说明文档

[1.项目概述及人员： 2](#_Toc15427)

[1.1项目概述 2](#_Toc20796)

[1.2 人员 2](#_Toc13816)

[2.需求分析 2](#_Toc21567)

[3.系统设计 3](#_Toc11688)

[3.1 总体设计 3](#_Toc30622)

[3.2 系统功能模块 3](#_Toc32349)

[4.代码框架及函数设计 8](#_Toc18096)

[4.1.实现环境： 8](#_Toc30902)

[4.2.主要函数设计 9](#_Toc8087)

[5.实验 11](#_Toc6728)

[5.1.实验环境： 11](#_Toc10663)

[5.2.实验步骤： 11](#_Toc21731)

[5.3实验结果分析： 12](#_Toc19)

[6.实验总结 12](#_Toc3337)

1.项目概述及人员：

## 1.1项目概述

针对乳腺癌患者，可以依据中医阶段治疗法进行治疗，在辨症的基础上围绕各个病程的特殊症候先症而治型。依据医学指南，将乳腺癌辨证统一化，分为六种症型。即患者在围手术期、围化疗期、围放疗期和内分泌治疗期，基本都会出现特定的临床症状。可以通过一定的规则，挖掘出各中医症素与乳腺癌TNM分期之间的关系。探索不同分期阶段的乳腺癌患者的中医症素分布规律，以及阶段病变发展（71，72，73），指导乳腺癌的中医临床治疗。

（1）对数据进行预处理；

（2）利用数据挖掘的算法对数据进行分析，得到中医症素与乳腺癌TNM 分期之间的关系。根据实际的业务分析建议最小支持度设定为：6%，最小置信度为：75%；

（3）对所构建的模型进行分析，得出你的结论。

## 1.2 人员

|  |  |  |
| --- | --- | --- |
| 姓名 | 学号 | 分工 |
| 刘昕 | 2016E8018661172 | 贝叶斯算法，文档 |
| 李博伟 |  | svm，预处理 |
| 蒋浩丰 | 201628018629015 | 神经网络 |
| 苏莉娅 | 201628018629065 | 决策树，框架设计 |
| 昝文 | 201628018629015 | k近邻,图形界面 |

# 2.需求分析

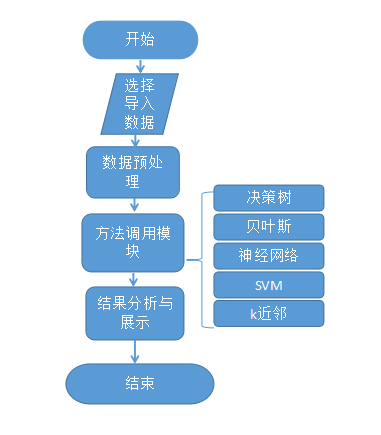
（1）功能需求：利用数据挖掘的算法对数据进行分析，得到中医症素与乳腺癌TNM 分期之间的关系。根据实际的业务分析建议最小支持度设定为：6%，最小置信度为：75%；对所构建的模型进行分析，得出你的结论。

（2）性能需求：满足多算法优势比较。满足正确率在90%以上，错误率低于10%，运算速度可接受。

采用图形化界面，有统一框架，可以集中调用算法进行输出。

# 3.系统设计

## 3.1 总体设计



## 3.2 系统功能模块

### 3.2.1 预处理

在预处理阶段，我们采用箱均值的方法对数据进行光滑或者直接去除空值的方法去噪。经过比较,去除空值的方法准确率较高，故最终采用了去空值的方案去去除噪声。

### 3.2.2 分类算法：

我们采用了决策树、朴素贝叶斯、SVM、K-近邻、神经网络的方法对训练样本进行分类学习。

#### 决策树：

###### 原理：

决策树是一个预测模型，代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表的某个可能的属性值，而每个叶结点则对应从根节点到该叶节点所经历的路径所表示的对象的值。决策树一般自上而下生成的。分割的目的一致：对目标类尝试进行最佳的分割。

