摘 要

随着信息技术的快速发展，人们可获取数据的规模越来越大，数据维度也越来越高。数据的高维度特征会导致“维数灾难”，同时会降低下游任务的性能。此外，获取大量数据的标签在实际中往往需要大量的成本。因此，如何对高维无标签数据进行特征选择，降低数据维度以提高后续任务的性能已经成为大数据挖掘的一个重要问题。

为了解决上述问题，本文提出了基于自适应超图的无监督特征选择方法。首先，根据不同的核函数方法，本文构建了多个超图来刻画特征间不同的高阶相关关系。其次，考虑到不同超图之间存在潜在的一致性，本文使用多个超图学习一个共同超图。此外，为了保留数据的全局信息，提高特征的判别力，本文对特征权重矩阵进行了正交约束。最后，本文将特征学习和自适应超图学习统一到一个框架中，相互促进，从而有利于选出最优特征子集。

为了验证本文所提方法的有效性，将该方法在 7 个公开数据集上和相关的特征选择方法进行比较。实验结果表明，本文提出的方法在两个评价指标上具有较优的性能。此外，本文也在一个实际信用数据集上验证了本文方法的有效性。

关键词：无监督特征选择；自适应超图学习；一致性信息

目 录

1 绪论

1.1 研究背景

近些年，随着网络信息技术和多媒体的发展，各行各业的数据都呈现指数级增长的状态，因此产生了海量的高维数据。各种各样高维数据集的出现可以使我们更加清晰地了解事物的表象，但高维数据中往往存在大量的冗余数据和无关数据，为数据集增添了很多没有意义的特征，这会降低机器学习和信贷风险评估算法的效果，若想从高维数据中获取有用的信息，则需要通过降维方法从高维数据中获取有用的特征[1]。因此特征选择方法成为了很多统计学和机器学习领域的研究者研究的焦点问题之一。从高维数据中挖掘有意义的特征，称为数据挖掘，它可以从原始的数据集中提取出数据内部隐含的，并对经济社会的生产和人们生活有价值的信息，并且借助计算机技术寻找数据中蕴含的规律。数据挖掘在模式识别，图像处理和信贷风险等领域有着广阔的应用前景，在数据分析中十分流行。

高维数据中含有十分详细的信息，这些信息体现在特征的高维度上，但特征维度高意味着可能存在冗余和没有相关性的特征，相关研究人员不能直接使用这些数据，必须对数据进行筛选和预处理，这是研究工作者需要面临的难题[2]。通过降维技术约简数据已成为数据预处理中经常遇到的一项任务，数据维数降低会简化机器学习算法的结构，最终可以提高模型的效果[3]。

目前有两种减少数据维度的方法，特征提取方法和特征选择方法。虽然这两种方法最终都能降低数据的维度，但是使用了不同的数据处理技术。特征提取实质上是对高维数据进行低维的映射或变换，用少量的特征表示原始数据中的信息，通过特征提取技术降低数据的维数，最终的特征和原始数据集中特征是不相等的，这些特征是新产生的特征[4]，因此可解释性较差。和特

征提取技术不同的是，特征选择方法能够从原始数据中选择出具有代表性和辨别性的特征子集，没有重新生成特征，保留了原始特征空间的结构[5]。所以，特征选择方法经常会在高维数据预处理阶段进行应用。特征选择通过正交约束或自适应谱图学习等方法能够去除原始数据中不相关和冗余的特征。降低数据的维数，能够降低数据的存储空间，简化机器学习算法的结构，有利于提升模型性能[6]，并且有利于增加实验结果的可解释性。

由于不同的数据集可能面临不同的处理任务，并且有些数据集是有标签的，而有些数据集是没有标签的。根据数据集有无标签，或者数据处理过程中是否需要标签，特征选择可分为两种，分别为有监督的特征选择方法[7]和无监督的特征选择方法。有类别标签的数据集可以使用有监督的特征选择方法，该方法利用了特征和类别标签之间的关联性进行特征选择。缺失类别标签的数据集则使用无监督特征选择方法，该特征选择方法利用数据内部结构选择特征，可以提高聚类的准确性。

随着经济和信息技术的迅速发展，真实世界中存在非常多的无标签数据，对于缺失数据较少的样本数据可以通过人工为数据贴上标签，进而将其转化为有监督数据。但现在大部分数据集的样本量非常大，并且人力成本昂贵，为数据贴标签没有现实意义。因此，针对没有标签的数据集，研究无监督特征选择方法是具有价值和意义的，并且有着广泛的应用场景，提升无监督特征选择方法的效果和稳健性，仍是未来值得机器学习领域的研究者研究的重要课题。

无监督特征选择本质上是利用数据内部的几何结构选择具有代表性的最优特征子集。以往很多无监督特征选择的研究者都基于谱图理论展开研究，主要通过原始数据学习到相似矩阵，保持数据的局部几何结构，然后对各变量进行模型优化，选出辨别性较强的特征。相似矩阵中元素即为数据样本点，每一列既可以理解为一条边，每列中出现的非零元素即表示边连接的样本点，根据相似性的不同为样本分配不同的权重值。当构建相似图时，有两种边连接的方式，分别为全连接图和不全连接的稀疏图。以往的研究大都采用不全连接图的方式构建图，比如利用 K 近邻（KNN）的思想，每条边连接离样本顶点最近的 K 个样本顶点，只计算这些样本之间的权重。不全连接图的构图方式，相比全连接图减少了边连接的顶点数量，得到一个稀疏的相似矩阵，

可以加速算法执行效率[8]。

1.2 研究意义

目前提出的无监督的特征选择方法具有一些问题，首先，很多研究采用简单图进行研究，而简单图只能保持数据的成对关系，而很多数据的特征空间中存在着多重关系，比如在研究论文和作者的关系，一篇论文通常是由多位作者编写而成，也就是论文和作者的关系往往是多重的。此外，很多研究者在实验时是构造了一种图，忽略了不同图之间具有的互补性和一致性信息。本课题基于以上两方面的问题，提出了一种基于自适应超图的无监督特征选择的算法模型。一方面，超图学习可以保持数据的高阶关系，另一方面，基于直方图交叉核、高斯核等不同的核函数方法，并结合 KNN 方法构建了多个超图，可以保留不同的超图之间一致性和互补性的信息。同时，本文提出了一种迭代优化算法，并将特征选择和自适应超图学习统一到一个框架中，相互促进，有利于输出最优的特征，进而提升实验效果。

总的来说，自适应超图学习的无监督特征选择研究方法能够选出具有代表性和辨别性的特征，能够提升聚类效果，加快机器学习算法的执行效率。该方法可以应用在很多机器学习和数据挖掘任务中的预处理，比如图像识别问题和信贷风险评估问题，使用该方法选出的特征，可以简化机器学习算法的结构，增加实验结果的准确性，以便从更多的高维数据中挖掘有用的信息，促进科技和经济社会健康有效地发展。

1.3 本文研究内容和组织结构

本文提出了一种新的嵌入式无监督特征选择方法—基于自适应超图的无监督特征选择方法（AHUFS）:1）从原始数据的特征空间构造相似度矩阵。利用超图保持其高阶局部几何结构。2）对特征矩阵使用 $l_{2,1}$ 范数来增加特征的稀疏性，有助于选出具有代表性的特征。3）利用多个超图自适应学习一个共同的超图，增强模型的鲁棒性，和特征选择过程相互促进，提升实验效果。其中超图的构建是基于不同的核函数，这些核函数包括直方图交叉核、χ^2核、

高斯核、欧几里得距离核和余弦核核函数。在大多数情况下，这些不同的超图之间包含一致性和互补性的信息。通过自适应学习共同的超图可以有利于输出可靠和信息丰富的特征，选出的特征又可以促进学习自适应超图的效果。

