摘 要

随着科学的进步和信息技术的发展，现代数据的类型和规模也在迅猛增长，它们往往具有高维特征，而高维数据会增加学习成本，导致模型训练速度下降，还会带来维度灾难等问题。同时，获取有标记的数据需要耗费大量的人力物力，因此从现实场景中获取的数据通常不具有完备的标签信息，即只有少部分样本有标签信息，大部分样本没有标签信息。因此，如何对具有少量标签信息的高维数据进行特征选择，已经成为大数据挖掘领域的研究热点。

现有的半监督特征选择方法大多存在以下问题：一是在训练时没有考虑样本的重要性，认为所有样本是同等重要的，这就会导致噪声样本和异常值对模型的性能产生影响；二是在训练时只考虑了特征的重要性，忽视了特征的冗余性，导致相似的特征包在最优特征子集中同时存在，这将对模型的性能产生影响。

为了解决上述问题，本文提出了一种基于自步学习的半监督特征选择算法。首先，分别通过岭回归和自表示学习来对有标签样本和无标签样本进行特征选择，并通过赋予权重的方式将他们融合到一个半监督特征选择框架中。接着，本文将混合式自步学习融入上述框架中，通过自步学习自动学习样本的重要性，在训练过程中控制训练样本的数量，可以有效地抑制噪声异常值对模型的影响，进而提高模型的性能。最后，为了解决特征冗余问题，本文引入了一个正则化项，主要思想是考虑特征之间的成对相似性，使相似的特征不太可能同时被选择，以保证选择出低冗余的特征。

为验证本文所提算法的有效性，将其与其他 6 种特征选择算法在 6 个真实数据集上进行了比较。实验结果表明，在大部分情况下，本文所提算法在特征选择率变化下以及标签比例变化下的性能都是优于其他特征选择方法的。

关键词：半监督 特征选择 自步学习 自表示学习

Abstract

With the progress of science and the development of information technology, the size and complexity of modern data are increasing rapidly, often containing high-dimensional features. High-dimensional data will increase the learning cost, lead to slow model training, and also bring problems such as the curse of dimensionality. At the same time, acquiring labeled data requires significant amounts of time, manpower, and resources. Therefore, data obtained from real-life scenarios usually lack complete label information, where only a handful of samples have label information, and the rest do not. Consequently, the selection of features from high-dimensional data with limited labeling information is a hot topic of research in the field of big data mining.

Most of the semi-supervised feature selection methods currently available suffer from the following issues: Firstly, they do not account for the significance of individual samples during the training process. As a result, all samples are treated equally important, leading to noise samples and outliers that can affect the model's performance. Secondly, only the importance of features is taken into account when training when training, but the redundancy of features is ignored. Consequently, similar features are selected and included in the optimal feature subset, which can impact the model's overall performance.

To address the aforementioned issues, this paper propose a semi-supervised feature selection algorithm based on self-paced learning. Firstly, we utilize the ridge regression model and self-representation learning model to select features for labeled and unlabeled samples respectively. These models are then integrated into a semi-supervised feature selection framework with assigned weights. Subsequently, we incorporate hybrid self-paced learning into the above framework, automatically determining the importance of samples through the learning process. By controlling the number of training samples during the process, we can effectively suppress noise outliers' influence on the model and

enhance its performance. Finally, to resolve feature redundancy issues, we introduce a regularization term. The main idea is to consider the pairwise similarity between features, preventing similar features being chosen simultaneously, thereby ensuring that low redundancy features are selected.

To validate the efficacy of the proposed algorithm, it was compared with six other feature selection algorithms across six real datasets. The experimental results reveal that the performance of the proposed method surpasses that of the other feature selection methods, particularly under varying feature selection rates and label ratios.

Keywords: Semi-supervised Feature selection Self-learning Self-representation

目 录

1.绪论 1
1.1 研究背景与意义 1
1.2 国内外研究现状 4
1.3 研究内容与组织结构 8
1.4 本章小节 9
2.相关理论知识 10
2.1 特征选择 10
2.1.1 过滤式 10
2.1.2 封装式 12
2.1.3 嵌入式 12
2.2 自步学习理论 13
2.2.1 课程学习 14
2.2.2 自步学习 14
2.3 本章小结 16
3.基于自步学习的半监督特征选择算法 17
3.1 模型构建 17
3.2 模型的优化及分析 20
3.2.1 优化算法 20
3.2.2 复杂度分析 24
3.3 本章小结 25
4.实验结果及分析 26
4.1 实验方案设计 26
4.1.1 实验数据集 26
4.1.2 比较方法 28

4.1.3 评价指标 29
4.1.4 实验设置 30
4.2 实验结果分析 30
4.2.1 特征选择比例变化下的算法性能分析 34
4.2.2 标签比例变化下的算法性能分析 36
4.2.3 收敛性分析 39
4.2.4 参数敏感性分析 40
4.3 本章小结 41
5.总结与展望 43
5.1 总结 43
5.2 展望 44
参考文献 45
致谢 49

1.绪论

本章将先介绍本文的研究背景与意义，接着再介绍一些国内外的研究现状，并对现有的方法进行简单的介绍，最后将介绍本文的研究内容和论文的结构安排。

1.1 研究背景与意义

自 1946 年第一台计算机诞生以来，计算机的软件与硬件都在不停地进行更新换代，当今计算机的性能相较于以前已大为提升。这一变化也悄然改变了人们的生活习惯与思维方式，使人们进入了以数据为核心的第四次工业革命时代。例如，在零售方面，传统的卖家根据以往的销售经验向不同风格的客户提供相应的商品，其销售成功率和卖家的个人属性挂钩；而现在大部分零售企业依托于互联网和历史留存数据，打破了传统零售的信息壁垒。在实际操作中，线上零售会留存用户的基本信息、生活方式、娱乐偏好等数据生成人群画像，给消费者提供针对性的推荐和个性化的服务，激发用户的购买欲望，进而提高用户的粘性，增加企业的收入。在疾病诊断方面，过去只能通过医生的经验来进行判断，其效果完全取决于医生的水平。如今医院将患者的各项检查数据存入医疗信息化系统，并根据这些医疗数据建立相应的疾病诊断模型，辅助并完善医生的诊断，减轻了医生的工作量。由以上两个案例可以得知，数据对于国民生产的方方面面都具有重大的意义。与此同时，随着国家对信息化建设的重视以及人们逐渐意识到数据的价值，各行各业都留存了大量的数据，据研究表明，截止到 2012 年，数据的计量单位已经从 G、TB 级别跃迁到 PB、EB 甚至 ZB 级别，面对体量如此庞大的数据，仅仅靠人工的方式来处理和分析数据，总结规律来指导决策是不现实的。因此，怎样有效地从大量的数据中挖掘出有用的知识，进而提高生产效率和经济效