首先利用训练数据集来构造一棵决策树，树建立起来就可为未知样本产生一个分类。决策树模型有很多的优点、便于使用，高效、决策树的大小独立于数据库、、可以对有许多属性的数据集构造决策树。但决策树模型也有过拟合等问题的缺点。

###### 实现：

本文采用的是ID3的决策树的方式进行实现。ID3算法是决策树的一种，它是基于奥卡姆剃刀原理，即用尽量用较少的东西做更多的事。在信息论中，期望信息越小，那么信息增益就越大，从而纯度就越高。ID3算法的核心思想是在决策树的每一个非叶子结点划分之前，先计算每一个属性所带来的信息增益，选择最大信息增益的属性来划分。因为信息增益越大，区分样本的能力就越强，越具有代表性。该算法采用自顶向下的贪婪搜索遍历可能的决策空间。

#### 朴素贝叶斯：

###### 原理：

贝叶斯分类器的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。

朴素贝叶斯模型拥有稳定的分类效率,所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。因为NBC模型假设属性之间相互独立，这个假设在实际应用中往往是不成立的，这给NBC模型的正确分类带来了一定影响。在属性相关性较小时，NBC模型的性能最为良好。

###### 实现：

朴素贝叶斯的三个常用模型：高斯、多项式、伯努利实现,本次的实验数据为连续值，适合采用高斯分布的方式进行计算。

本次采用这个公式为基础实现朴素贝叶斯的高斯模型。

#### SVM：

###### 原理：

支持向量机(support vector machine)是一种分类算法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。

SVM是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，即支持向量机的学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解.

###### 实现：

我们实现SVM的方法是调用libsvm的库。LIBSVM是台湾大学林智仁(LinChih-Jen)副教授等开发设计的一个简单、易于使用和快速有效的SVM模式识别与回归的软件包。由于libSVM程序小，运用灵活，输入参数少，并且是开源的，易于扩展，因此成为目前国内应用最多的SVM的库。

#### K-近邻：

###### 原理：

K最近邻(k-Nearest Neighbor，KNN)分类算法的思路是：

如果一个样本在特征空间中的k个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别。

KNN算法中，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

KNN方法在类别决策时，只与极少量的相邻样本有关。由于KNN方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别，KNN方法更适合类域的交叉或重叠较多的待分样本集。

KNN算法的主要不足是，当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的K个邻居中大容量类的样本占多数。KNN算法的另一个不足之处是计算量大。该算法比较适用于样本容量比较大的类域的自动分类，而那些样本容量较小的类域采用这种算法比较容易产生误分。

###### 实现：

设定参数k,并计算范式距离。维护一个大小为k的的按距离由大到小的优先级队列，用于存储最近邻训练元组。随机从训练元组中选取k个元组作为初始的最近邻元组，分别计算测试元组到这k个元组的距离，将训练元组标号和距离存入优先级队列，遍历并比较训练元祖集并不断更新，确保集内为范式距离最小的K个元祖。遍历完毕，计算优先级队列中k 个元组的多数类，并将其作为测试元组的类别。

#### 神经网络：

###### 原理：

神经网络分为前向传输和逆向反馈，我们实现的是逆向反馈及BP神经网络。训练一个BP神经网络，实际上就是调整网络的权重和偏置这两个参数，BP神经网络的训练过程分两部分：

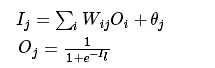
前向传输，逐层波浪式的传递输出值；

逆向反馈，反向逐层调整权重和偏置；

(需要补充,等待相关的代码)

###### 实现：

根据下面的公式



由上面两个公式算出隐藏层节点进而算出输出层节点

根据下面的公式算出输出层节点的偏差值

IMG_256

中Ej表示第j个结点的误差值，Oj表示第j个结点的输出值，Tj记录输出值

根据下面的公式算出隐藏层节点的偏差值

IMG_256Wjk表示当前层的结点j到下一层的结点k的权重值，Ek表示下一层的结点k的误差率

再根据下式更新各个节点的权值和偏置

IMG_256

IMG_256

如此对每一个训练节点进行训练，不断更新各个节点的权值，重复N次。

### 3.2.3 测试

我们采用的是交叉验证的方法划分测试集与训练集进行分类器的训练进行测试。

交叉验证的基本思想是把在某种意义下将原始数据(dataset)进行分组,一部分做为训练集(train set),另一部分做为验证集(validation set or test set),首先用训练集对分类器进行训练,再利用验证集来测试训练得到的模型(model),以此来做为评价分类器的性能指标.

