统计图中，我们也可以看出，数据抽取是独立随机的。

我们首先统计初筛结果的各个具体数值。

下表给出每个危险因素在人群中所占的比例。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | dfStroke | TIA | dfHypertension | dfAF | dfSmoking | dfLDL | dfGlycuresis | dfSportsLack | dfOverweight | dfStrokeFamily |
| positive | 1.96% | 3% | 21.20% | 2.70% | 18.10% | 31.30% | 5.80% | 19.70% | 17.30% | 7% |
| negtive | 98．04% | 97% | 78.80% | 97.30% | 81.90% | 68.70% | 94.20% | 80.30% | 82.70% | 93% |

对初筛表中数据的挖掘统计结果如下表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 高危 | 中危 | 低危 | 总和 |
| 发病人数 | 10603 | 4541 | 1718 | 16862 |
| 划分的总人数 | 129963 | 272709 | 459572 | 862244 |
| 所占人群的比例 | 8.16% | 1.67% | 0.37% | 1.96% |
| 所占脑卒中人群比例 | 62.88% | 26.93% | 10.19% | 100% |

但是，通过统计，有265145个被调查者不具备任何危险因素，占所有人群的30.75%，其中有656个脑卒中患者不具备任何危险因素。这656人占所有脑卒中患者的3.89%。



统计数据库中各个年龄段占总人数的比例，并且统计每个年龄段发病概率，做出上面的统计图。可以看到，随着年龄的上升，发病比例也在升高，70岁年龄段的发病比例在3.96%

1. 关联规则挖掘

由上面的结果可知，每个危险因素在人群中的比例都很低，而组合的规则比例更低。我们定义零事务，如果一个数据条目即不包含备选的危险因素也不包含是脑卒中，那么它就是一个零事务。零事务对支持度的影响是很大的。因为有大量的零事务出现（不包含我们需要测试的规则且不是脑卒中患者），因此，规则挖掘的过程中最小支持度要设置的比较小。

我们首先对每个单独的危险因素计算其支持度置信度和提升度。

单独因素的结果如下表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 因素 | support | confidence | lift | Kulczynski度量 |
| dfTia | 0.55% | 16.93% | 8.66 | 0.23 |
| dfHypertension | 1.34% | 6.31% | 3.23 | 0.37 |
| dfAF | 0.22% | 7.95% | 4.07 | 0.10 |
| dfSmoking | 0.42% | 2.34% | 1.20 | 0.12 |
| bloodfat | 0.99% | 3.15% | 1.61 | 0.27 |
| dfGlycuresis | 0.36% | 6.30% | 3.22 | 0.12 |
| dfSportsLack | 0.58% | 2.94% | 1.50 | 0.16 |
| dfOverweight | 0.63% | 3.64% | 1.86 | 0.18 |
| dfStrokeFamily | 0.48% | 6.92% | 3.54 | 0.16 |

由上表中，可以看到，除了高血压以外的危险因素支持度都低于1%，而如房颤（AF），其支持度低于0.2%，也就是说在整个数据库中，只有1897人是同时具有房颤和脑卒中发病的。

