1 前言

1.1 国内外研究背景与现状

通常，数据集包含有大量的特征。并且随着世界的快速发展，数据的维度和大小在大多数领域里都是快速成长的。而提取数据集中相关特征的方法被称为特征选择。属性约简是一种重要且有用的方法，可以有效的减少特征空间的维度，并且在表示原始数据时保持相对较高的精度，还能增强模型的可解释性，加速学习过程并且提高学习性能。

高维数据集对机器学习来说是一个巨大挑战。一些最具实用性和高影响力的应用程度（例如基因表达数据）很可能有超过10000个特征。而很多特征可能与当前的任务无关或者与其他特征无关。在这种情况下，学习会带来很多问题，比如对不相关数据的过度拟合，以及处理许多提供冗余信息的相似特征的计算负担。因此，找到包含有效变量的较小子集，也就是属性约简，是一个重要的研究方向。

属性约简是特征工程中的重要问题，人们通常说：数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。随着新技术的不断出现，在模式识别系统中，数据的显著特点是样本少、维数高，这给传统的分类学习带来了挑战。在许多新兴领域，如生物信息学，他们所面临的数据集有大量的特征（高达上千个），大量的噪声、非线性依赖并且常常只有数百个样本。

在这种情况下，发现功能关系和设计有效的分类器似乎是一项巨大的挑战。近来技术的进步，如微阵列技术，使得在一个生物样本中同时检测数千个基因成为可能。由此可见，生物信息学中经常遇到的两个分类问题是怎样区分肿瘤类型和如何在微阵列基因表达谱上预测医学治疗的效果。若果我们将这种预测任务形式化为监督分类问题，那么我们会看到以基因数目为代表的输入变量的数目是巨大的（大约数千个），而以临床实验代表的样本是十分有限的（大约几十个）。

因此，在生物信息学中的分类技术需要有处理多变量少样本数据集的能力（也就是特征-样本比高的数据集）。在众所周知的数字和统计精度问题上，通常在开始一个学习过程之前，需要对减少变量的数目。此外，特征选择可以增加一个模型的可解释性，同时减小度量和存储需求。大量的实验表明，不相关和冗余的特征可以极大的降低由数据所生成模型的预测准确性。所以如何去除高维数据集的冗余，提取出有用特征，以免发生维数灾难的问题，这是当前研究人员面临的一道难题。

属性约简是机器学习的一个主题，其目标是在一组输入变量中选出可以最佳化预测模型的变量。我们可以将这个问题表述为随即优化问题，因为搜索最佳子集必须根据与给定子集相关的准确性统计估计来进行。有两种知名的属性约简方法将搜索策略与随机评估函数相结合：过滤式方法和封装式方法。在封装式方法中，评估函数是学习算法本身的验证结果。在过滤式方法中，有些评估函数是概率距离、类间距离、信息理论或者是概率依赖度。通常，这些衡量标准是数据的固有属性，因为它们由原始数据计算得到，并不需要一种平滑分布或降噪的学习模型。

特征选择算法的分类有很多种，根据样本中是否含有类别信息可将其分为有监督特征选择和无监督特征选择。根据子集的生成方式可将其分为搜索性算法和非搜索性算法。此外，根据所采取的评价准则与后续分类器是否联合一起使用，可将其分为Filter（过滤式）算法、Wrapper（封装式）算法、Embedded（嵌入式）算法。其中Filter算法具有较强的通用性，省去了对分类器的训练，所以算法不复杂，当数据维数较大时，计算量也相对较低，所以Filter（过滤式）算法的应用比较广泛。Filter算法的准则基本上是基于信息理论的，比如DISR算法、MRMR算法、JMI算法、FCBF算法等，这些算法都是有针对性的，有针对相关性的，有针对冗余性的，也有针对互补性的。

由于数据集中变量数量的增加，搜索算法已经从穷举搜索发展到启发式搜索算法，并且评估函数也从学习算法发展到快速的相关性度量，比如互信息。为了使搜索和渲染算法更快，相关性和冗余性也被正式定义了。