和目前存在的无监督特征选择方法相比，本文提出的 AHUFS 方法具有以下贡献：

(1) 本文提出了无监督特征选择基于自适应超图学习，以往很多关于无监督特征选择的研究都是基于简单图，简单图只能保留数据集中样本的成对关系，这样会忽略原始数据样本之间的多重关系，而超图可以保留这种多重关系，能够更加充分的保持原始数据中复杂的局部几何结构。

(2) 本文构建了多个超图自适应学习一个共同的超图，充分利用了不同超图之间的互补性和一致性的信息，增强了模型算法的鲁棒性。其次，将自适应学习超图和特征选择统一在一个框架当中，相互促进，有利于选出最优的特征权重矩阵和超边权重矩阵。

(3) 本文提出的模型开发了一种有效迭代优化算法，收敛性由实验证明。并且通过在真实数据集上大量实验表明，本文所提方法都优于目前先进的基线模型。

本文共有五章，对文章的组织结构做了如下安排：

第一章，绪论，介绍了本文的研究背景和研究意义，引出了本文研究目的。其次，介绍了本文的研究内容和组织结构，说明了本篇论文的研究逻辑和框架。

第二章，对特征选择的国内外研究现状做了一个概述。为引出我们的研究，分别描述了基于简单图和超图的无监督特征选择，同时，对特征选择在经济学中的应用作了简要介绍，为本研究方法的展开打好铺垫。

第三章，该章节主要研究了我们提出的算法模型，引言对本章的主要内容做了概述，之后提出和分析了算法模型，并对其进行了优化，最后说明了实验结果的有效性和实验的收敛性。

第四章，本章主要基于第三章提出算法的有效性，进一步结合自身的学科背景，在信用数据集上进行了信贷风险领域的研究，通过实验结果的比较，说明了我们提出算法的有效性。

第五章，对本文实验过程进行了总结，并阐述了本次研究中的不足之处，

对今后的研究进行展望。

1.4 本章小结

本章首先介绍了本文的研究背景，从现实角度引出了无监督学习的必要性，随后，介绍了本文的研究意义，阐明了本文研究方法的可取之处，接着，对本文的研究内容做了简要的介绍，引出了本论文提出的方法。最后对文章的组织结构进行介绍，使文章的逻辑结构更有条理性。

2 国内外研究综述

2.1 特征选择概述

得益于信息技术的高速发展，我们进入了数字化时代，生产生活中产生的数据和信息都通过数据化的形式保存下来，由此各行各业产生了很多大规模的高维数据。数据维数呈指数级增长，但同时需要面临的问题则是很多机器学习算法在处理高维数据时，需要耗费大量的存储和计算成本，算法执行效率低下。一个优秀的特征选择算法可以很好的解决这些问题，显著地降低数据的维数，将具有代表性和辨别性的特征选出，简化了机器学习算法的结构，提升了算法运行效率和实验结果的准确性。因此越来越多的研究者关注并研究特征选择方法，因此关于特征选择的方法在近些年得到了很大的发展。

数据维度的增加使得针对高维数据的处理，变得越来越迫切，这要求机器学习算法的性能能够有较大的提升，现实需求极大的促进了大数据技术的发展。特征选择则是提升机器学习算法性能的重要途径之一，即在机器学习实验的数据预处理阶段进行特征选择，在此基础上进行试验可以大大节省数据的存储成本，为高效开展后续实验提供了保障。

在数据量和特征量较少的时候，完成学习任务并不需要特征选择，只需通过分析确定学习任务的影响因素，然后将能够代表这些因素的特征变量找出并量化，然后将这些特征输入到合适的机器学习或统计模型中便可以成功解决我们遇到的问题。但是现实生活中遇到的数据量和维数都非常大，如果通过简单的处理和变换就将其输入到机器学习模型当中，将浪费大量的时间成本，即使得出实验结果，也难以保证它的质量。特征选择方法变得更为重要，因为特征选择方法旨在消除特征之间的冗余度，降低数据的维度，去除掉和研究目的无关和影响较小的特征，只保留包含有用信息的少量最优特征子集。

深度学习算法过去由于计算机性能对数据的限制导致没有得到很好的使用。随着计算机算力的提高，深度学习技术再次兴起，只需将原始数据集输

入到神经网络模型当中，最终模型会自动输出最优的特征子集。但是这并不是说特征选择的机器学习方法不再需要，深度学习是一个黑箱，对实际问题没有进行分析，对于实验结果可解释性较差，相反经过特征选择算法选出的最优特征具有很好的可解释性，并且可以理解算法原理，使特征选择结果更加简单和直观，增加实验的可靠性。

2.1.1 特征选择的定义

在以往的统计课程的学习中，对特征选择有另外的叫法，叫做属性选择或者变量选择，这个叫法更为统计学专业的学生熟知。它指的是将原始数据输入到统计模型当中，输出最优特征子集的过程。根据被解释变量和解释变量的关系，以及解释变量自身具有的相关性，特征选择可以分为 4 种类型。分别为相关特征、无关特征、冗余特征和非冗余特征。

相关特征可以理解为代表性特征，和类别标签具有内在联系的特征称为相关性特征。无关特征就是指原始数据集中与类别标签没有联系的特征。冗余特征指的是具有相似信息的一些特征，这些特征在学习任务中是不需要的。相关和无关特征属于有监督学习研究的问题，而冗余和非冗余特征属于无监督学习研究的问题。

相关特征指的是原始数据集的代表性特征。无关特征指的是对学习任务没有意义的特征。冗余特征是指重复信息，指特征空间当中包含多余的特征。有监督学习中特征分为两种，相关特征和无关特征，无监督学习中特征分为，冗余特征和非冗余特征。

如果原始矩阵我们用 X 表示，$X \in R^{m \times n}$，m 为样本数量，n 为特征数量，f_i表示 X 的第 i 个特征，X 为原始的特征空间：$X = \{f_1, f_2, \cdots, f_n\}$。特征选择的目的是为了从原始数据矩阵 X 中选出具有代表性和辨别性的最优特征子集，从而组合形成维数更低的特征集合 S，进一步提升机器学习算法的性能，也可以增加机器学习算法的聚类和分类准确性。

对于有监督学习，特征选择是为了识别并删除无关特征。对于无监督学习，特征选择的目的是识别并删除冗余的特征子集。特征选择可以降低数据的维度，简化了算法结构，大大降低算法的计算复杂度，可以提高算法的学

习速度，提升实验结果的可解释性，能够提高数据的预测准确率。

如果数据的样本数量远远比特征的数量小时，特征选择的搜索空间是稀疏的，这个时候，模型对于相关特征的区分能力变得非常差，使特征选择算法的性能下降。

2.1.2 特征选择的分类

有多种方式可以对特征选择方法进行分类，根据数据样本的类别标签判断信息是否可用，特征选择方法可分为有监督，无监督和半监督三种。按照学习算法的不同可分为过滤式、包裹式和嵌入式的特征选择方法。下面将对这些方法进行具体介绍。

(1) 按数据标签是否需要的原则划分

根据特征选择算法是否需要类别标签，可以将特征选择方法分为有监督的特征选择方法，无监督的特征选择方法和半监督的特征选择方法。

有监督特征选择中的特征选择方法主要用来处理分类或回归问题，根据特征和标签的关联性，选出和标签的相关性强的那些特征。在进行特征选择实验之前，需将数据按照一定的比例拆成训练集和测试集，在训练集进行模型的学习，特征选择的过程既可以与分类或回归模型关联，也可以不和这两类模型关联，独立的进行特征选择，或者可以在机器学习模型中嵌入特征选择算法，在学习模型的同时也在进行特征选择。利用选出的特征子集在测试集上进行预测，最后将预测的标签和真实的标签进行比较，验证模型算法的有效性。