因此，在我们的计算中，选取最小支持度=0.1%，最小置信度=10%。在这个支持度下，最少覆盖862人，即脑卒中发病人数5%。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 频繁项集 | Support | Confidence | Lift | Kulc | IR |
| dfTia | 0.55% | 16.93% | 8.66 | 0.23 | 0.28 |
| dfTia,dfHypertension | 0.37% | 22.43% | 11.47 | 0.21 | 0.09 |
| dfTia,dfAF | 0.11% | 23.94% | 12.24 | 0.15 | 0.66 |
| dfTia,dfSmoking | 0.11% | 18.98% | 9.71 | 0.12 | 0.55 |
| dfTia,bloodfat | 0.28% | 15.99% | 8.17 | 0.15 | 0.05 |
| dfTia,dfSportsLack | 0.19% | 18.95% | 9.69 | 0.14 | 0.34 |
| dfTia,dfOverweight | 0.17% | 20.43% | 10.45 | 0.15 | 0.42 |
| dfTia,dfStrokeFamily | 0.16% | 21.52% | 11.00 | 0.15 | 0.49 |
| dfHypertension,dfAF | 0.15% | 11.60% | 5.93 | 0.10 | 0.21 |
| dfHypertension,dfGlycuresis | 0.27% | 10.24% | 5.24 | 0.12 | 0.16 |
| dfHypertension,dfStrokeFamily | 0.36% | 11.81% | 6.04 | 0.15 | 0.24 |
| dfTia,dfHypertension,bloodfat | 0.20% | 21.73% | 11.11 | 0.16 | 0.38 |
| dfTia,dfHypertension,dfSportsLack | 0.14% | 24.78% | 12.67 | 0.16 | 0.59 |
| dfTia,dfHypertension,dfOverweight | 0.13% | 24.52% | 12.54 | 0.16 | 0.61 |
| dfTia,dfHypertension,dfStrokeFamily | 0.11% | 27.17% | 13.89 | 0.16 | 0.68 |
| dfTia,bloodfat,dfSportsLack | 0.11% | 18.58% | 9.50 | 0.12 | 0.57 |
| dfHypertension,bloodfat,dfGlycuresis | 0.17% | 11.66% | 5.96 | 0.10 | 0.17 |
| dfHypertension,bloodfat,dfStrokeFamily | 0.21% | 12.50% | 6.39 | 0.12 | 0.08 |
| dfHypertension,dfGlycuresis,dfOverweight | 0.12% | 11.77% | 6.02 | 0.09 | 0.34 |
| dfHypertension,dfSportsLack,dfStrokeFamily | 0.12% | 13.20% | 6.75 | 0.10 | 0.39 |
| dfHypertension,dfOverweight,dfStrokeFamily | 0.14% | 12.52% | 6.40 | 0.10 | 0.27 |

|  |  |  |
| --- | --- | --- |
| **各危险因素** | **出现频次** | **在出现规则中占百分比** |
| dfTia | 13 | 0.62% |
| dfHypertension | 13 | 0.62% |
| dfAF | 2 | 0.10% |
| dfSmoking | 1 | 0.05% |
| bloodfat | 5 | 0.24% |
| dfGlycuresis | 3 | 0.14% |
| dfSportsLack | 4 | 0.19% |
| dfOverweight | 4 | 0.19% |
| dfStrokeFamily | 6 | 0.29% |

1. 新关联规则筛选与分析

我们的目的是更加准确的识别高危人群，在保证规则简单易操作的前提下，我们提出了新的危险划分规则。

这个规则的目的是，减少划入低危人群的发病可能，增加被划入高危人群的发病可能。这样一来，如果被诊断为低危，那么其发病可能尽量小。但如果被划入高危，那么就要更加注意。从上表中可以看到，高血压的出现频率是12次，仅次于TIA的13次，吸烟的出现频率最低，只有一次。而支持度超过20%，提升度超过10的规则中，都含两项或以上危险因素，且有TIA或高血压。因此，我们提出的新的规则如下：

1. 患有TIA的病人直接划入脑卒中高危人群
2. 在除了TIA和吸烟的7项危险因素中，有3项以上（包括3项）的，且患有高血压的患者，划入高危人群
3. 在除TIA外的8个危险因素中，有一个以上（包括1个），划入中危人群
4. 除了TIA，高血压，家族史，其余危险因素有1个以下（包含1个）划入低危人群

按照新的危险因素和危险划分方式，我们得到以下结果：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 高危 | 中危 | 低危 | 总和 |
| 发病人数 | 6667 | 9539 | 656 | 16862 |
| 划分的总人数 | 49853 | 547246 | 265145 | 862244 |
| 所占人群的比例 | 13.37% | 1.74% | 0.25% | 1.96% |
| 所占脑卒中人群比例 | 39.54% | 56.57% | 3.89% | 100% |

实际上，根据前面的统计我们知道，划入低危人群的656人属于无法判别组，即他们没有任何一个危险因素。

为了对新旧规则进行比对，我们分别计算新旧规则的提升度，结果如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 高危 | 中危 | 低危 |
| 旧规则发病率 | 8.16% | 1.67% | 0.37% |
| 新规则发病率 | 13.37% | 1.74% | 0.25% |
| 旧规则提升度 | 4.1719 | 0.4244 | 0.1912 |
| 新规则提升度 | 6.8385 | 0.8913 | 0.1265 |

提升度低于1的，表示其规则和结果不是正相关。可以看到，中危和低危的提升度都小于1。而高危的规则中，新规则的提升度明显高于旧规则，提升度高说明分类准确率高。

将结果展示如下图所示：

可以看到，新规则在高危和中危的判别提升度中都有显著提高，低危的判别提升度下降。在发病比率中，划分为高危的人群发病率也提升很多，低危人群的发病比率下降，也符合我们的预期。