UC Berkeley近日提出了一种新型特征选择方法条件协方差最小化（Conditional Covariance Minimization， CCM），该方法基于最小化条件协方差算子的迹来进行特征选择。研究者的实验证明该方法在多个合成和现实的数据集上达到了不输当前先进方法的性能。其论文《Kernel Feature Selection via Conditional Covariance Minimization》被NIPS 2017接收。

1.2 研究内容与目标

高维数据集会因特征维数过高而引发“维数灾难”，过大的计算复杂度会影响数据集的有效使用。另外，可能并非所有维度的数据或概念都同样包含重要的预测信息。因而，特征选择算法作为数据的降维技术显得尤为重要。总体来讲,特征选择算法中的核心问题包括计算：（1）特征子集与预测类别的相关性；（2）特征间包含分类信息的冗余性；（3）平衡（1）和（2）二者的权重比例。

本次所选课题的研究目标是要求针对特定的高维集合，在MATLAB环境下，使用基于信息论的相关概念如信息量、熵、条件互信息等分析样本参数之间的互补性，并对其有效利用，实现基于特征互补性的高维数据集属性约简算法，给出实验数据和实验波形的分析并得出相应结论。将所研究的内容与应用相结合，探讨研究内容的理论和应用意义。

本次研究的主要内容是用DISR算法对iris、semeion、arcene、gisette数据集达到降维的目的，并且用SVM进行分类，对SVM分类器进行参数寻优，并与其他算法进行比较，结合正确率和时间复杂度等对算法的性能效果进行思考、研究和比较。

在毕业设计完成的过程中，熟悉数字信号处理的相关知识、熟悉信号与系统的专业知识，并对所研究的内容进行MATLAB仿真，对仿真数据和仿真图形进行分析、研究和对比。同时，对算法原理、实现、有效性等方面进行研究和探讨。

1.3 本文章节安排

本文的章节安排如下：

第一章粗略的讲解了特征选择算法的国内外研究背景和特征选择算法的研究意义，简单概括了本次课题的研究目标和主要内容。

第二章介绍进行特征选择算法时运用的信息论和特征选择的相关知识，如信息量、条件熵、互信息等，以及特征选择方法的定义、步骤和分类。

第三章主要介绍了本次课题所用到的DISR算法的原理、代码实现和调用方式等，并介绍一些用于进行性能对比的其他特征选择算法的原理。

第四章主要是分析实验结果，对DISR算法对不同数据集的选择效果进行对比，然后与其他特征选择算法得到的结果进行对比分析。

2 基于信息理论的属性约简算法

2.1 属性约简算法概述

在机器学习和统计学中，特征选择也叫做变量选择、属性选择或者变量子集选择，是一个选择相关特征（变量、预测变量）的子集的过程，并将其用在模型的构建中。运用特征选择技术主要由以下四个原因：

1. 简化模型来使研究人员或者用户更容易解释，
2. 缩短训练时间
3. 避免维数灾难
4. 通过降低过适应来增强泛化能力。

使用特征选择的中心前提是数据中包含许多冗余或者不相关的特征，使之在不造成大量信息损失的情况下被移除。冗余或者不相关的特征是两个截然不同的概念，因为一个相关特征可能在另一个与它强相关的特征面前是冗余的。

2.1.1 特征选择简介

特征选择可以看作是搜索技术和评价方法的结合，前者用于提出新的特征子集，后者可以对不同的特征子集进行评分。

最简单的算法使测试每个可能的特征子集，找到使错误率最小化的一个特征子集。这是对空间的彻底搜索，并且除了最小的特征集之外，其他都是难以计算的。评估准则的选取对算法产生很大的影响，并且根据评估准则将特征选择算法分为三个主要的类别：封装式、过滤式和嵌入式的方法。

2.1.2 特征选择算法分类

封装式方法用一个预测模型来对特征子集打分。每个新的自己用来训练一个模型，并在一个固定的集合上进行测试。通过计算在固定集合上错误数量（模型的错误率）来给这个自己打分。由于封装方法为每一个子集训练一个新模型，所以它的计算量特别大，但是通常可以为特定类型的模型提供最佳性能的特征集。