益逐渐成为研究的重点。

为了从海量的数据中排除不相关、有干扰的信息，找到真正影响业务的核心因素，机器学习（Machine Learning）应运而生。机器学习是一门多学科交叉融合的专业，它包含数学、统计学和计算机知识，通过建立模型来获取数据中的知识。从有无标签的角度来划分，机器学习可以分成以下三类：无监督学习、有监督学习和半监督学习。具体来说，无监督学习中使用的数据不包含标签信息，通过发现数据的内在逻辑与结构来训练模型以解决问题，比如常见的聚类问题和降维问题，有图形图像的特征降维，信号的处理，文档主题的挖掘等。有监督学习是指使用标签完备的数据集来训练模型以解决问题，比如常见的分类问题和回归问题，有垃圾邮件识别，疾病诊断，房价预测等。但是，在现实生活中获取标签完备且准确的数据是十分困难的，因为生活中大都是无标签的数据，要获取有标签的数据需要对数据进行标记。在业界，数据标记被称为人工智能背后的“人工”，该项工作需要耗费大量的人力物力。因此，有监督学习的成本较高。但面对标签信息较少的数据，传统的有监督学习方法很难获得比较好的性能。同时，无监督学习虽然不需要有标记的数据，但由于该学习方法缺乏先验知识可能导致模型性能的下降，所以人们希望能够利用少量有标签的数据和大量无标签的数据进行训练，这就是半监督学习，它可以在一定程度上解决上述问题。因此，对半监督学习的研究具有一定的现实意义，该方法可以减少数据标注的代价，同时利用小部分标记数据提升模型性能。

随着大数据的兴起和信息技术的发展，数据的类型和规模也在迅猛增长，近现代的数据往往具有特征维度高、样本数量多的特点，如基因数据、高像素的图片数据、视频数据等。海量和高维的数据虽然带来了更加丰富全面的信息，但过多的数据样本和高维特征对科学研究也是一个挑战。显而易见的是高维度的数据会增加运行时间，导致算法效率下降；并且随着数据维度的增加，数据在高维特征空间中会变得稀疏，这会导致模型效果下降，理论上可以通过增加样本数量来解决这一问题。然而，要解决上述问题，样本数量需要随特征数进行指数增长，这在实际生活是不现实的。同时，高维数据也带来了大量不相关和冗余的信息，这会给模型带来干扰，增加学习难度，进而导致模型性能的下降。由此可以得知，过高的特征维度会带来“维度灾难”

等问题，而降维后可以有效地节省计算时间，减少噪声和不相关信息对模型的影响，提高模型的精度，因此对高维特征进行降维是十分有必要的。常见的降维方法有两种：一是构建优秀的低维子空间，该方法通过一些规则，将原始高维特征空间映射到低维特征空间。该方法虽然降低了原始特征空间的维度，但是在空间转换过程中会丢失原始特征的含义，因此该类方法对需要解释特征含义的问题没有实际意义。二是进行特征选择，该方法会保留原始特征。具体操作是从所有的特征集合中根据特定规则保留重要的特征或删除不相关、冗余的特征，构建最优的特征子集。这种方法不仅进行了特征筛选，还保留了特征的原始含义，对实际应用有较强的指导意义。在降维时，特征选择的比例也是一个值得探讨的问题，选择的特征较少会导致信息的损失，从而导致选出的最优特征子集不能很好的反映全体特征的情况，而如果选择较多的特征则可能会增加计算成本，甚至包含了噪声，导致模型性能下降。

海量的数据不可避免的会包含噪声。对于有标签信息的数据来说，噪声包含样本噪声和标签噪声。样本噪声是数据集中的数据信息不准确、存在误差等原因导致的。标签噪声一般是由于人工标注错误或者定义不明而导致的，常见的标签噪声可以分成随机噪声和跨类别噪声两种，随机噪声是指不属于数据集中任何一个类别的样本被标注了类别信息，即有大量的和该数据集每个类别都不相关的样本被标注了信息；跨类别噪声是指属于某一类的噪声却被误分到另一类中。对于模型来说，有噪声的数据集会降低模型的稳健性，使模型在优化过程中陷入局部最优，对模型的性能产生影响。因此，在训练模型时如何解决噪声问题也是至关重要的。同时，数据集中的每个样本对模型性能的影响是不一样，有的样本是重要样本，对模型性能影响大，有的样本是次要样本，这些样本会影响模型的性能，但其对模型的影响介于噪声样本和重要样本之间，把这类样本看作噪声样本或是重要样本放入训练都是不合适的。因此，在训练模型时，考虑样本的重要性也是十分有必要的，而混合式自步学习可以很好地解决上述问题。自步学习是通过模仿人类在学习中的认知机理来进行学习的，人在学习中会先选择容易理解的，简单的内容，再学习自己认为复杂的内容。将自步学习融入特征选择方法，可以自动学习样本的重要性，在训练过程中控制训练样本的数量，有效地抑制噪声对模型的影响，进而提高模型的性能，改善算法的稳健性。因此，本文提出的基于

自步学习的半监督特征选择算法有一定的现实意义。

1.2 国内外研究现状

随着科学技术的发展，人们已经进入了信息化时代。虽然这极大地方便了人们的日常工作与生活，但与之相关的数据信息也正以指数形式增长，如高清图像有着更丰富的信息，流媒体的发展扩宽了人们的视野，人脸识别技术方便了人们的生活，与此同时也带来了大量的高维数据，过多的特征可能会造成信息冗余，提高计算成本，并产生“维度灾难”等问题。因此，对高维数据降维是十分有必要的，近几年在降维问题上，国内外学者都开展了深入的研究，提出了许多行之有效的方法。