通常在聚类问题中使用无监督的特征选择方法，特征选择的准则为选择使聚类准确性得到显著提高的特征子集。很多数据集是没有标签的，而人工标注数据的成本甚至大于收益。开发一种新的具有优良性能无监督特征选择算法则可以解决这一难题。由于没有标签，无监督特征选择无法根据特征和标签的关联而选择特征。无监督特征选择实质上是通过数据集内部的全局或局部结构来进行特征选择的一种方法，在实验过程中需要从原始数据中进行学习。特征选择可以依赖算法，也可以独立于学习算法，或将其嵌入算法的学习当中。有时也会遇到部分样本含有标签的数据集，此时便需要采用半监

督的特征选择算法。

(2) 基于特征选择和学习算法的关联性划分

根据特征选择和学习算法的关联性关系，可以把特征选择分为过滤式、包裹式和嵌入式方法。如图 1 所示，过滤式特征选择方法一般包括两个步骤：第一步，对特征或特征子集进行排序，将排好序的特征利用单一的评价指标进行特征评估，选择评估效果最好的特征或特征子集。对单个特征进行排序不需要考虑其他的特征，对特征子集进行排序，则需要将特征分为多个批次的特征子集。

过滤方法和学习任务是相互独立的，该方法采用特征排序的方法对特征或特征子集的重要程度进行评估，常用到的排序方法包括方差[9]，拉普拉斯分数[10]，特征相似度[11]和踪迹比[12]。He 等人采用了拉普拉斯分数的方式对特征进行排序，主要思想为类别相同的两个数据样本，他们之间离得应该很近。刘等人提出了一种基于信息论的方法，通过分层凝聚的方法对特征聚类，以此进行特征选择[13]。在文献[14]提出了一种特征选择的相似性保持准则，该准则超越了许多广泛使用的准则。Wang 等人采用了最大投影最小冗余度的特征选择方法选择特征[15]。Roffo 等人主要考虑了特征的分布，通过研究很多个特征分布的问题选出有效的特征[16]。

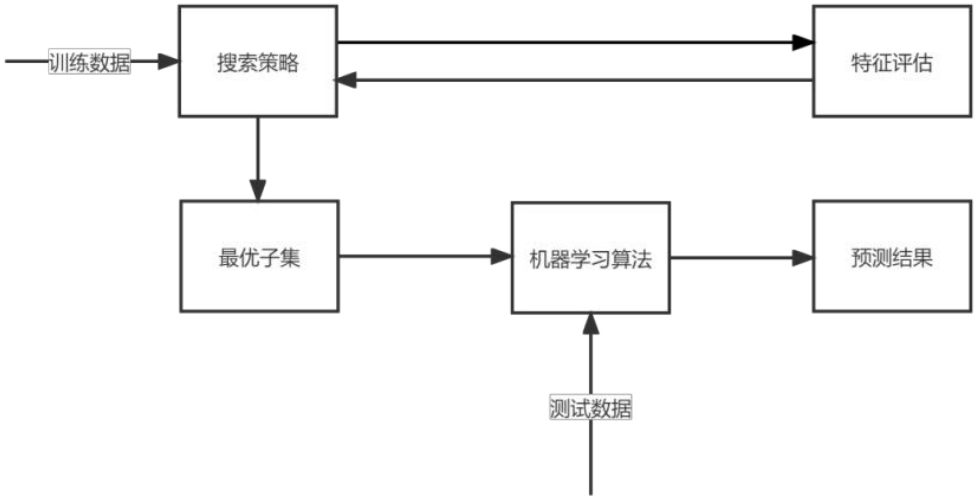

图 1 过滤式特征选择算法流程

过滤式特征选择方法具有较高的计算效率。但其学习算法是不固定的，

导致根据该方法选出的特征不是最优特征。其次，过滤式方法对特征进行单独操作，忽略了特征之间的相关性特点，因此不能有效地去除冗余特征。

包裹式方法选择特征是基于模型算法的聚类性能[17]，该方法将特征选择作为学习任务的一个中间环节，其目的是帮助提高学习任务的性能。具体来说，包裹式方法选择特征是为了更好地服务于给定的学习任务，提高学习性能。如图 2 所示包裹式特征选择方法的实验步骤是，首先，对特征子集进行搜寻；然后，评价特征子集，重复执行上诉两步操作，直到迭代终止条件满足则终止执行。终止条件一般为准确率的一个阈值，或者设定合适的特征数量。最终选择的特征子集，使得学习算法的准确率达到最高。

Dy 等人采用期望最大化（EM）聚类算法，利用离散可分性和最大似然估计的原理进行特征选择，选出最优的特征子集[18]。在验证子集中利用误差去除冗余特征[19]。一般情况下，包裹式特征选择模型优于过滤器模型[15]。但包裹式方法存在一个缺点，搜索所有的次数会随着特征数量的增加而指数级增加，因此不适用于高维数据。也有研究人员尝试解决这个问题，比如提出了顺序搜索或随机搜索等新的搜索策略，这可以一定程度上缓解包裹式特征选择不能用于高维数据特征选择的问题。但在面对维数过高的数据时，大多数包裹式方法在优化问题的计算上仍然是困难的，包裹式方法仍具有局限性。

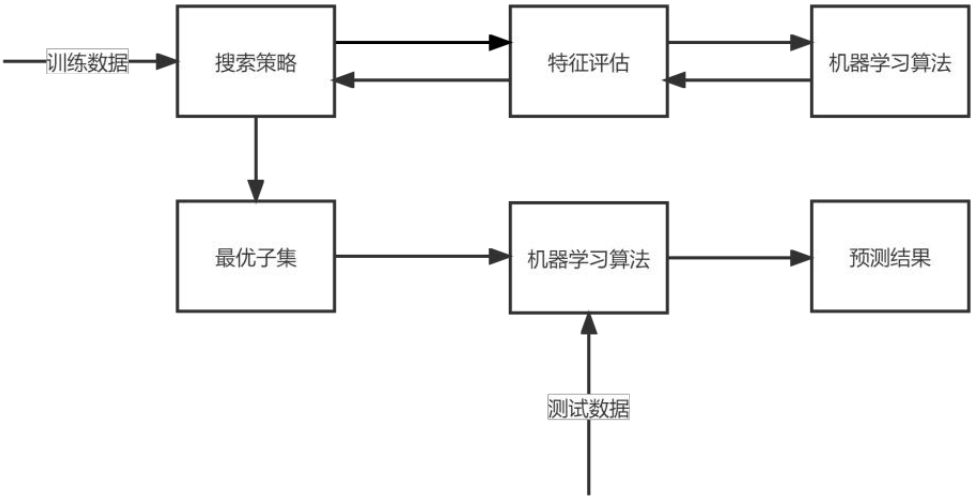

图 2 包裹式特征选择算法流程

嵌入式方法将特征选择和学习器训练过程结合在一起，二者在同一个优

化过程中完成，学习器在训练时自动的进行特征选择。具体而言，如图 3 所示，嵌入方法的学习模型通过训练所有特征而成，然后在保持学习模型的前提下去除部分冗余特征，从而达到模型的最佳精度，并且已被证明优于过滤和包装方法。嵌入式方法继承了包裹式方法和过滤式方法的优点，一是特征选择和机器学习统一到一个框架中，有利于选出更优的特征；其次，嵌入式方法比包裹式方法更高效，因为不需要迭代评估特征子集。嵌入式方法的步骤如下：

