# 绪论

## 选题背景

从二十一世纪开始，科技高速发展，伴随着人们在科技上的进步，人们在网络上留下的数据也逐渐增大。面对如此庞大的数据，如何从这些大量冗余的信息中提取出对于我们来说有价值的知识，如何通过这些数据给用户提供更精确更全面的服务，是我们一直在追求的目标。与此同时，随着人工智能这一名词的兴起，其理论基础也逐渐发展，作为人工智能的核心之一，机器学习(Machine Learning)也成为众多学者们研究的课题。瞩目于机器学习领域，集成学习(ensemble Learning)不仅是机器学习的重点研究方向之一，也在实践领域中得到了广泛的应用。集成学习是一种构建并将多个弱学习器结合在一起来完成任务，从而起到强学习器作用的一种提高模型泛化能力的方法，有时也被成为多分类器系统等。如何将这些弱学习器结合在一起，使得其得到更好的效果，这其中的策略就是学者们需要不断研究的方向之一，从集成学习这一概念出现开始，就有无数学者提出了不同的结合策略。

目前在聚类的领域，有许多经典的聚类算法被不断提出，包括基于划分的聚类：K均值聚类(K-Means)、围绕中心点的划分(PAM)、K众数聚类(K-Mode)、K原型聚类(K-prototype)；基于层次的聚类：凝聚的层次聚类(AGNES)、分裂的层次聚类(DIANA)；基于密度的聚类(DBSCAN)；基于网格的多分辨率聚类(STING)，基于模型的神经网络聚类(SOM)等。但是，由于现实数据的不确定性和难分辨性，仅仅使用单一的聚类方法并不能得到很好的效果。因此，研究者们近年来将精力不断转移到了多个学习器集成的方法研究上，旨在提高学习器的精度。时至今日，集成学习的研究热度仍是只增不减。就目前为止许多集成学习算法被提出且在应用中取得了不错的效果，并被广泛应用于字符识别、人脸识别、图像分类等领域。由此可见，集成学习的研究具有深刻的理论意义和极高的应用价值。

在集成学习理论框架上，系统的预测性能大多是通过增加基学习器的个数来实现的，但是这样会带来一定的负面影响，例如算法运行时间的增加、系统使用空间的变大等；更有可能因为数据的噪声过多，导致其基学习器均不能有好的学习结果。为了更好地解决这些问题，研究者们引入了动态规划，即选择性集成的概念，在众多学习器中有选择地集成一部分学习器，使得集成系统可以达到更好的效果。

综上所述，要在集成系统训练的时候使各个学习器“好而不同”，如何提升集成学习系统的差异性、提高学习的精度、改善运行的效率，仍就是集成学习领域中备受关注但仍未解决的问题。

## 研究意义

随着大数据概念的出现，人工智能逐渐成为现在研究的主流方向。机器学习作为人工智能领域中重要的一块，引发了学者们孜孜不倦地研究。作为机器学习的重点研究课题——集成学习，其相关的研究非常丰富。

学习器的结合可以带来三个方面的好处：第一，使用单个学习器可能会因误选而导致学习器的泛化性能差，结合使用多个学习器可能可以减小该风险；第二，学习器算法可能会陷入局部极小的情况，而该点所对应的泛化能力可能不佳，集合多个学习器可以降低陷入糟糕的局部极小点的风险；第三，结合多个学习器可以扩大相应的假设空间，从而学习到更好的近似结果。因此，对于如何进行集成这一问题，无数学者进行了多方面的研究，同时也发现了许多的结合策略。

在现实生活中，我们能获得的数据可以分为两种：有标签的和无标签的。通常我们将对有标签的数据进行分类，这个被称作监督学习；将无标签的数据进行聚类，这个被称作是无监督学习。但是，现实中的数据总是多样的，我们可能无法简单地从其形状或是特征判断出哪种算法在这个数据上可以达到非常好的效果；其次，现实中的数据可能会有很严重的噪声现象，即有相当多的噪声数据以及离群数据，这会导致我们在进行单一聚类的时候，出现聚类不理想的情况。

基于这一研究情况，论文望引入动态规划的概念，对聚类集成框架上进行有选择的集成，达到尽可能消除数据噪声影响、且使集成系统的差异性和精度尽可能提升的目的。

## 国内外研究现状

Perrone和Cooper[1]于1993年首次在集成学习中应用了平均法的结合策略，它在集成学习中具有特别的意义，加权平均法可以被认为是集成学习研究的基础。一种常见的结合策略——投票法，多数使用于在1996年Breiman[2]提出的Bagging算法框架中。当训练数据很多的时候，一种更为强大的结合策略——学习法Stacking就此出现，它在普通的集成学习的框架上将另外的学习器，也称为元学习器，作为结合策略，且将其他弱学习器称为初级学习器，1992年Wolpert[3]和1996年Breiman[4]提出的学习法是典型代表。