顾名思义，降维是将原始的高维特征空间转化成一个低维的特征子空间，在这个过程中希望能够获取原始高维空间中重要的、具有代表性的特征，去除掉冗余的、会对后续分析产生干扰的特征，使得降维后的特征子空间能够更好地反映原始高维空间的结构。常见的降维方法有以下两种：一是特征提取，即对原始的特征按照一定规则进行变换、组合，得到一组低维的、非冗余的特征，也可以说是获得一个高维特征空间投影后的低维特征子空间。通过这种方法降维所获得的特征失去了实际意义，存在可解释性弱的问题，但该方法更有利于探究特征空间的内在结构。常见的方法有主成分分析方法（Principal Component Analysis），简称 PCA [1]，该方法属于无监督降维方法中的一种，在高维数据建模的场景中被广泛使用，如图像处理、变量选择等。它的主要思想是通过投影变换，将高维空间中的数据样本投影到低维空间，即从原始的特征空间中构造出能够反映原始特征空间的主成分来进行降维，它的第一个主成分是在原始数据空间中方差最大的方向，第二个主成分是在跟第一个主成分正交的平面中方差最大的，以此类推，获得剩余的主成分。通过上述方法获得的前几个主成分可以包含原始特征空间的大部分信息，从而实现了降维；还有线性判别分析方法（Linear Discriminant Analysis），简称 LDA [2]，该方法属于有监督降维方法中的一种，主要应用于模式识别领域，例如，人脸的识别和检测、风险欺诈识别和目标跟踪检测等等。该方法最早在 1936 年由 Fisher 提出，它的主要思想是将原始特征空间投影到一个新的

低维子空间，并利用有监督的信息，希望同类数据在低维子空间中的距离近，异类数据在低维子空间中的距离远。简单地说，就是类内距离最小化、类间距离最大化。还有因子分析方法，简称 FA（Factor Analysis），该方法最早是由英国的心理学家斯皮尔曼在研究学生成绩时发现的，主要用于人文社科以及社会研究方面。该方法的主要思想是探究原始高维特征空间中特征的相关性，找出潜在的共性因子，将相关性高的、本质相同的特征都归入同一个因子中，进而减少原始特征的数量，从而达到降维的效果。还有独立成分分析方法，简称 ICA (Independent Component Analysis)，该方法是无监督降维方法中的一种，主要应用于信号处理领域，可以在混合信号中提取源信号。

以上是一些常见的特征提取方法。接下来介绍第二种降维方法——特征选择，该方法根据特定的规则从原始的特征集合中挑选出最优的特征子集，该操作可以有效地节省计算时间，减少噪声和不相关信息对模型的影响，提高模型的精度[3]，同时它没有改变原始的特征，保留了特征的原始含义，在后续的分析过程中具有更好的解释性。根据数据样本是否拥有标签信息，特征选择方法可以分为以下三类：

第一种是有监督的特征选择方法[4]，该方法使用的数据集拥有完整的类别标签信息。有监督的特征选择算法一般是通过与类标签或目标的相关性来进行研究，进而确定特征的重要性，该方法的准确率较高，因而近年来有许多关于有监督特征选择算法的研究。Song 等人（2007）提出了一个过滤式特征选择方法的框架，该框架使用希尔伯特-施密特独立标准作为标签和特征之间相关性的度量，在此标准下，特征越好，其值越大[5]。Jose 等人（2010）提出了一种基于度量的，既可以应用于离散型数据又可以应用于连续型数据的有监督特征选择方法。该方法是基于信息论的，通过计算特征与类标签之间的条件互信息来进行特征选择。同时，Jose 等人在文章中还提出了一种新颖的方法，该方法可以有效地克服向前向后算法的嵌套问题[6]。Murthy 等人（2012）对文本分类数据中的特征选择方法进行了研究，提出了一种新的有监督特征选择方法，该方法利用文本数据中的词和文本所属类别标签之间的相似性来进行特征选择，文本数据中的每一个单词都会依据和类别的相似性被赋予一个分数，分数较大的单词将会被作为重要的特征[7]。Chen 等人（2018）提出了有监督的分层特征排序方法，首先根据有标签的数据去识别

每个特征对每一类的重要性，并将特征进行聚类，然后根据学习到的子空间权重对不同特征簇中的特征分别进行排序，使得提取出的特征既丰富又多样性[8]。Wu 等人（2021）提出了一种带有特征加权的基于正交最小二乘回归模型的有监督特征选择方法，该方法将特征加权引入了正交回归，与传统回归方法相比，该方法能够保留更多的信息。同时为了解决不平衡的正交普鲁克问题，该方法采用广义幂迭代方法来求解回归矩阵 [9]。

第二种是无监督的特征选择方法[10]，该方法不需要使用类别标签信息，而是通过构造伪标签、探寻内在结构等方法来进行特征选择。同时无标签这个特性也与实际情况十分适配，在现实生活中，大部分数据都是没有标签的，而想要获取完整的有标签数据需要耗费大量的人力物力。因此，无监督的特征选择方法被广泛研究。Mitra 等人（2002）提出了一个通过衡量特征之间相似度来进行特征选择的方法，该方法的复杂度比较低，计算起来十分高效，适用于大样本和高维度数据。该方法中衡量相似度的指标被称作最大化信息压缩指数[11]。Qian 等人（2013）将具有鲁棒性的非负矩阵分解和局部学习等方法与特征选择方法相结合，提出了一个具有鲁棒性的无监督特征选择方法。这种方法与以往生成伪标签的方式不同，它通过局部学习和非负矩阵分解来学习伪聚类标签，同时在标签学习的过程中，为了减轻噪声和异常值的影响，将$L_{2,1}$范数加入特征选择学习框架中[12]。Wang 等人（2015）通过稀疏学习直接将特征选择方法嵌入到聚类算法中。这种方法优化了以前的某些特征选择算法，比如之前的某些算法由于无监督特征选择缺乏标签信息，就先通过聚类算法来生成数据的标签信息，再根据这些获得的标签信息将原始的无监督特征选择转化成有监督特征选择[13]。Tabakhi 等人（2017）提出了一种基于蚁群优化算法的无监督特征选择方法，该方法是通过特征之间的相似度来计算特征之间的相关性，再通过多次迭代来寻找最优的特征子集，从而减少了最优特征之间的冗余[14]。Ding 等人（2020）将潜在表示学习嵌入到特征选择中，提出了一种新颖的无监督特征选择方法。该方法解决了传统无监督特征选择中存在的两个问题：第一个问题是传统的无监督特征选择算法常常假设数据之间有着相同的分布，且数据之间是相互独立的，即数据之间不存在依赖关系，这些假设显然与真实情况不相符。在现实生活中，数据样本之间往往是相互联系的，很少是独立的；第二个问题是在传统的无监督特征选择