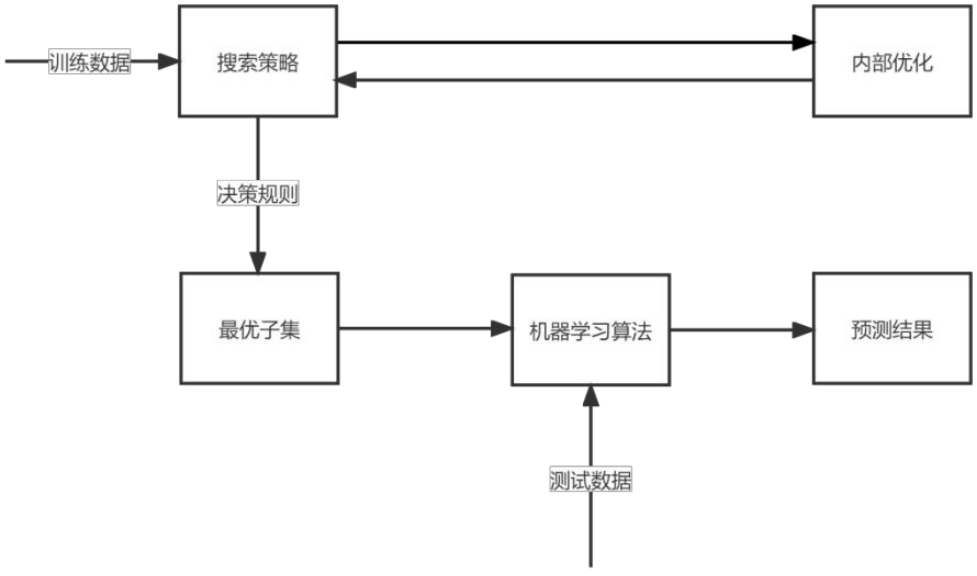

图 3 嵌入式特征选择算法流程

许多嵌入式方法首先采用简单的图来构造度量数据样本之间的成对关系，以此来构建相似性矩阵，保持数据样本，然后将所得到的图进行正则化和稀疏限制来选择特征。例如，基于支持向量机（SVM）的递归特征消除方法[20]，正则化最小二乘分类器（RLSC），K-Means 的单集频谱稀疏化（BSS）[21]，以及基于 K-Means 聚类算法的特征随机选择（RFS）算法[22]。在近些年的很多研究中，出现了很多基于稀疏学习的特征选择方法，这些方法在目标函数中使用了稀疏的正则化项，有利于选出性能优良和具有较强解释性的特征。这些方法背后都围绕着一个基本原理，即稀疏性正则化可以对特征的重要性进行约束。为了在特征选择过程中使特征权重矩阵更加稀疏，现如今已经开发了许多稀疏性诱导的规范，例如稀疏逻辑回归[23]和组稀疏性[24]。比较典型

的方法有：采用多个聚类方法进行特征选择（MCFS）[25]，将特征选择和子空间学习统一到一个框架中进行学习（FSSL） [26]，利用无监督鉴别的方式进行特征选择（UDFS）[27],以及基于特征相似性学习的无监督特征选择算法（FSFS）[28]。文献[6]在特征选择过程中采用了自我表征的方法，并显示出了优越的结果。自我表征方法的可行性是因为数据集中的每个特征都可以通过与它相关的其他特征的组合进行重构，并且具有稀疏性约束地稀疏矩阵可以用作特征权重。很多研究已经证实，保持数据的局部几何结构对于无监督的特征选择特别重要[24]。为此，在许多的嵌入式方法中，图拉普拉斯正则化项已经被广泛用于保持局部几何结构[25]。然而，以往方法中使用的传统相似度图，即后文提到的简单图只能表示数据的成对关系，而不能描述高阶关系，因此原始数据中隐含的复杂结构不能被充分利用。此外，以往的很多方法都是在原始数据的特征空间中进行特征选择，这样会使最终的算法模型性能受到原始数据中噪声的影响。

当前的嵌入式方法仍有局限性，需要在后续的研究中加以解决。一方面，嵌入式方法的两步策略（即学习相似矩阵和进行特征选择）可能会降低特征选择的性能，因为相似矩阵学习的目的是获得最优的相似关系，而不是特征选择的结果。另一方面，现有的嵌入式方法是从原始数据构造相似度矩阵，这些数据通常包含冗余和不相关的特征，因此可能会选择没有信息的特征。除此之外，通过一个简单的图来构造相似度矩阵，该图可以衡量训练数据之间的成对关系，而没有考虑他们之间的高阶关系，从而不足以捕捉训练数据中的复杂结构。

需要另外说明的是，在有的研究中将特征选择方法总共分为 4 种，除了前面提到的过滤，包装和嵌入方法，还加入了一种混合式的特征选择方法。顾名思义，混合式特征选择方法其实就是把多个特征选择的方法进行组合。因为之前的每个特征选择算法都有其优点和缺点，所以混合式特征选择方法想要利用这些方法的优点，并且克服不同特征选择模型的缺点。该方法适用于小样本的高维数据，当小样本的高维数据受到微小扰动时，如有使用前面提到的方法，产生的特征子集会发生较大的改变，不能够选出最优的特征子集。但是使用混合式的特征选择方法，将特征选择结果组合在一起，则可以增加选出特征的可靠性和可信度。

(3) 依据不同数据类型划分

大数据具有数据价值密度比较低，需要的计算机处理速度要快，数据类型较多，时效性要求较高的特点。这些对特征选择方法的研究带来了新的挑战。如今社交平台，网络电视平台产生了各种的流数据，其数据的固定结构和数据的特征是变化的，传统特征学习不适用于流数据。而且，有时数据流的规模很大，是不能够一次性地将其储存在计算机的内存当中，因此只能对数据流划分不同的批次，按照批次进行加载和处理。而这需要开发针对流数据的特征选择模型。

另外，传统的统计模型方法在进行特征选择时都假设数据样本是同构的，即每个样本之间都是独立同分布的。但是数据的来源有可能是不同的，文本、图像和视频产生的数据的数据分布不是相同的。所以需要采用新的特征选择方法，这种新的特征选择方法可以利用数据的内部结构和数据之间的相关性来选择最优的特征子集。来自不同数据源的特征具有不同的内在结构，可能具有组的结构，也可能具有树和图结构的结构。所以，当面对来自多个渠道的数据时，必须考虑数据内部结构，从而使特征能够在目标学习任务上进行高效的使用。特征的先验知识也可以帮助提高特征选择算法的性能。

总之，按照数据类型不同可以将特征选择划分为两种类型，一种是流数据，另一种是静态数据。这两种数据适用的特征选择方法是不相同的。

2.1.3 特征选择的步骤

一般来说，特征选择分为 4 步。首先，需要生成特征子集；其次，需要对子集进行评价；然后，需要根据终止条件进行模型迭代；最后，需要对实验结果进行验证。下面将对每个步骤的具体作用进行说明。

(1) 子集生成

特征子集的生成需要具备两个条件，一个是搜索起点，另一个是搜索策略。搜索起点和搜索策略可以相互影响，一方面，当新的特征子集生成时，会产生新的搜索起点，这会影响之后的搜索策略。另一方面，当使用的搜索策略不相同时，后续的特征子集的搜索也会受到影响，进而影响搜索起点的确立。搜索策略主要分为完全、随机和启发式的搜索策略。

1) 所谓完全搜索策略指的是通过列举的方式，将所有的特征集合都列举出来，在此基础上选择最优的特征子集。运用数学中的排列组合原理可以准确的计算候选子集的个数。比如数据集 X 中一共包含了 n 个特征，则会产生2^n个特征子集，即搜索复杂度为$O(2^n)$。可以看出来，当数据集中的样本个数增加，搜索的次数将会呈现指数级增长。这大大增加了模型的计算量，在实际应用中会付出大量的成本，因此不适合大规模的数据处理过程。

2) 所谓随机搜索策略就是指当面对的数据是均匀分布的数据时，搜索的轨迹在目标区域内是均匀且随机的。退火算法、遗传算法和进化策略等搜索算法为常见的随机搜索算法。这些随机搜索算法基于概率统计和随机采样的学科知识背景，具体步骤为：首先，为每个特征分配权重；其次，设定筛选特征的阈值；最后根据阈值选出特征，可以提升模型的分类性能。