过滤式方法用一个错误率的替代方法来对特征子集打分。这个方法被用来进行快速计算，同时仍然可以保证特征集的有效性。常用的方法包括互信息、点间互信息、皮尔森积矩相关系数、基于Relief的算法等。过滤式方法通常比封装式方法的计算量小，但是它产生的特征集合没有调整到预测模型的特定类型。许多过滤式方法仅提供了一个特征排序，而不是一个显性的最佳特征子集。

嵌入式方法是一种综合的技术，它可以将特征选择作为模型构建过程中的一部分进行执行。这种方法的范例是构造线性模型的LASSO方法，它用L1惩罚因子来惩罚回归系数，使其中的很多值减少到0。其它不是0回归系数的特征就被LASSO算法“选择”出来了。这种方法的计算复杂度介于过滤式和封装式之间。

本课题所使用的DISR算法属于过滤式方法。

2.2 属性约简算法中的信息论基础

信息论是研究信息的量化、存储和通信的学科，运用概率论与数理统计的方法研究信息、信息熵、通信系统等问题。信息论最早是由香农在1948年提出的，目的是找到信号处理与通信操作之间的基本限制。信息论的基本应用包括无损压缩（例如ZIP文件）、有损压缩（例如MP3和JPEG）和信道编码（例如数字用户线路）。

其中，熵是信息论中的重要度量。熵量化涉及随机变量值或者随机过程结果不确定性的数量。举个例子，确定一个均质硬币的翻转结果（有两种等可能性的结果）所提供的信息（熵）比确定投掷一枚骰子的结果（有六种等可能性的结果）所提供的信息（熵）要小。信息论中其他重要的度量有互信息、信道容量和相关熵。

2.2.1 信息熵

假定*X*为一个随机变量，*P*(*x*)表示随机变量*X*取值为*x*的概率，那么信息熵*H*(*X*)可以表示为：

 （2.1）

*H*(*x*)表示为*X*取值的不确定性，用来量化*X*分布中预设的不确定性，在数据集分析中，可以理解为，一个特征的信息熵越大，表明该特征所含的信息量就越多，对分类作用就越大，反之，其信息熵越小，则说明该特征所含的信息量就越少，对分类作用就越小。

2.2.2 条件熵

条件熵则是计算在另一个变量给定的基础上，该变量的不确定性程度，用于量化一个以另一个变量为条件的变量的不确定性，即该变量对另外一个变量的依赖强弱程度。

假定Y为另一个随机变量，那么随机变量*X*关于*Y*的条件熵可以表示为：

 （2.2）

这里*p*(*x*,*y*)是随机变量*X*和*Y*的联合概率分布，而*p*(*x*|*y*)是在给出变量*Y*的基础上*X*的条件概率分布。条件熵*H*(*X*|*Y*)代表在给出随机变量*Y*的基础上，计算随机变量*X*的不确定性程度。

2.2.3 联合熵

联合熵是用来计算多个变量共同含有的信息量的多少。两个随机变量*X*和*Y*，假设其联合随机变量为(*X*,*Y*)，概率分布对应为*p*(*x*,*y*)，则随机变量*X*和*Y*的联合熵为*H*(*X*,*Y*)可以表示为：

 （2.3）

这里*p*(*x*)和*p*(*y*)分别为随机变量*X*和*Y*的边缘概率分布，*p*(*x*,*y*)是随机变量*X*和*Y*的联合概率分布。如果*X*和*Y*是相互独立的，则它们的联合熵是它们各自熵的和。

2.2.4 互信息

互信息度量了通过观测另一个随机变量而获得的给定随机变量的信息量，定义了两个随机变量的相互关联程度的大小，它是用来计算两个变量间含有的相同信息量的多少。

假定*X*和*Y*均为随机变量，*X*和*Y*之间的互信息*I*(*X*;*Y*)定义为：

 （2.4）

这里的*p*(*x*)和*p*(*y*)分别为随机变量*X*和*Y*的边缘概率分布，*p*(*x*,*y*)是随机变量*X*和*Y*的联合概率分布。

互信息代表*X*和*Y*之间的统计依赖关系。

2.2.5 条件互信息

条件互信息是用来计算两个随机变量在已知某个的基础上，相互之间还存在的相互关联程度的强弱。

假定随机变量*Z*是已经给定的，随机变量*X*和*Y*之间的条件互信息表示为：

 （2.5）

其中，*p*(*x,y,z*)是联合概率分布，*p*(*x|z*)、*p*(*y|z*)和*p*(*x,y|z*)对应为条件概率分布。

2.2.6 相关性

给定三个随机变量*X*，*Y*和*Z*以及它们的联合概率分布。如果，那么在给定*Z*的情况下，*X*和*Y*的相关性表示为，为0。而当时，在给定*Z*的情况下，*X*和*Y*的相关性定义为：