方法中往往使用简单的相似矩阵来描述数据之间的关系，这种方法只能反映数据之间的成对关系，而不能有效的获取数据中复杂的高阶信息[15]。

第三种是半监督的特征选择方法[16]，该方法根据少量的有标签信息和大量无标签信息来进行特征选择。这种方法在一定程度上解决了有监督特征选择方法和无监督特征选择方法存在的一些问题。对于有监督特征选择方法，获取全部有标签的信息需要耗费大量的人力物力，而半监督特征选择方法只需要少量的标签信息，这就会大大降低人力物力的投入；而对于无监督特征选择方法，该方法在进行选择特征时没有标签信息，需要通过探求数据的潜在结构或构造伪标签等方式来进行特征选择，其精度往往低于有监督的特征选择方法。因此，近年来半监督特征选择方法也被广泛研究，下面具体来介绍一下。Zhao 等人（2007）为了解决数据中只有少量有标签样本的问题，他们建立了一个正则化框架，通过该框架同时利用小部分有标签的数据和大部分无标签的数据进行特征选择[17]。Zhao 等人（2008）提出了一种新的半监督特征选择算法，该算法通过最大化不同类别数据样本之间的距离来利用有标签数据，通过探寻数据空间中的几何结构和内在特征来利用无标签数据[18]。Xu 等人（2010）为了解决半监督问题中有标签样本过少而不能进行特征选择的问题，他们提出了一种基于流行正则化思想的半监督特征选择方法。该方法选择特征的标准是最大化不同类别数据样本之间的间距，并同时利用有标签数据和无标签数据的概率分布信息[19]。Chen 等人（2017）继承并发展了原始的最小二乘回归模型，并由此提出了一种新的半监督特征选择方法。该方法通过添加一个新的变量重新调整了最小二乘回归模型中的系数，并通过这个新的变量对特征进行排序，选取那些排名靠前的特征放入最优特征子集，同时这个变量具有很好的解释性。该方法还解决了之前某些方法不能同时求解投影矩阵全局解和稀疏解的问题[20]。Li 等人（2021）提出了一种基于广义不相关约束的半监督特征选择方法，该方法解决了将岭回归直接用于半监督学习而不能得到封闭解的问题，同时将流形结构统一到上述框架下，更好地利用了数据中的信息[21]。Zhong 等人（2021）提出了一种新的半监督特征选择方法，该方法不使用原始数据来生成相似度矩阵，而是在迭代的过程中自适应的生成相似度矩阵，这样就可以在一定程度上减少噪声对算法产生影响。同时，他们还通过调整$L_{2,p}$范数来控制相似度矩阵的稀疏性[22]。

1.3 研究内容与组织结构

本文研究的是半监督特征选择问题，提出了一种基于自步学习的半监督特征选择算法，主要包含以下内容：首先，本文将有监督特征选择方法和无监督特征选择方法通过赋予不同的权重的方式进行组合，生成一个半监督特征选择的框架。通过岭回归对有标签的样本进行特征选择，通过自表示学习对无标签的样本进行特征选择，通过网格搜索法来确定使模型效果达到最好的权重。接着，为了减轻噪声和异常值对模型的影响，考虑样本的重要性，避免目标函数过早陷入局部最优，提高模型性能，本文将混合式自步学习融入到半监督特征选择框架中。最后，为了解决特征的冗余性问题，我们引入了一个正则化项，利用特征之间的成对相似性来解决特征冗余问题。

接下来，给出本文的组织结构。本文一共由五个章节组成，主要内容如下所示：

第一章是绪论，首先介绍本文的研究背景与研究意义，接着再介绍一些国内外关于降维的研究情况，并对这些方法做了简单的介绍，最后再介绍本文的研究内容和组织结构。

第二章是相关理论知识，该章节主要介绍了与本文研究相关的一些基本理论知识，包含特征选择和自步学习理论这两个部分。

第三章是基于自步学习的半监督特征选择算法，在章节中，首先介绍了模型的目标函数，并阐述了构建模型的思想；接着给出模型的优化算法，并对算法的复杂性进行分析。在优化算法时，本文是通过交替迭代更新法进行的。

第四章是实验结果及分析，主要介绍实验方案设计和实验结果分析。首先，在实验方案设计一节中，本文将介绍实验数据集、对比模型、评价指标和实验设置。接着，在实验结果分析这一节，本文将从不同标签率和不同特征选择率来比较算法的性能，最后，还将对该算法进行收敛性分析和参数敏感性分析。

第五章是总结和展望，对全文的内容进行了总结，并对文章中可以改进的地方进行了展望。

1.4 本章小节

本章首先介绍了现代社会数据的特点——样本多、纬度高、有噪声、标签获取困难等，阐述了本文提出的基于自步学习的半监督特征选择算法的现实意义。接着详细介绍了降维的两种方法：特征提取和特征选择，并给出了三类特征选择方法的一些研究现状。最后介绍了本文的研究内容和组织结构。

2.相关理论知识

本章将分别介绍特征选择和自步学习理论的相关概念。在特征选择部分，本文将分别介绍过滤式特征选择方法、封装式特征选择方法和嵌入式特征选择方法。在自步学习理论部分，本文将先介绍课程学习，再介绍自步学习并给出一些常用的自步正则项。