3) 所谓的启发式的搜索策略是指划分特定的信息节点，这些信息节点有可能是完成目标任务所需要的最好的路径。该搜索策略方法需要评估每一个搜索位置，当评估完成后将继续从该位置进行搜索，直到实现指定的目标则停止搜索。按照这种启发式的搜索策略进行搜索，可以省略很多没用的搜索路径，从而加快搜索速度。上述提到的对位置的评估是非常重要的一个环节，常用到的位置评估准则包括常识或有根据的判断等方法。

(2) 子集评价

子集评价也是特征选择的一个重要步骤，经过评价，如果是包含重要信息的特征，或者具有代表性的特征，则将这些特征进行保留，否则就剔除这些特征。在之前的研究中经常采用的子集评价方法如下：

1) 互信息（Mutual Information）指标可以评估两个特征之间的相互依赖程度。互信息也可以用在有监督和无监督特征选择方法中，在有监督的特征选择场景下，互信息可以衡量特征与标签信息之间的相关性。在无监督特征选择的场景下，互信息对特征变量之间的相关性。互信息的计算公式如下所示：

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (2\text{-}1)$$

其中，$p(x,y)$表示的是联合概率密度，$p(x)$和 $p(y)$表示的分别是 x 和 y 的边缘概率密度函数。

2) 相关系数（Correlation coefficient）指标可以计算出两个特征变量之间的线性相关程度。该系数的绝对值越大，则说明变量之间的相关程度越大，如果该系数的符号为正，则表示变量之间是正相关关系，如果符号为负，则表示变量之间是负相关关系。在有监督特征选择方法中，可以通过计算特征和标签的相关系数，衡量二者的相关程度。在无监督特征选择方法中，则可以计算特征之间的相关系数，从而反映特征之间的相关性。相关系数的计算公式如下：

$$P(a,b)=\frac{cov(a,b)}{\sqrt{Var(a)}\sqrt{Var(b)}} \tag{2-2}$$

其中，$Cov(a,b)$表示 a 和 b 之间的协方差，$Var(\cdot)$表示方差。

3) 分类准确率（Classification rate）是通过对特征使用分类算法进行分类而得到的，得到分类标签和真实样本类别一致的比率。分类准确率可以计量数据样本分类的准确率，该值越大，则说明分类越准确，用于评价有监督学习中的特征子集。分类准确率的计算公式如下：

$$CLF=\frac{1}{m}\sum_{i=1}^{m} g(l_I),r_I \tag{2-3}$$

其中，l_I是第 i 个样本的真实标签，r_i是第 i 个样本的分类标签；函数$g(l_I,r_i)$的取值取决于真实标签和分类标签是否相等，若相等则取值为 1，若不想等则取值为 0。

4) 冗余率（Redundancy rate）指标是用来度量特征变量之间的冗余度，该比率的值越高则说明冗余度越高，反之冗余度则越低。Rr 计算公式如下：

$$Rr=\frac{2}{n(n-1)}\sum_{f_I,f_j\in S,i>j}\left|\beta_{I,j}\right| \tag{2-5}$$

其中，S 表示的是选择的特征子集，n 表示该特征子集包含的总的特征个数，$\beta_{i,\ j}$表示两个特征f_i和f_j之间的皮尔逊相关系数。

以上对常见的特征评价指标进行了介绍，但评价指标除了上面介绍的还有很多，例如 Laplacian 分数和 Fisher 分数等。

(3) 迭代终止条件

迭代终止条件决定了什么时候停止迭代，并且上述论文提到的特征子集的生成策略和评价函数会对终止迭代的时间产生一定的影响。终止迭代的策略总共有 4 种，下面将一一列举：

1) 迭代次数达到指定的次数；

2) 不管是增加还是删除特征，不会选出更好的特征；

3) 选出的特征子集中的特征个数和指定的特征个数相等；

4) 选择通过评价函数评价最好的特征子集。

想要终止迭代的话，一种常见的做法是设定一个阈值，只要特征子集评价的结果达到或者超过阈值则终止迭代，反之继续搜索。

(4) 结果验证

把通过特征选择方法选择的特征子集输入到机器学习模型当中，检验对预测准确率和算法性能提升效果，该方法称为结果验证。交叉验证是常见的验证方法，其原理是将数据集分割成若干部分，将其中一份作为测试集，剩余的特征作为训练集，重复多次这样的操作，每次使用训练集进行模型的训练和学习，将学习到的特征在测试集上进行验证。交叉验证的过程中，上次出现在训练集中的样本，下次就可能出现在测试集上。按照不同的切分方式，将交叉验证分为两种，如下所示：

1) 简单交叉验证：这里说的“简单”，其实是指数据分割的份数最少。首先随机的将样本分为两部分（70%的数据样本充当训练集，30%的数据样本充当测试集），分好后接着利用训练集来训练模型，在测试集上利用模型进行验证。然后，再次打乱样本，重新按原有的比例选择训练集和测试集，和之前步骤一样，在新的训练集上学习模型，在测试集上验证模型。最后选择能使损失函数值最小的模型参数。

2) K 折交叉验证：K 折交叉验证首先会把数据分为相等份额的 K 份，将其中一份作为测试集，其余数据样本作为训练集。当上一次训练和验证模型完成之后，重新选择 $K-1$ 份数据作为训练集，剩下的 1 份数据作为测试集再次训练。经过 K 次后，选择使损失函数值最小的模型参数。

2.2 基于简单图学习的无监督特征选择

在许多机器学习任务中，数据高维一方面意味着有丰富的信息，另一方面也可能有维数灾难问题。事实上在很多现实世界的应用中，只选取具有区别性和富含信息的特征中的一部分子集的聚类效果就会比选取所有特征的效

果好。人类认知的准则对于机器学习模型的设计具有重要的意义。通过模仿人脑机制设计的认知系统可以处理很多计算任务，比如：分析，处理，评估等。数据降维方法可以看作是一种分析数据内部特征的认知方法[29]。

一个好的特征选择方法可以选择到一个特征子集，这个子集不仅可以减少噪音和冗余特征的影响，而且还可以保存数据的内部结构。由于数据标签的缺失，无监督特征选择必须充分利用数据的结构信息。谱聚类方法在保存无标签数据的内部结构是非常有效[30]。谱聚类方法需要一个相似图探索数据的结构信息。传统的谱聚类方法总是在原始的高维数据中利用 K 近邻（KNN）算法构建一个相似图。He 等人提出了 Laplacian Score（LS）度量保存特征局部结构的能力，并寻找那些可以代表图结构的特征[10]。Li 等人采用了回归模型和谱聚类进行非负判别特征选择[27]。它通常基于原始数据构造了相似图，并且采用了非负矩阵分解（NMF）增强了谱聚类学习的能力[31]。Shi 等人提出了一种鲁棒的谱特征选择（RSFS），该方法利用一个鲁棒的谱回归方法提升了图结构的鲁棒性[32]。此外，一些研究更侧重于修改模型和利用范数约束[33]导致的稀疏性来提高特征选择的性能。然而，上述提到的所有方法都有一个共同问题：相似图不可靠。事实上，原始高维数据可能包含了很多的冗余和噪声特征，这些特征会破坏相似图，使得在模型学习的过程中如果使用这样的相似图会不可避免的降低模型的性能。