 （2.6）

根据定义，相关性是一个大于等于0小于等于1的函数，度量了在*Z*值给定的条件下，由*X*提供的*Y*不确定性的相对减小量。

2.2.7 互补性

两个随机变量*Xi*和*Xj*关于输出*Y*的互补性定义为：

 （2.7）其中。

如果两个变量关于*Y*的互补性是一个正值，那么我们将它们定义为是互补的，反之，如果它们的互补性是一个负值，那么我们称它们在预测*Y*时是冗余的。

2.3 属性约简原理框图

特征选择的一般过程为，首先从特征全机种产生出一个特征子集，然后用评价函数对该特征子集进行评价，评价的结果与停止准测进行比较，若满足停止准则就停止，否则就继续产生下一组特征子集，继续进行特征选择。选出来的特征一般还要验证其有效性。

特征选择通常包括特征子集产生过程，评价函数，停止准则和验证过程四部分。

过滤式方法通过分析特征子集内部的特点来衡量其好坏。过滤式方法一般用作预处理，与分类器的选择无关。过滤式方法的原理如下图所示。

训练数据

完整的特征集合

过滤式功能子系统

搜索方法

目标函数

特征子集

信息量

机器学习算法

最终特征子集

图2.1 过滤式方法原理

首先将完整的特征集合输入到过滤式功能子系统中，然后按照一定的搜索方法产生特征子集，然后将特征子集送入目标函数中进行评估，然后将评价结果进行反馈，判断是否继续进行搜索，最终得到特征子集。将最终的特征子集输入机器学习算法，来评价特征选择算法的优劣。

2.4 常用算法

2.4.1 mRMR算法

mRMR算法是基于最小冗余最大相关性准则的算法。代表已被选择的特征子集，代表未被选择的特征子集，代表第*i*个特征，代表第*j*个特征，*Y*代表类别。在未选定特征的中选择变量，使最大化，其中是关联项，是冗余项。换句话说，是对输出*Y*的单独关联，是对每一个已选定的()的平均冗余。

 （2.8）

 （2.9）

 （2.10）

在每一步里，这个方式选的变量都是含有最佳折中相关冗余的。这个选择标准是快速和有效的。在向前搜索第*d*步中，计算*n-d*评价，这里，每一个评价都要估量(*d+1*)个双变量密度估计。

2.4.2 CMIM算法

CMIM算法是基于条件互信息最大化准则的算法。这种方法提出，表示已被选择的特征子集，表示未被选择的特征子集，表示第*i*个特征，表示第*j*个特征，*Y*表示类别。选择出()，在特征()已选定的条件下与*Y*相互关联性最小。并且条件互信息最小值取最大值的那个特征。然后这个具有最大最小条件相关性的特征就被选择了出来。用公式表达如下：

 （2.11）

2.5 小结

本章主要介绍属性约简算法及其信息论基础，并且给出了属性约简的原理框图，并介绍了一些常用算法。

其中，在属性约简算法概述中主要介绍了属性约简的优点及其分类：过滤式、封装式和嵌入式。在属性约简算法中的信息论基础中介绍了一些信息论度量，例如信息熵、条件熵、互信息等，也介绍了在DISR算法中使用的相关性和互补性的概念。然后给出了过滤式属性约简算法的原理框图，简单介绍了算法的流程。最后给出了两种常用的属性约简算法，一种是基于最小冗余最大相关性的mRMR算法，另一种是基于条件互信息最大化的CMIM算法。