2.1 特征选择

随着科技的进步和信息技术的发展，我们进入了大数据时代，现阶段的数据往往有以下几个特点：体量大、维度高、价值密度低等。而高维数据意味着该数据拥有大量的特征，但其中的某些特征对后续研究并没有意义，是多余的，这样的特征不仅会增加后续研究的时间成本，甚至还会给后续的研究带来干扰，进而产生负面影响，导致模型性能下降。同时，高维数据还会带来“维度灾难”等问题。因此，对高维数据进行降维是十分有必要的，特征选择就是一种有效的降维方式。顾名思义，特征选择是从原始数据集的所有特征中选择出一部分特征。具体操作是：首先从数据集的所有特征集合中，按照一定标准挑选出相应的特征子集，接着对挑选出的特征子集进行评价，如果满足相应的标准，则将该特征子集作为最优特征子集，否则重复上述操作。根据在选择特征时采用的策略和评价函数的不同，特征选择方法可以分为过滤式特征选择方法、封装式特征选择方法和嵌入式特征选择方法。

2.1.1 过滤式

过滤式特征选择方法（Filter methods）是根据一定的评价标准来分析数据本身或类标记信息，然后再评估所有特征的重要性，选择出重要的特征，

最后再将选出的特征放入最优特征子集。可以看到该方法在选择特征时不需要对模型进行训练，是通过探究数据本身的特点来选择特征，这样就会大大降低计算复杂度，进而提高时间效率。最常见的过滤式方法是 Fisher Score 特征选择方法，这是一个非常经典的算法，通过对数据集中样本的每一个特征进行打分，然后再根据得出的分数进行排名，最后选出符合要求的特征作为最终的特征子集。该方法利用的思想是同类事物往往具有相同的属性，但也存在不是同一类的事物拥有相同的属性。因此，要对事物进行有效地分类，就要选择在同一类别中都存在，而在不同类别中不存在的属性，这样就可以更高效地进行分类。因此，在 Fisher Score 特征选择方法中，对于优秀特征的定义是在同一个类别中取值相近，在不同类别中取值有较大差异。常见的过滤式特征选择方法还有根据相关性来进行特征选择，比如 Pearson 相关系数，它是用来度量两个变量之间的线性相关关系；比如构建卡方检验，来检验变量之间的相关性；比如借用信息论里的互信息的思想，互信息越大，则代表两个变量的相关性越高[23]。Kira 等人提出了 Relief 方法[24]，该方法利用样本间的近邻距离来给特征赋权重，简单地说就是不同类别的近邻距离与同类别的近邻距离之差越大越好，最后将权重超过阈值的特征作为最优特征。同时，Relief 方法作为一个经典的特征选择算法被广泛研究，并拓展出了 Relief-F 方法。Relief-F 方法是为了解决 Relief 方法只能处理二分类的特征选择而提出的，它的主要思想就是将拥有多个类别的数据看成二分类数据进行计算。以上介绍的一些方法大部分都在考虑如何筛选出重要的特征，而忽视了筛出的重要特征可能存在冗余的问题，过多重要但相似的特征并不能给模型带来性能的提升，删除重复的特征也不会导致分类结果的巨大变化。为了解决特征选择中存在冗余的问题，Peng 等人提出了 mRMR 方法[25]，该方法的全称是 Max Relevance and Min Redundancy，通过上述名字我们可以得知该方法的主要思想是寻找相关性最大的特征，即特征和类别变量之间要有比较强的相关性，同时寻找冗余度最小的特征，即选择的特征之间要尽量是不相关的，进而来形成最优的特征子集。

2.1.2 封装式

封装式特征选择方法（Wrapper methods）在选择特征时要根据对应的算法来进行评估，该方法选出的特征是使特定算法的效果达到最好的特征。封装式特征选择方法一般采用以下的方法来选取最优特征子集，比如前向搜索方法（Sequential Forward Selection）、后向搜索方法（Sequential Backward Selection）、递归特征消除方法（Recursive feature elimination）等。前向搜索方法首先定义候选特征集为空集，接着分别评估每个特征在模型中的效果，选择一个使得模型效果最好的特征放入候选特征集，接着再从剩余的特征中选择出一个特征，和候选特征集中特征组合在一起评估模型效果，将使模型效果达到最好的特征放入候选特征集，接着重复上述的操作，直至模型的效果无法产生明显的提升或者达到了指定的阈值，此时的候选特征集就是最优特征子集。那类似的，后向搜索算法是特征集合由多变少的过程。递归特征消除法首先将所有的特征作为候选特征集合放入模型中进行计算，并通过交叉验证的方法来验证模型的效果，同时得到候选特征集合中每个特征的重要性；接着，从上述的候选特征集合中删除掉重要性最低的那个特征，再次重复上述操作，直到候选特征集合里没有特征；最后，通过每次交叉验证得到的模型效果来确定最终的特征集合，即选择使得模型效果最好的候选特征集合作为最优特征子集。而上述的这几种方法都属于贪心算法，在进行特征选择时，都是根据模型的效果或者其他条件做出当下最优的选择，并没有从整体的角度上来考虑，这样就容易陷入局部最优，进而找不到最优秀的特征子集。同时，封装式特征选择方法需要进行多次训练，导致整个特征选择的时间较其他方法更长，效率更加低下。因此，封装式特征选择方法在实际研究中使用较少。

2.1.3 嵌入式

嵌入式特征选择方法（Embedded methods），也被称作混合式特征选择方法。该方法是在优化模型的过程中进行特征选择，最后同时得到训练好的模型和最优的特征子集，该方法融合了过滤式特征选择方法和封装式特征选择