为了克服固定图的缺点，Nie 等人已经开发了一个模型可以同时进行特征选择和图学习，该模型名字为 SOGFS[30]。该方法利用每次迭代中学习到的图为数据构造图形，以降低冗余和噪声的影响，但是，它过多的考虑了子空间的结构，而忽略了原始流行的贡献，使得图偏离了内在结构。这样的问题也出现在 Li 等人的研究中，其学习到的图形受到最大熵的约束[34]。因为基于原始高维数据的相似度图可以在一定程度上反映内在结构。Nie 等人将误差重构和受约束的拉普拉斯秩（CLR）结合起来，得到了具有期望连通分量的自适应图，并且用于聚类任务。基于此，Peng 等人提出了一种将联合图学习和无监督特征选择算法（JGUFS），该算法可以根据原始数据构造的融合相似图来调整自适应图[35]。然而，自适应图的质量对于超参数是敏感的，更为重要的是，自适应图不能反映学习到的低维子空间的结构。下面介绍几种基于简单图无监督特征选择的目标函数构造方法。

(1) 拉普拉斯评分（LapScore）[10]

LapScore 是一个典型的过滤式无监督特征选择方法，该方法通过拉普拉斯分数和局部结构保存能力来度量特征的重要性。LapScore 基于的假设为：如果两个数据点是彼此接近的，则这两个数据点可能和相同的主题有关联。L_r定义第 r 个特征的拉普拉斯分数，令f_{ri}表示第 r 个特征的第 i 个样本。通过构造最近邻算法来寻找能够反映局部几何结构的特征，利用特征权重矩阵 S 模拟数据空间的局部结构，对于第 r 个特征$\mathrm{f_r} = [\mathrm{f_{r1}}, \mathrm{f_{r2}}, \cdots, \mathrm{f_{rm}}]^{\mathrm{T}}$,拉普拉斯得分计算如下：

$$\mathrm{L_r} = \frac{\tilde{\mathrm{f}}_{\mathrm{r}}^{\mathrm{T}} \mathrm{L} \tilde{\mathrm{f}}_{\mathrm{r}}}{\tilde{\mathrm{f}}_{\mathrm{r}}^{\mathrm{T}} \mathrm{D} \tilde{\mathrm{f}}_{\mathrm{r}}} \tag{2-6}$$

其中 $\mathrm{D} = \mathrm{diag}(\mathrm{S}\mathbf{1}),\ \mathbf{1} = [1,1,\cdots,1]^T$，图拉普拉斯矩阵 $\mathrm{L} = \mathrm{D} - \mathrm{S}, \tilde{\mathrm{f}}_{\mathrm{r}} = \mathrm{f}_{\mathrm{r}} - \frac{\mathrm{f}_{\mathrm{r}}^{\mathrm{T}} \mathrm{D}\mathbf{1}}{\mathbf{1}^{\mathrm{T}} \mathrm{D}\mathbf{1}}\mathbf{1}$。

(2) 鲁棒的近邻嵌入（RNE）[36]

RNE 是一种典型的嵌入式特征选择方法，该方法首先利用每个数据点和其近邻点都接近局部线性的现实，故采用局部线性嵌入算法获得特征权重矩阵。其次，使用l_1范数来使损失函数最小化，损失函数能够抑制离群值和噪声的影响。此外，RNE 是非平滑的凸函数，则利用乘法交替更新法（ADMM）来解决。目标函数如下所示：

$$\arg\min_{\mathrm{H}} \|\mathrm{AH}\|_1 + \frac{\alpha}{4}\|\mathrm{H}^{\mathrm{T}}\mathrm{H} - \mathrm{I_m}\|_{\mathrm{F}}^2 + \mathrm{tr}(\beta \mathrm{H}^{\mathrm{T}}) \tag{2-7}$$

其中，$\mathrm{A} = (\mathrm{I_n} - \mathrm{W}^{\mathrm{T}})$X,W 为特征权重矩阵，X 为数据集，$\mathrm{H} \in \mathbb{R}^{\mathrm{d} \times |l|}$为指示矩阵。AH 可以抑制离群点和噪声的影响[37]。$\alpha > 0$ 是惩罚项系数，能够使得$\|\mathrm{AH}\|_1$足够小并且$\|\mathrm{H}^{\mathrm{T}}\mathrm{H} - \mathrm{I}_m\|_F^2$不会特别大。$\beta \in \mathbb{R}^{d \times m}$是拉格朗日乘子，使得 $\mathrm{H} \geq 0$。第二项是对$\mathrm{H}^{\mathrm{T}}\mathrm{H}$ 和I_m差异的惩罚。

(3) 结合判别分析和$l_{2,1}$范数最小化的无监督特征选择方法（UDFS）[38]

UDFS 方法基于数据标签可以通过线性分类器预测的假设，采用把判别型分析和$l_{2,1}$范数最小化整合成一个框架的无监督特征选择方法。目标函数如下所示：

$$\min_{\mathrm{W}^{\mathrm{T}}\mathrm{W}=\mathrm{I}} \mathrm{Tr}(\mathrm{W}^{\mathrm{T}}\mathrm{MW}) + \gamma \sum_{\mathrm{i}=1}^{\mathrm{d}} \|\mathrm{w}^{\mathrm{i}}\|_2 \tag{2-8}$$

其中，W 矩阵为特征权重矩阵，w^i为 W 矩阵的第 i 行。由上式可看出最

优的 W 矩阵的很多行会缩减为 0。因此，$W^{T}x_i$即可以用少量特征表征x_i。根据$\|w^i\|_2$对所有特征进行降序排序，最后选择排序靠前的特征子集。

2.3 基于超图学习的无监督特征选择

图拉普拉斯矩阵通常被用来保存原始数据的局部结构，并且该类算法在很多情境下都有很好的表现。但是，基于简单图的无监督特征方法选仍然存在三个方面问题。首先，基于嵌入式方法的两步走策略（学习相似矩阵和构造特征选择矩阵）可能会降低特征选择的效果，因为相似矩阵的学习是为了得到最优的相似矩阵，而不是获得最优的特征选择结果。其次，他们通常认为数据样本的分布是相同的，并且数据样本之间是独立的。即使数据实例的来源都一样，它们也经常受到外部条件的干扰和影响，例如：人脸图像中光照变化。除此之外，在之前的方法中，简单图保存的局部几何结构只能描述数据的成对关系，不能够捕捉数据的高阶关系，以至于隐藏在数据背后的复杂结构不能被有效的利用。而在真实世界中，数据的特征空间往往有着多种关系，而不只是简单的成对关系，数据的成对关系不能够充分的表示高阶关系。

能够挽救信息损失的一个天然方法是将数据用超图而不是简单图表征。超图允许顶点通过超边进行多重连接，因此可以捕捉要素之间的多重或更高阶的关系。由于具有表征多重关系的有效性，基于超图的方法已经被用于很多实际的问题，比如划分电路网表，聚类和图像分割等。对于多标签分类，Sun 等人通过构建了一个超图，利用包含在不同标签中的相关性信息。在这个超图中，实例对应于顶点，每个都包括了用公共标签标注的所有实例。利用该超图表征，可以探索相同标签下多个实例之间的高阶关系。根据谱图嵌入理论，它们通过线性转换把数据转换到一个低维空间，超图保存了样本标签的关系。Huang 等人使用超图算法解决了一个无监督图像分类问题，在这篇研究中超图根据形状和外观特征表征无标签图像的复杂关系。具体来说，他们首先提取每幅图像的感兴趣区域（ROI），然后根据 ROI 中的形状和外观特征在图像之间的超边。超边被定义为每个顶点（图像）形成的群，或者它的 K 个近邻（基于形状和外观描述符）。通过这种方式图像分类的任务转换

成了使用一个超图切割算法可以解决的超图分割算法。下面介绍几种基于超图无监督特征选择的目标函数构造方法。

(1) 自适应超图无监督特征选择（AHLFS）[39]