和以上几种将所有初级学习器训练完再进行结合的静态结合策略不同的是，2006年Soares等人[5]提出了基于差异性(KNN and selection method)和准确度(Clustering and selection method)的动态集成方法，前一种方法使用KNN的思想，找到待分类样本的最邻近的k个测试样本，然后基于这k个测试样本对于初级分类器，进行分类，最后找到分类精确度高、差异性大的几个分类器作为待分类样本的初级分类器；后一种方法使用聚类的思想将验证集划分成不同的簇，找到距离待测样本最近的簇，然后从分类器中选取对于该簇分类精度最高的几个分类器中差异性最大的几个作为初级分类器。动态选择集成通过在测试集训练时对分类样本的领域的分类准确度排序，选择出最适该待分类样本领域的分类器，从而在最初阶段就可以减少无效（或是说分类结果不好）的分类器分类所占用的时间和空间，达到“剪枝“的目的。2010年，Woloszynski等人[6]提出了一种基于RRC(Randomized Reference Classifier)的动态集成选择算法DES\_RRC，他们先计算所有基于RRC的分类器分类能力，然后再选取高于随机分类器分类正确的概率的分类器，最后使用多数投票法结合。2012年，Woloszynski等人[7]又提出一种新的分类器度量方法MCR，即从基分类器集合中选择比随机分类效果其更好的分类器子集，去除掉对集成系统起负面影响的分类器，并在该度量方法上提出了两种集成系统：DES-P和DES\_KL。

在集成学习中，针对使用的数据是否有标签，又可以为分类集成和聚类集成。

基于动态集成的理论的提出，朱雪[8]针对动态集成选择算法对动态选择数据集分布敏感，尤其是动态选择数据集中含有噪声、高度重叠的样本实例分类效果不理想、能力区域不能充分代表待测样本的问题，引进了被编辑最近邻和自适应最近邻两种选择机制，提出了改进的动态集成选择算法。在应用上，童珂凡等人[9]利用动态集成分类器的方法，根据基分类器在北侧样本邻域内的表现选择满足 要求的分类器来集成，并在烟卷感官质量的预测上得到了较好的效果。.Cheriguene等人[10]在青光眼诊断数据中，采用多样性度量和随机子空间算法组成的继承分类器进行视网膜图像的分类，取得了显著的效果。

同时，在聚类集成算法的研究上，宋敬环[11]采用模糊聚类算法作为遗传共识策略中的基聚类器，引入自适应因子，提出了一种可以提高集成速度的新的遗传共识聚类集成算法。徐丽[12]利用监督信息，研究了一种结合传统的基于随机子空间的聚类集成算法而设计出的一种基于约束选择的加权随机子空间聚类降级成算法。谷鹏花[13]基于选择性聚类聚成的理论，提出了一种基于数据挂链的聚类集成方法(DRBCE)，先提取出在聚类成员中体现出关联的数据对象组成新的类，然后在对这些类再次聚类的到最终的集成结果。不仅仅是在理论上的发展，聚类集成在实际应用中也有很广泛的应用，在进行交叉口车量行驶路径的提取上，尹卓等人[14]提出一种基于聚类集成的路径信息提取方法，且有效地提高了路径提取的精度。

参考文献：

[1] Perrone, Michael & Cooper, Leon. (1993). When Networks Disagree: Ensemble Methods for Hybrid Neural Networks. Neural networks for speech and image processing. 10.1142/9789812795885\_0025.

[2] Breiman, L. Bagging predictors. Mach Learn 24, 123–140 (1996).

[3] Wolpert, D., Stacked Generalization, Neural Networks, Vol. 5, 1992, pp. 241–259.

[4] Breiman, L. Stacked Regressions. Machine Learning 24, 49–64 (1996).

[5] A. Santana, R.G.F. Soares, A.M.P. Canuto, M.C.P. de Souto, A dynamic classifier selection method to build ensembles using accuracy and diversity, in: 2006 Ninth Brazilian Symposium on Neural Networks, SBRN ’06, 2006, pp. 36–41.

[6] M. Kurzynski, T. Woloszynski, R. Lysiak, On two measures of classifier competence for dynamic ensemble selection—experimental comparative analysis, in: 2010 International Symposium on Communications and Information Technologies (ISCIT), 2010, pp. 1108–1113.

[7] Tomasz Woloszynski, Marek Kurzynski, Pawel Podsiadlo, Gwidon W. Stachowiak,A measure of competence based on random classification for dynamic ensemble selection,Information Fusion,Volume 13, Issue 3,2012,Pages 207-213,ISSN 1566-2535.

[8] 朱雪. 动态集成分类方法的研究[D].大连海事大学,2019.

[9] 童珂凡,张忠良,雒兴刚,曾鸣,汤建国.基于动态分类器集成系统的卷烟感官质量预测方法[J].计算机应用与软件,2020,37(01):66-70+81.

[10] Cheriguene S., Azizi N., Dey N. (2016) Ensemble Classifiers Construction Using Diversity Measures and Random Subspace Algorithm Combination: Application to Glaucoma Diagnosis. In: Dey N., Bhateja V., Hassanien A. (eds) Medical Imaging in Clinical Applications. Studies in Computational Intelligence, vol 651. Springer, Cham.

[11] 宋敬环. 聚类集成算法研究[D].哈尔滨工程大学,2015.

[12] 徐丽. 基于粒度计算的聚类集成算法研究[D].中国矿业大学,2018.

[13] 谷鹏花. 聚类集成及差异性的研究[D].西南交通大学,2012.

[14] 尹卓,许甜,陈阳舟,卢佳程.基于集成聚类的交叉口车辆行驶路径提取方法[J].计算机应用与软件,2020,37(07):180-187+193.