方法的某些特点，相较而言，可以说该方法具备过滤式方法的高效率和封装式特征选择的高性能。因此，人们对嵌入式特征选择方法进行了广泛深入的研究，如 Shi 等人为了更好地对矩阵进行稀疏，将$L_{2,1}$范数替换成$L_{2,\frac{1}{2}}$范数[26]。Tang 等人将 LIR 模型扩展到半监督领域[27]。为了解决之前某些传统无监督特征选择方法单独计算每个特征分数而忽略特征之间相关性的问题，Cai 等人基于流形学习和正则化的思想提出了一种新的无监督特征选择算法，提升了特征选择的效果[28]。在特征选择时，某些学者利用拉普拉斯矩阵去学习数据的局部几何结构，但传统的这些方法主要是根据原始特征空间来生成相似矩阵，如通过 K 近邻等方式来生成，但由于原始特征空间有噪声和异常值，因而由此生成的相似矩阵将会无法正确表达数据的相似结构并导致模型性能的降低。为了解决这一问题，许多学者提出改变生成相似矩阵的方式，由简单的利用原始特征空间生成转变成自适应学习。如 Huang 等人通过自适应学习的方式获取相似矩阵，这一举措在获得数据相似结构的同时减轻了噪声和异常值的影响；同时，他们提出了一种新的具有不相关约束和正则化的广义回归模型以解决在特征选择的过程中相似结构不稳定和存在冗余的问题[29]。由此可以得知，如何减轻噪声和异常值，减少特征的冗余，也是特征选择方法中至关重要的问题。众多学者对于上述问题也做了不少的研究。Battiti 为了找到相关性强、冗余度小的特征，通过特征与类和特征与特征之间的互信息构建目标函数[30]。Zhao 等人通过构建包含样本相似度的目标函数来减少特征的冗余[31]。Shao 等人将$L_{2,1}$正则化合并到加权非负矩阵分解中，使得模型对噪声和异常值具有鲁棒性[32]。Nie 等人通过在训练模型时学习局部结构来降低噪声对模型性能的影响[33]。Li 等人定义了新的广义回归模型，并通过添加不相关约束和$L_{2,1}$范数正则化来减少冗余，通过自适应地学习相似性矩阵来减轻噪声和异常值的影响[34]。Wu 等人通过引入高相似度惩罚机制来减少冗余，通过根据数据情况灵活分配权重来减轻噪声和异常值的影响[35]。

2.2 自步学习理论

在这一小节中，将介绍关于自步学习理论的相关知识。本节将先介绍课

程学习的背景、提出者、主要内容和优缺点等内容，然后再引入自步学习理论，并对其进行介绍。

2.2.1 课程学习

在信息革命之后，计算机和互联网蓬勃发展，人工智能也逐渐成为了人们关注的焦点。人工智能是由麦卡锡等人提出的，其主要内容是讨论如何让机器像人一样进行思考、学习。其中，机器学习是人工智能中最重要的一个领域，也是现在热门的研究方向。课程学习就属于机器学习，它是由 Bengio 等人在 2009 年的 ICML 会议上提出的[36]。它的主要思想是基于人类学习的方式：从简单的知识开始学习，然后逐步提高知识的难度，去学习复杂的知识。在课程学习中会从简单的样本开始学习，然后再学习复杂困难的样本。当课程学习被应用到具体的算法时上，我们可以简单地定义简单样本就是使模型性能更高、损失更小的样本。当然，课程学习对于简单样本或者是复杂样本的定义会随着学习目标的不同而不同。Bengio 不仅在会议上创造性地提出了课程学习这种方法，还提出了该方法的优点：可以加速模型的训练，减少训练时的迭代次数，并且拥有更好的性能[36]。但我们可以明显的看到课程学习存在的问题：需要利用先验知识去人为定义样本的复杂度。

2.2.2 自步学习

通过上文的介绍，我们得知课程学习需要人为地预先定义样本的难度，这个操作在现实中比较困难，同时需要耗费较多的人力物力。因此，Kumar 等人[37]在 2010 年提出了自动确定样本难易程度的方法，即自步学习。该方法继承和发展了课程学习的思想，它的核心思想仍然是通过模拟人学习的方式来进行学习，不同的是，课程学习会事先对学习内容进行评估，把内容由难到易进行分类，然后再根据难易程度进行学习，先学习简单的，再学习复杂的；而自步学习取消了预先对内容进行难易分类这一步骤，在学习的过程中自动判断知识的难易程度，根据自己的水平和能力进行学习，先学习自己认为简单的知识，然后再学习自己认为困难的知识。在机器学习领域，自步

学习不同于课程学习在训练前确定样本的难易程度，而是在模型训练过程中自动评估样本的难度，自动决定学习样本的先后顺序，先学习简单的样本，再学习复杂的样本，最终得到效果好的模型。下面给出经典的自步学习模型介绍：

$$\min_{w,v} \sum_{i=1}^{n} v_i L(y_i, f(x_i, w)) + g(v_i, k) \tag{2.1}$$

其中，$L(y_i, f(x_i, w))$是损失函数；v_i是第i个样本的权重，该权重的大小决定了样本学习的先后顺序，权重大的样本首先放入模型进行训练；k是用于控制参与训练的样本数，当k值较大时，自步学习倾向于选择较少的样本用于训练，当k值较小时，自步学习倾向于选择更多的样本进行放入模型进行学习；$g(v_i, k)$是自步正则项。

常见的自步正则项有硬权重式的，公式如下所示：

$$g(v_i, k) = -\frac{1}{k}\sum_{i=1}^{n} v_i, v_i = \begin{cases} 0, l_i > \frac{1}{k} \\ 1, l_i \leq \frac{1}{k} \end{cases} \tag{2.2}$$

上述的自步正则项采用的是类似支持向量机中“硬间隔”的方式：若样本满足选择要求，则该样本的权重就是 1；若样本没有满足选择要求，则该样本的权重就是 0。上述方法是通过 0-1 来代表样本是否被选择，这样就可以避免将异常值放入模型进行训练，进而提高模型的稳健性。但上面的选择方式存在以下的问题：当一个样本介于 0 和 1 之间时，即该样本可能是一个与目标相关性较低的样本，但它又存在有利于模型训练的信息，这时将该样本的权重赋值成 0 和 1 都不太合适，此时，我们就考虑使用连续权重的方式，即软权重，在学习过程中，每个样本都会获得一个值，而这些值就可以细化地代表样本的潜在重要性。因此，采用软阈值权重替代硬阈值权重，这样更有利于模型的训练。基于此，Lu 等人就提出了软阈值权重的方式[38]，但该方法是基于线性的。为了同时保留软硬自步正则项的特性，Zhao 等人[39]使用了混合式自步正则项，公式如下所示：

$$g(v_i,k)=\sum_{i=1}^{n}\frac{\eta^2}{v_i+\eta k},v_i=\begin{cases}0,l_i>\dfrac{1}{k^2}\\1,l_i\le\left(\dfrac{\eta}{\eta k+1}\right)^2\\\eta\left(\dfrac{1}{\sqrt{l_i}}-k\right),otherwise\end{cases}\tag{2.3}$$