AHLFS 将相似矩阵的学习和子空间学习（通过学习一个动态超图）整合在一个框架中，并且使用不同的方法进行特征选择使得该方法能够更加有效和鲁棒的选择有信息的特征子集。目标函数如下所示：

$$\min_{S,H,D_e,D_v,W} \mathrm{tr}(S^{T}XLX^{T}S) + \alpha\|W\|_2^2 + \beta\|S\|_{2,1} \qquad (2\text{-}8)$$

$$\text{s.t.} \quad w^{T}\mathbf{1} = 1, \quad w_i > 0, \quad S^{T}XX^{T}S = I$$

其中，S 是特征权重矩阵，X 是原始数据集，W 是超边权重。通过 S 矩阵的$l_{2,1}$范数使得 S 的行稀疏化，从而选择有信息的特征。并且变量 S 可以同时选择有信息的特征（通过稀疏性限制），构造子空间学习（通过 2-8 式第一项）和全局信息（通过正交约束）。

(2) 联合超图学习和稀疏回归（JHLSR）[40]

JHLSR 不是简单的通过超图学习刻画数据的局部关系，同样也使用预定义的相似性度量的方法保存样本之间的相似结构。因此 JHLSR 将超图学习和相似样本学习整合在一个新的框架中，目标函数如下所示：

$$\min_{S,W_H} \|(AXS)(AXS)^{T} - K\|_F^2 + \mu\mathrm{tr}\left(S^{T}X^{T}\hat{L}_H XS\right) + \lambda\|S\|_{2,1} + \gamma\|\mathrm{diag}(W_H)\|^2$$

$$\text{s.t.} \quad \sum_{j=1}^{m} W_i^j = 1, \quad W_H^j > 0, \quad j=1,\ldots,m \qquad (2\text{-}9)$$

其中，$A \in \Re^{l\times n}$是一个二项选择矩阵，K 是预定义的相似矩阵，$S \in \Re^{d\times k}$中，d 表示数据集特征的数量，k 表示转换后数据的维度。（2-9）式的第一项，通过计算样本之间成对的相似性可以保存数据的全局结构，第二项则探讨了数据的局部结构，第三项$l_{2,1}$范数促进了行稀疏性，最后一项可以使没用的超边设置为 0。

(3) 基于联合超图嵌入和稀疏编码进行数据表征（JHESC）[41]

JHESC 方法把超图嵌入和稀疏编码整合在一个统一的框架中。超图嵌入能够提供数据的低维表征，能够保存数据的高阶关系。同时，对数据进行稀疏表征能够捕获数据的内部几何结构。JHESC 采用两步走的策略分解数据矩阵，首先，基于 K 近邻方法构建超图；其次，通过嵌入超图学习新的基础矩阵 B；最后我们对矩阵 C 的每一列进行稀疏约束。JHESC 框架具体展示如下：

1) 基础矩阵学习

$$\mathcal{L}Y^{T} = D_{v}Y^{T}\Lambda, Y = X^{T}B \quad (2\text{-}10)$$

2) 系数表征

$$\min_{C} \|X - BC\|_{F}^{2} + \alpha \sum_{i=1}^{n} \|c_{i}\|_{1} \quad (2\text{-}11)$$

其中，$\mathcal{L}$表示超图拉普拉斯矩阵，矩阵 Y 的每一行表示原始数据点的低维嵌入，D_v表示超图的度矩阵。根据非负矩阵分解原理[42]，把矩阵 X 分解为非负矩阵$B \in \mathbb{R}^{m\times d}$和$C \in \mathbb{R}^{d\times n}$,即它们的乘积可以近似原始数据矩阵 X。学到的 B 矩阵能够捕获数据点的高阶关系。同时，新的表征矩阵 C 将是稀疏的，能够学习到数据的内部几何结构。因为将 B 和 C 二者其一固定，另一矩阵是凸函数，因此采用迭代更新算法优化 B 和 C。参数α是权衡参数，能够平衡稀疏性和重构误差的大小。

2.4 基于特征选择的经济学案例分析

特征选择方法常用在作为机器学习和数据挖掘的预处理阶段，在生物信息、工业故障识别与诊断、图像识别和信用风险评估领域都有广泛的应用。本节会对部分比较典型的应用领域进行介绍，同时，也会对在这些领域的算法进行简要阐述。

特征选择方法在医学图像分割中具有重要的应用价值，李卫伟提出了一种特征自选择机制的特征选择方法进行图像分割[43]，该方法主要是基于互信息量和交叉验证的理论基础，实验表明该方法对于不同的分割对象都可以选择最佳的特征组合，显著提高了分割效果。在图像检索领域的特征变得更加精细化，产生了很多不相关和冗余的特征，特征选择技术不可或缺。李国祥等人使用一种有效的连通图特征选择方法[44]，该方法结合特征距离和特征尺度等多种属性，构建特征分离图，把问题转换为在确保图像匹配精度的前提下，使特征分离图的阶最小。并通过是实验证明了该方法可以提高图像的检索精度。

生物信息领域会产生很多高维小样本数据，只有很少部分的基因与疾病有关，对基因数据进行分析的首要任务就是进行特征选择，因此特征选择方

法是生物信息领域也有很大的发展空间。张维健针对生物信息数据样本少，维度高的数据提出了一种特征选择方法 ReliefF-WS [45]，ReliefF 是一种快速有效的过滤式特征选择算法，ReliefF-WS 在 ReliefF 的基础上应用了类重叠样本加权的思想，可以降低 ReliefF 算法在更新特征权重时差样本带来的影响。

近些年，互联网金融发展迅速，很多商业银行和相关金融机构出现了越来越多的信用风险问题。信贷风险评估模型通过对客户的信息和活动数据进行识别，可以发现客户潜在的风险。李霜提出一种多种过滤器结合 NSD（New Separable Degree）指标的特征选择方法[46]。Lin 等人提出将风险指标和破产风险扩展到计分卡中，用于企业经营绩效评估，因此建立了一种融合混合过滤器-包装子集选择（HFW），随机向量功能链接网络（RVFLN）和蚁群优化（ACO）的融合机制，进行经营绩效预测[47]。Hall 等人[48]同时考虑了特征变量对目标变量的判别程度和特征之间的冗余性，提出了基于相关性的特征选择方法（Correlation-based Feature Selection，CFS），CFS 方法将该特征变量子集作为考察对象，有研究者已经将该方法应用到风险管理领域当中，Duma 等人[49]将该方法应用到保险数据集中，并从中找出了影响保险风险的最优特征子集。从多种角度对信贷客户的特征进行度量和评估，在一定程度上，避免了使用单一过滤器容易忽略信贷客户的多方面信息的问题。实验表明该方法选择的最优特征子集可以显著提高信贷客户风险评估的分类精确率。

2.5 本章小结

本节主要是介绍了国内外研究综述，首先，对特征选择方法进行了概述，使我们对特征选择方法的各个方面有比较全面的了解。其次，我们介绍了以往基于简单图的相关研究，并分析了其方法的局限性。最后，由于简单图存在的问题，我们提出了利用超图进行无监督特征选择的方法，能够解决简单图存在的问题，进一步引出了本文基于自适应超图的无监督特征选择方法研究。最后，结合学科背景和实验内容，对经济学中的相关应用研究进行了描述。

3 基于自适应超图的无监督特征选择方法研究

无监督特征选择可以降低原始高维数据的冗余度和噪声。之前的研究大致分为基于简单图或者单个超图的无监督特征选择方法。简单图只保持了数据之间的成对关系，然而很多数据集的样本之间存在多重关系，超图可以表示这种多重和高阶关系，可以更好的保留数据的局部几何结构。本章提出了一种超图学习模型，考虑到不同的超图之间具有互补性和一致性信息，我们利用多个超图自适应的学习一个共同超图，并将自适应学习超图和特征选择整合到一个框架，相互促进，有利于选出最优的特征。最后本文在求解目标函数时使用了一种高效的迭代优化算法。在 7 个数据集上的进行试验，实验结果表明，本章所提出的模型的性能优于其他先进的基线模型。