3 基于互补性的属性约简算法

3.1 DISR算法原理

3.1.1 互信息的下界

在这一节中，我们获得了数据集*Xs*和目标变量*Y*之间的互信息的下界。这表明这个互信息比其它用*Xs*的子子集所计算出来的互信息还要低。

定理：给定是*X*中含有*d*个变量的子集，是一个不含有变量*Xi*的*Xs*的子集，那么

 （3.1）

上述定理表明了一个子集*S*和一个目标变量*Y*互信息的下界被限制在，其是*Xs*所有子子集*Xs-i*的互信息量的平均值。

接下来，我们将用这条定理来作为下述启发式搜索的一个理论支撑：没有任何关于*d*变量的子集应该如何结合的附加知识，最佳子集是有最好效果的（*d-*1）变量子集的组合。

3.1.2 准则

给定一个固定数量*d*的变量，我们可以将特征选择问题写成如下的形式：

 （3.2）换句话说，特征选择的目标就是找到一个*d*变量的子集，使得与输出*Y*的互信息最大化。

如果将的最大值替换成也就是下界的最大值：

 （3.3）而右边的式子又可以由它的下限值来代替并且递归到只有两个变量的子集：

 （3.4）

也就是说，在没有任何关于怎么结合超过两个变量的子集信息的条件下，最佳的子集（最佳边界）是所有两变量结合的互信息的最大总和。

3.1.3 对称相关性

给定两个随机变量*X*，*Y*以及它们的联合概率分布*p*(*x*,*y*)，则对称相关性*SR*(*X*,*Y*)定义为：

 （3.5）这个定义说明对称相关性是一个的函数，表明了互信息集中包含在*p*(*x*,*y*)里。

因此，DISR的准则如下：

 （3.6）

使用这个准则选择变量的主要优点是一个已选变量的互补变量比其它准则更容易被选出。在前向搜索的第*n*步，搜索算法计算*n*-*d*个评估值，其中每个评估值需要计算*d*重三元密度。

3.2 DISR算法的实现及其流程图

3.2.1 DISR算法的实现

第一个特征的选取准则是按照特征与类之间的互信息最大来进行的。接下来对未被选择的特征进行处理。此部分代码实现如下：

|  |
| --- |
| for n = 1 : totalFeatures  classMI(n) = mi(featureMatrix(:,n),classColumn);  if classMI(n) > highestMI  highestMI = classMI(n);  highestMICounter = n;  end  end  answerFeatures(1) = highestMICounter;  unselectedFeatures(highestMICounter) = 0; |

我们要做的就是生成一个对称相关性的矩阵，根据式（3.5）和式（3.6），我们首先将待选特征与已选特征形成一个联合特征矩阵，然后计算联合特征与类之间的互信息与熵，然后根据式（3.5）用联合特征与类之间的互信息除以熵得到待选特征的对称相关性。此部分代码实现如下：

|  |
| --- |
| unionedFeatures = joint([featureMatrix(:,answerFeatures(m)),featureMatrix(:,j)]); tempUnionMI = mi(unionedFeatures,classColumn);  tempTripEntropy = h([unionedFeatures,classColumn]);  featureSRMatrix(m,j) = tempUnionMI/tempTripEntropy; |

其中，joint(input1, input2)是用来联合input1和input2特征矩阵的函数，mi(input1, input2)是计算两个输入input1和input2之间互信息的函数，h(input1, input2)是计算input1和input2之间熵的函数。