其中，η是一个区间控制参数，用于控制 0-1 之间的模糊间隔。

2.3 本章小结

本章主要由两部分组成，分别是特征选择和自步学习理论。第一部分介绍了本文的主要研究内容——特征选择，并详细介绍了过滤式特征选择方法、封装式特征选择方法和嵌入式特征选择方法；第二部分介绍了自步学习理论，先介绍了课程学习，再基于课程学习存在的缺点而引入对自步学习理论的介绍，并给出一些常用的自步正则项。

3.基于自步学习的半监督特征选择算法

本章将从模型构建和模型的优化及分析两个方面来介绍。在模型的构建部分，本文将详细阐述构建模型的思路并给出最终的目标函数；在模型的优化及分析部分，本文将具体介绍模型的优化算法，并对算法的复杂性进行分析。

3.1 模型构建

从上文研究背景中可以得知，当今社会各行各业的数据往往都是高维的，但在后续分析过程中这些高维数据会带来“维度灾难”等问题，给生产生活带来不便；同时，由于获取成本及人工标注等原因，在现实生活中往往很难得到完备而准确的有标签数据集。因此，本文将对半监督特征选择问题进行研究。由于半监督特征选择方法是由有监督特征选择方法和无监督特征选择方法发展而来，且有监督特征选择算法往往比无监督特征选择算法拥有更好的性能，因此，本文先来介绍经典的有监督特征选择方法。

在介绍传统的有监督特征选择方法前，我们首先对符号进行介绍。在本文中，我们用大写字母来表示矩阵，而数据集就可以看成是一个矩阵，因此，我们用大写字母X来表示数据集，即$X \in R^{d\times n}$，其中d代表数据集X共有d个特征，n代表数据集X共有n个样本，上述数据集X还可以细化成由n个样本和d个特征来表示，即$X = [x_1, x_2, \dots, x_n] = [w_1; w_2; \dots; w_d]$。同样的，我们通过大写字母$Y$来表示标签矩阵，即$Y = [y_1, y_2, \dots, y_n]^T \in R^{n\times c}$，其中$n$代表总的样本个数，$c$表示该数据集类别的个数，$y_i$表示对应样本$x_i$的类别标签。对于有监督的数据，传统的方法通常使用回归模型来学习一个低维的子空间，使得投影后的结果尽可能接近原始样本标签，即缩小预测标签值和实际标签值的差异，进而训练出相应的投影矩阵W，模型简单的定义如下所示：

$$\min_{W} L(f(X,W),Y) \tag{3.1}$$

上面公式中的L代表损失函数，简单来说，是通过一个函数来衡量预测值和真实值之间的差异，这里本文采用是 Frobenius 矩阵范数的平方。对一般的回归模型来说，为了保证模型的稳定和防止过拟合，会给模型添加一个正则化项，常见的正则化项有L_0范数，L_1范数，L_2范数等，本文采用是L_2范数。下面我们给出经典的有监督的回归模型[40]，具体的定义如下所示：

$$\min_{W} \|X^T W - Y\|_F^{\,2} + \beta \|W\|_F \tag{3.2}$$

其中，β是正则化项$||W||_F$的参数，可以通过改变β的大小来改变特征矩阵W的稀疏程度。

但是在实际生活中，获取有标签的数据往往十分困难，需要耗费大量的人力物力，虽然无标签的数据在训练模型方面不如有标签的数据有效，但无标签的数据本身也蕴含着大量的信息，可以通过探究其内在结构等方法去探寻与模型相关的信息。因此，无监督特征选择算法也被广泛研究。由于无监督特征选择算法缺少标签的信息，有的学者想到可以构造原始数据的伪标签来进行学习，即将标签矩阵Y用伪标签矩阵F来替代，还有的学者就希望充分挖掘数据本身的信息，通过特征的自表示来构建模型[41]。在本文中，我们采用自表示的方法来进行特征选择，并对其添加L_2正则化项，无监督特征选择方法如下所示：

$$\min_{W} \|X^T W - X^T\|_F^{\,2} + \beta \|W\|_F \tag{3.3}$$

对于半监督特征选择方法，之前是通过聚类等无监督学习算法去预测无标签样本的标签信息，进而再把其转换为有监督特征选择方法。显然的，这种方法操作比较麻烦，需要额外进行训练获取数据的标签信息。现在比较常见的方法是类似于无监督特征学习中构建伪标签矩阵的方式，半监督特征选择方法将有标签和无标签的样本同时纳入一个伪标签矩阵，并给予有标签的样本较高的权重，但它将预测信息放入模型进行训练，模型结果可能会受到标签预测错误样本的影响，进而降低模型的效果。在本文中，我们为了同时利用有标签的数据和无标签的数据，将上文提到有监督模型和无监督模型进行组合[42]；在组合的过程中，我们考虑到有监督模型和无监督模型的重要程度是不一样的，基于这一点，我们给这两个模型施加不同的权重，以获得最

好的模型效果，这样就可以得到以下的模型：

$$\min_{W_u,W_l,v} \left[(1-\theta)\|X^T W_u - X^T\|_F{}^2 + \theta\|X^T W_l - Y\|_F{}^2\right] + \beta\left(\|W_u\|_F{}^2 + \left||W_l|\right|_F^2\right) \tag{3.4}$$

其中，θ是权重参数。

上述模型可以筛选出重要的特征，但原始数据中往往存在异常值和噪声，而在传统的特征选择中，会将所有的样本放入模型进行训练，这样就会影响模型的效果。同时，不同样本对目标函数的影响不同，因此在训练模型时，对样本的选择也尤为重要。自步学习可以在模型训练过程中自动且有序地对样本的难度进行评估并选择样本进行学习，首先会选择它认为重要的样本来构建初始模型，然后再选择次重要的样本来构建模型，以此来提高模型的泛化能力[43]。因此，我们将混合式自步学习融入到特征选择中，可以得到以下的目标函数：

$$\min_{W_u,W_l,v} \left[(1-\theta)\sum_{t=1}^{u} v_t \|x_t - x_t W_u\|_F{}^2 + \theta\sum_{j=1}^{l} v_j \left\|y_j - x_j W_l\right\|_F{}^2\right] + \sum_{i=1}^{n}\frac{\eta^2}{v_i+\eta k} + \beta\left(\|W_u\|_F{}^2 + \left||W_l|\right|_F^2\right) \tag{3.5}$$