3.1 引言

在有监督学习中，数据样本的标签可以提供鉴别性信息，有利于探索数据的局部几何结构。而无监督学习缺乏标签信息，则如何利用数据集内在的结构以保存数据的局部流形结构就显得尤为重要。目前无监督选择的方法中，嵌入式方法比过滤式方法和包装器方法有更明显的优势，其主要通过学习到一个特征子集来提高模型的准确性。很多嵌入式方法首先会构造一个相似度矩阵，并且通过简单图来度量数据之间的成对关系，进而保存数据的局部和全局几何结构。并且会使用图正则化项加上稀疏约束（比如：l_1范数正则化项或者$l_{2,1}$范数正则化项）选择有信息的特征。

但是当前的嵌入式无监督特征选择算法仍然有缺陷。大多数方法采用的两步走策略：首先，学习相似矩阵；其次，构造特征选择矩阵。但旨在实现最优相似关系的相似矩阵而不是实现最优特征选择的方法会降低特征选择的性能。除此之外，当前通过构造相似矩阵实现无监督特征选择的方法通常都

会包含冗余和不相关的特征。并且一般通过简单图构造的相似矩阵，只度量了数据的成对关系，而不会考虑数据的高阶关系，导致其不能够充分的挖掘到数据的复杂结构。

为此，本章提出了一种全新的无监督特征选择方法：自适应超图无监督特征选择（AHUFS）。超图学习可以保存数据的多重或者高阶关系，我们基于不同的核函数构造了不同的超图关联矩阵。为了充分的利用不同超图间的一致性信息，将多个超图学习一个自适应超图，这样可以使特征选择结果更具有鲁棒性。总的来说，本章主要做了如下贡献：

(1) 利用超图学习构造关联矩阵，挖掘数据的多重或几何结构，并且利用多个超图自适应学习一个共同超图，使得学到的超图更加具有代表性和鲁棒性。

(2) 本章的实验将自适应超图学习和特征选择整合到一个框架中，相互促进，有利于选出最优的特征权重矩阵和超边权重矩阵。

(3) 本文提出了一种有效的迭代优化算法，并且分析了该算法的参数敏感性和收敛性。并在 7 个数据集上进行了实验，实验表明提出算法的性能在所有数据集上均优于其他先进的基线算法。

3.2 自适应超图的无监督特征选择

3.2.1 符号表示

这篇论文把矩阵定义为加粗的大写字母，把向量定义为加粗的小写字母，标量用斜体表示。一个矩阵 $\mathrm{X}=[\mathrm{x}_{\mathrm{i,j}}]$的第 i 行和第 j 列可表示为$x_i$或$x_j$，它的弗朗贝尼乌斯范数和$l_{2,1}$范数分别表示为 $\|\mathrm{X}\|_{\mathrm{F}}=\sqrt{\sum_{\mathrm{i}}\sum_{\mathrm{j}}\mathrm{x}_{\mathrm{i,j}}{}^{2}}$ 和 $\|X\|_{2,1}=\sum_{\mathrm{i}}\sqrt{\sum_{\mathrm{j}}\mathrm{x}_{\mathrm{i,j}}{}^{2}}$，同时也把矩阵 X 的迹、转置和逆表示为 tr（X），X^{T}和X^{-1}。表 1 总结了一些重要的概念和下文使用的相应定义。

表 1 本文重要标注的列表

标注	定义
X	原始数据矩阵 $X\in \Re^{n\times m}$，其中 n 是样本数量，m 是特征数量
G=（V,E,W）	超图
c	超图个数
V	超图的度 $V\in \Re^{n\times n}$
E	超图的边 $E\in \Re^{n\times n}$
W	原始数据的特征权重矩阵 $W\in \Re^{n\times n}$
$H(v_i,e_j)=\begin{cases}1, & v_i\in e_j,\\ 0, & \text{其他}\end{cases}$	超图的关联矩阵 $H\in \Re^{n\times n}$
$\delta(e)=\sum_{v\in V}H(v_i,e_j)$	超边度矩阵的计算
D_e	超边的度矩阵$D_e\in \Re^{n\times n}$
W_H	超边权重矩阵$W_H\in \Re^{n\times n}$

3.2.2 超图学习

图学习[50][51]和超图学习[52, 53][54]已经在很多应用中取得了不错的成绩，在简单图中，一个顶点代表了一个样本，并且基于一些测距方法（比如：欧几里得距离），一条边连接了两个顶点。因此在简单图学习的方法中，只能表示样本之间的成对关系。这种做法无法体现原始数据的复杂关系。传统的方法使用简单图来表示数据的成对关系，例如简单图可以体现甲和乙共同写了一篇文章，乙和丙共同写了一篇文章，但我们更关注的是如何体现数据的高阶关系，例如甲、乙和丙共同写了第一篇文章，乙，丁和戊写了另一篇论文。通过构造超图可以很容易表示这种高阶关系。

不同于简单图的学习，根据已有的一些文献，在一个超图中超边可以连接超过两个顶点。从这个意义上说，超图学习可以自动构造复杂的数据关系，这是超图学习的一个优势。近几年，超图学习已经被广泛用于很多应用。比如，Gao 等人发明了一个基于超图的 3D 目标追溯和识别方法，并且已经取得

了最新的结果。在此方法中，不同样本的关系放在一个超图结构中表示。基于不同样本的聚类结果，构造多个超图，并为了充分利用这些信息，学习一个自适应超图来估计样本之间的高阶关系。所以这篇论文通过构建超图来保存数据的局部结构，从而可以表示比简单图更复杂的关系。

在本文中，我们将超图表示为 $G = (V, E, w)$，其中 $V = [v_i]$，$E = [e_i]$，分别表示顶点的集合或者超边的集合，$w = [w_i]$指的是超边的权重。超图的构造需要进行 3 个连续的步骤：

步骤 1：关联矩阵 H 能够表示顶点和超边的二项关系，即每一个元素表示规则如下：

$$H(v_I, e_j) = \begin{cases} 1，\ v_i \in e_j， \\ 0，\quad 其他 \end{cases} \tag{3-1}$$

步骤 2：这个权重向量 w 可以衡量超图的重要性；

步骤 3：构建超图的拉普拉斯矩阵，即利用构建的超图来构建正则化的拉普拉斯矩阵。

传统的简单图的边表示的是顶点和顶点的关系，而超图的关联矩阵可以表示顶点和超边之间的关系。为了实现这个目标，根据原始数据 $X = R^{c \times n}$，其中 c 和 n 分别表示特征的数量和样本的数量。每一个样本可以看成一个顶点，我们尝试让每个顶点都拥有一个超图。具体而言，通过下式的规则产生超边：

$$e_i = \left\{ v_i \middle| KNN\left(\theta(x_i, x_j)\right) \right\}，\ i，\ j = 1，\cdots，\ n \tag{3-2}$$

其中，$\theta(x_i, x_j)$度量了x_i和x_j之间的相似性（比如分别利用欧几里得距离与高斯核函数、多项式核和直方图交叉核等核函数求相似性），我们采用 KNN 算法构建超图，即将每个样本的 K 个近邻作为一个超边。

其次，通过构造好的关联矩阵 H 来学习超边的权重 w。然后，通过计算 $\delta(e_i) = \sum_{v_i \in E} h(v_j, e_i)$ 我们可以得到超边的度 $\delta(e_i)$，通过计算 $d(v_j) = \sum_{v_i \in e_I, e_i \in E} w(e_i) h(v_j, e_i)$可以得到顶点的度 $d(v_j)$。

最后通过计算下式得到拉普拉斯矩阵：

$$L = I - D_v^{-\frac{1}{2}} H W_H D_e^{-\frac{1}{2}} H^T D_v^{-\frac{1}{2}} \tag{3-3}$$

其中，$I \in \mathbb{R}^{n \times n}$是一个 n 阶的单位阵，$D_e$，$D_v$和 W 分别表示对角矩阵$\delta =$