然后根据计算出的对称相关性对待选特征进行打分，选出对称相关性最大的特征。经过循环，对数据集中的特征进行排序。

3.2.2 DISR流程图

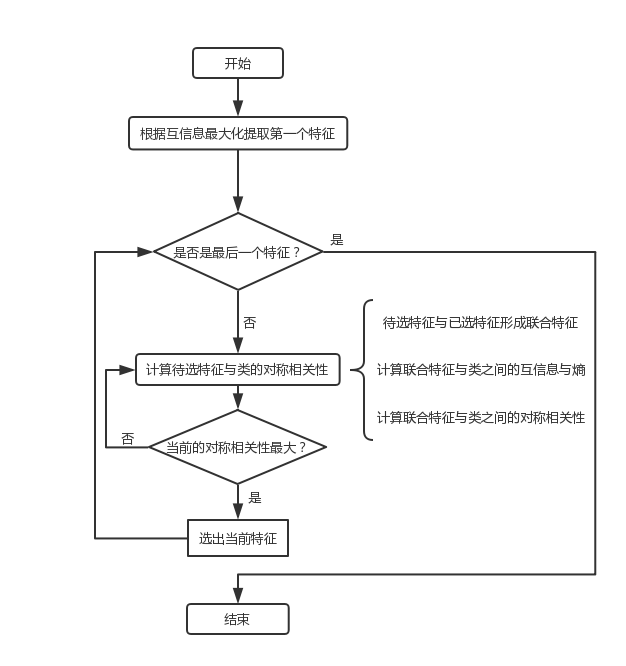


图3.1 DISR算法流程图

3.3 DISR的特征排序

3.3.1 数据集

本文在实验过程中使用了4种数据集，分别是iris数据集、semeion数据集、arcene数据集和gisette数据集。

Iris数据集是一个常用的分类实验数据集，由Fisher收集整理。Iris也叫鸢尾花卉数据集，是一种多重变量分析的数据集。该数据集包含150个样本，分为3类，分别是山鸢尾、杂色鸢尾和维吉尼亚鸢尾，其中第一个种类和后两个种类是线性可分离的，后两者是非线性可分离的。其中每个样本包含4个属性，分别是花萼长度、花萼宽度、花瓣长度和花瓣宽度，根据这4个属性可以预测属于哪种花。表3.1给了该数据集得而部分分数据。

表3.1 iris数据集的部分数据

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 花萼长度/cm | 花萼宽度/cm | 花瓣长度/cm | 花瓣宽度/cm | 类别 |
| 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 4.9 | 3 | 1.4 | 0.2 | Iris-setosa |
| 7 | 3.2 | 4.7 | 1.4 | Iris-versicolor |
| 6.4 | 3.2 | 4.5 | 1.5 | Iris-versicolor |
| 6.3 | 3.3 | 6 | 2.5 | Iris-virginica |
| 5.8 | 2.7 | 5.1 | 1.9 | Iris-virginica |

Semeion数据集收集了大约80个人的1593个手写数字。手写数字被放置到一个16\*16的正方形格子中。并且每个图像的每个像素都用固定的阈值限定在一个布朗型值（1/0）。因此，semeion数据集共含有1593个样本，分为10类，分别是0,1,2，…，8,9的手写体数字的16\*16点阵图。



图3.2 semeion数据集中的手写数字

Arcene数据集是用来从质谱数据中区分癌症细胞和正常细胞的一个数据集，是一个具有连续输入变量的二分类问题。该数据集包含900个样本，每个样本包含10000个特征，特征中添加了一些被称为“探针”的干扰特征，没有预测能力，每组中都有3000个探针。该数据集分为3组数据，分别为训练集、验证集和测试集。在本文的实验中我们仅用到前两组数据，每组数据包含100个样本，其中44个正样本和56个负样本。

Gisette数据集也是一个识别手写数字的数据集，但它区分的是具有高度混淆性的数字“4”和“9”。该数据集包含13500个样本，每个样本中包含5000个特征，该数据集中仍然加入了没有预测能力的探针，每个样本中都包含2500个探针。这些数字已经被尺寸标准化并且位于28\*28固定大小图像的中央。原始数据已经被修改来适应属性约简。该数据集分为三组，分别是训练集、验证集和测试集。在本文的实验中我们仅用到前两组数据，其中训练集包含6000个样本，正负样本各3000个，验证集包含1000个样本，正负样本各500个。

3.3.2 排序结果

在实验过程中，我们对iris数据集进行了双分类和多分类的实验，其中双分类用的是山鸢尾和杂色鸢尾两类，排序结果如下：

同样的，我们也对semeion数据集进行了双分类和多分类的实验，其中双分类用的是数字“0”和数字“1”两类进行的，排序结果如下：

在本实验中，我们对arcene数据集和gisette数据集进行了双分类的实验。

3.4 本章小结

通过上述实验的排序结果我们可以看出，DISR算法对不同数据集的不同特征的排序是不同的。也就是说特征对分类的结果有着显著的影响。