本次实验我们采用的是K折交叉验证，初始采样分割成K个子样本，一个单独的子样本被保留作为验证模型的数据，其他K-1个样本用来训练。交叉验证重复K次，每个子样本验证一次，平均K次的结果最终得到一个单一估测。

### 3.2.4 功能集成：

我们通过web的图形界面进行分类器的集成和结果展示.

实现：

我们引入了flask的库实现web界面的开发。Flask是一个使用 Python 编写的轻量级 Web 应用框架，易于上手。

# 4.代码框架及函数设计

## 4.1.实现环境：

代码实现语言：python

编码环境：windows

## 4.2.主要函数设计

### 4.2.1数据预处理:

**代码路径：**

LIBOWEI/preprocessing.py

**主要功能函数：**

1>fillNullWithMean# 用均值填补空值

2>deleteOutliers #计算统计异常值边界,并去掉边界值

### 4.2.2分类算法

#### 决策树：ID3算法

**代码路径：**

suliya/ID3ForContinuousValue.py

**主要功能函数：**

1>calcShannonEnt(dataSet):#计算数据集中的香农熵，看数据的标签分布

2>chooseBestFeatureAndValueToSplit(dataSet,labels):#找到最大熵增益的特征及特征取值

3>createTree(dataSet,labels):#创建决策树，使用熵增益评判

#### SVM:

**代码路径：**

libowei/svm.py

**主要功能函数：**

1>get\_SV(self):获得支持向量

2>svm\_train(arg1, arg2=None, arg3=None):SVM训练函数

3>svm\_predict(y, x, m, options=""):SVM预测函数

#### K近邻法：

**代码路径：**

zanwen/knn.py

**主要功能函数：**

1> classify(inX, dataSet, labels, k, lp):分类函数

2> getLpDistances(lp, inX, dataSet):# 计算两个向量的lp范数

#### 朴素贝叶斯：

**代码路径:**

liuxin/NB.py

**主要功能函数:**

1>summarizeByClass(dataMat):#通过计数统计获得分类器

2>calculateClassProbabilities(summaries, inputVector):#通过高斯函数计算概率。朴素贝叶斯高斯模型的核心算法。

#### 神经网络：

**代码路径：**

handle\_data.py

bp\_network\_test.py

**主要功能函数：**

handle\_data():#对xls数据进行处理，获取前六列与第八列数据放入一个TXT中，如果相应的行有空值，则舍去该行值。

input\_data():#用于获取TXT中的数据，并将其对变量赋值，随机选取900条数据中的700条作为训练数据，200条数据作为预测数据。

train():#用于训练数据，最终得到神经网络各个节点对应的参数矩阵。

predict():#用于预测数据，最终得到一个预测成功的数据的个数。

### 4.2.3 展示界面：

##### 代码路径:

zanwen/gui.py

# 5.实验

## 5.1.实验环境：

操作系统：windows

编程语言： python

## 5.2.实验步骤：

### 1)数据处理

本文的数据中包含着一些空值，需要通过数据预处理的方式处理噪音。

### 2)特征选择：

按照题目要求，本文选择的特征为“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”，并以TNM分期为目标进行分类学习。

### 3)调用分类方法

在对数据进行预处理去除噪音之后，划分训练集和测试集，分别调用KNN，朴素贝叶斯,决策树，神经网络,SVM方法训练分类器

### 4）运行并测试

对测试集运行分类器，检验分类结果。并通过交叉验证法测定正确率和覆盖率，检验分类器的效果。

### 5）展示：

通过图形界面集中展示实验结果。

## 5.3实验结果分析：

### 5.3.1准确性分析：

### 5.3.2可扩展性：

### 5.3.3效率：

### 5.3.4合作性：

### 5.3.4新颖性：

# 6.实验总结

参考文献