$$s.t. 0 \le v_i \le 1, i = 1,2,\dots,n$$

其中，v_i是第i个样本的权重，该权重的大小决定了样本学习的先后顺序；k是用于控制参与训练的样本数；η是一个区间控制参数，用于控制 0-1 之间的模糊间隔。

在特征选择方法的发展过程中，许多判定特征重要性的方法被构建，但在特征选择过程中还存在一个重要的问题，即特征的冗余性问题。如果仅依据特征的重要性进行特征选择，就可能选出高度相似的特征。在后续操作中，我们通常会按照特定的比例或者特定的数量来选择特征放入后续的模型，如果不对重复且重要的特征进行去重，相似重要的特征就会占据最优特征子集，却不能带来模型效果的提升，同时让其他的特征无法进入最优特征子集，会损失一部分信息，降低模型效果。基于此，Liu 等人提出了一个解决特征冗余性的方法，即通过考虑特征之间的成对相似性来解决特征之间的冗余问题[44][45]。首先，他们通过向量乘积的思想来描述不同特征之间的相似性，基于

原始数据集X定义了一个相似性矩阵S，即$S = XX^T$，s_{ij}的值越大，意味着第i个特征和第j个特征越相似，而对于两个相似的特征，只想选取一个放入最优特征子集。接着，通过投影矩阵W的行和来定义每个特征的重要性，并通过最小化$tr(S^T W1W^T)$来解决特征的冗余性，其主要思想是如果s_{ij}的值比较大，那第i个特征和第j个特征不太可能同时被选择，这样就可以很好的解决特征冗余的问题。在本文中，我们也借鉴该思想来解决特征的冗余性问题，最后就得到了如下所示的目标函数：

$$\begin{aligned}\min_{W_u,W_l,v} &\left[(1-\theta)\sum_{t=1}^{u} v_t \left\|x_t - x_t W_u\right\|_F^2 + \theta \sum_{j=1}^{l} v_j \left\|y_j - x_j W_l\right\|_F^2\right] + \\ &+ \sum_{i=1}^{n} \frac{\eta^2}{v_i + \eta k} + \beta\left(\left\|W_u\right\|_F^2 + \left||W_l|\right|_F^2\right) + \alpha tr(S^T W1W^T) \\ &s.t. 0 \le v_i \le 1, i = 1,2,\dots,n\end{aligned} \tag{3.6}$$

其中，$W = [W_u, W_l]$。

3.2 模型的优化及分析

在上一节中，我们介绍了基于自步学习的半监督特征选择算法的构建过程。在这一节中，我们将介绍该模型的优化过程，交代算法中的每一个变量是如何求解的；还将给出模型的复杂度分析，用来评估模型的性能。

3.2.1 优化算法

本文提出的基于自步学习的半监督特征选择算法一共有三个需要求解的变量，分别是W_u、W_l和v。W_u是无监督特征选择方法中的投影矩阵；W_l是有监督特征选择方法中的投影矩阵；v是表示每个样本权重的向量，其取值范围在 0~1 之间，数值越接近 1，代表该样本越重要；数值越接近 0，代表该样本越不重要。由于本文同时将无监督特征选择方法和有监督特征选择方法纳入模型，对于衡量样本重要性的变量v，也需要分别对有标签样本和无标签的样本进行求解。

由于本文的目标函数是一个非凸问题，不可以对其进行直接求解，因此，

本文采用交替迭代更新法来求解相应的变量。交替迭代更新法是在更新某一变量时，固定其他变量不变，即把其他变量当作常量。在更新时，目标函数中仅有这次需要更新的变量是“变量”，对其进行求解。按照上述步骤依次对需要求解的变量进行更新，直到目标函数值收敛，得到各个变量的结果。这样就可以把一个非凸问题拆解成多个凸问题进行求解。比如说，在本文中，我们就可以分三个部分对本文提出的算法进行优化。下面给出具体的更新步骤：

（1）固定其他变量，更新W_u

当W_l和v固定时，上文的目标函数（3.6）可以改写成以下形式，即保留与W_u相关的部分，可以将与W_l和v相关与W_u无关的部分看成是常量：

$$\min_{W_u}(1-\theta)\sum_{t=1}^{u} v_t \|x_t - x_t W_u\|_F^2 + \beta\|W_u\|_F^2 + \alpha tr(S^T W_u 1 W_u^T) \tag{3.7}$$

为了方便进行求解，我们将上述式子转化成矩阵的形式，令$U = diag(\sqrt{v}), G = UX$，可以得到以下的式子：

$$\min_{W_u}(1-\theta)\|G - GW_u\|_F^2 + \beta\|W_u\|_F^2 + \alpha tr(S^T W_u 1 W_u^T) \tag{3.8}$$

为了求解上述最小化问题，我们对上述式子进行求导，并令求导之后的结果为 0，可以得到以下方程：

$$2(1-\theta)G^T(GW_u - G) + 2\beta W_u + 2\alpha S W_u I = 0 \tag{3.9}$$

经过化简移项，可以得到最后的结果：

$$W_u = ((1-\theta)G^T G + \beta I + \alpha S)^{-1}(1-\theta)G^T G \tag{3.10}$$

（2）固定其他变量，更新W_l

当W_{un}和v固定时，上文的目标函数（3.6）可以改写成以下形式，即保留与W_l相关的部分，可以将与W_{un}和v相关与W_l无关的部分看成是常量：

$$\min_{W_l}\theta\sum_{j=1}^{l} v_j \left\|y_j - x_j W_l\right\|_F^2 + \beta\left||W_l|\right|_F^2 + \alpha tr(S^T W_l 1 W_l^T) \tag{3.11}$$

为了方便求解，我们将上述式子转化成矩阵的形式，令$U = diag(\sqrt{v})$, $G = UX$，$Q = UY$，可以得到以下的式子：

$$\min_{W_{un}}\theta\|Q - GW_l\|_F^2 + \beta\|W_l\|_F^2 + \alpha tr(S^T W_l 1 W_l^T) \tag{3.12}$$

为了求解上述最小化问题，我们对上述式子进行求导，并令求导之后的