思想在于“Many Could be Better Than All”，即对于大量的个体学习器的效果，找出部分性能良好的学习器比全部的学习器组合集成更加有效，因此可以基于某种度量标准进行选择后再集成，以期实现更好的性能。

1.3 标签噪声的研究意义及研究现状

一些针对标签噪声的文献[9][10]表明，标签噪声对于分类器的预测性能的影响是难以忽略的，甚至训练样本集中小部分标签噪声的引入也会对分类器的决策产生较大大的偏差和失真。常用机器学习算法，如感知器、K 近邻算法、决策树算法和支持向量机等，分类性能都会受到标签噪声的影响。K 近邻算法中，特别是当 K 值取 1 时，一个样本标签的错误标记可能会使得周围邻近的一些样本的错误分类。相对来说，由于支持向量机的良好性能取决于精确的决策面的构建，如果有一个距离分类决策面较远的样本的类别标记错误，也会使得决策面发生位移和扭曲，分类器的预测性能更容易受其影响。

标签噪声的存在也会导致训练样本数目的增加和学习模型复杂度的变大。Quinlan [11]和 Abellán and Masegosa[12]发现，当训练数据集中有标签噪声存在时，决策树的尺寸会相应地增加，其结点数目相对较多，使得训练得到的模型异常复杂。

同样地，集成学习也受到标签噪声的影响。Dietterich[13]通过实验发现，Bagging 和 Boosting 两种常用的集成方法在处理有标签噪声的数据集时，展现了不同的效果。数据集中随机加入少量的噪声样本，Bagging 方法中基分类器的差异性会随之增加，最终构建的学习模型具有更好的分类性能。相对来说，Boosting 方法更容易受到标签噪声的负面影响，特别是 AdaBoost 算法，随着迭代次数的增加，由于算法会更多地关注于错分类的样本，必然会使得噪声样本的权值越来越大，进而增加了模型复杂度，降低了算法性能。

因此非常必要针对标签噪声的处理进行相关的研究。目前标签噪声主要有两方面的处理办法：一种对标签噪声鲁棒的算法；另一种方法尝试滤除噪声样本来改善训练数据的质量。前者的有效性来源于分类器的学习不会受到标签噪声敏感的影响；而后者则是在训练发生前处理含噪的样本，查找到噪声样本后或者重标记，或者进行简单的剔除。一般来说，剔除噪声样本的方法简单有效且容易执行，但更容易去除过多的样本，造成训练集信息的丢失。

针对 AdaBoost 算法的标签噪声敏感性，很多相关的改进鲁棒算法被提出。Domingo 和 Watanabe[23]于 2000 年提出了 MadaBoost 算法，通过调整权值更新公式，对样本权值设置阈值上限，减缓样本权值的过度增加，避免算法对噪声样本的过拟合。2008 年，Bradley and Schapire 提出了 FilterBoost 算法，算法采用了对数损失函数代替

了指数损失函数，与 MAdaBoost 类似，放慢了样本权值的增加速度。2012 年，Manwani 和 Sastry[22]两人从风险损失函数入手，证明了预测模型鲁棒性与标签噪声的关系，发现 0-1 损失函数和最小均方损失函数建立的模型鲁棒性较好，而指数损失函数、对数损失函数和合页损失函数鲁棒性较差，对应于 AdaBoost、逻辑回归（Logistic Regression）和支持向量机（SVM）算法。

当训练数据集中错误地添加了标签噪声，首先应该对训练数据集进行清理，查找出错分类的样本并将其剔除，这种思想和异常点检测是一致的。噪声样本的剔除仍依赖于训练数据集中样本之间的相关性。

Brodley 和 Friedl[24][25]提出了集成去噪的方法，又称投票去噪，在 K 折交叉验证的基础上，通过集成多分类器的预测结果投票选择出噪声样本。投票去噪方法主要有多数投票和一致投票两种方法，多数投票方法强调，集成多个分类器预测结果时，超过半数的分类器错分的样本才被视为标签噪声样本，而一致投票则是要求所有的分类器错分的样本才是标签噪声样本。相对来说，一致投票比多数投票更加保守，多数投票更容易剔除较多的所谓的“标签噪声样本”。

2003 年，Zhu 和 Wu 等人[35][36]提出数据集分块去噪的概念，然而只适用于相对较大的数据集。训练数据集被划分为多个数据子集，一个子集中的标签噪声样本由剩余的多个子集训练得到的基分类器集成投票决定。

此外，还有基于实例选择的方法可以用于标签噪声样本的剔除。Wilson[37]提出了 Edited Nearest Neighbors 的方法，该方法在 K 近邻方法的基础上，对于每个样本，找出其距离最近的 K 个样本，若该样本与多数类别标签不一致，则被视为噪声样本。

1.4 论文研究内容与结构安排

本文的主要研究方向为通过集成学习方法解决数据集中不可避免的标签噪声问题。 本文共包含 6 个章节的内容，以下为详细的结构安排：

第一章简单分析了现实数据中常遇到的标签噪声的研究意义及研究现状，紧接着讲解了集成学习算法的研究现状，最后再对论文的框架安排作出了简要的介绍。

第二章详细地介绍了集成学习算法方面的内容，其中着重讲解了集成学习方法中三种最重要的算法——Bagging、随机森林和 AdaBoost 算法，对两种算法的理论原理进行了分析，最后分析了集成学习在标签噪声中的应用。

第三章主要讲解了基于 Condensed Nearest Neighbors 和集成的标签噪声鲁棒算法，算法利用 Condensed Nearest Neighbors 对于标签噪声的敏感性和基于实例选择的特性，将其改进为训练子集的提取采样方法，不仅训练子集中标签噪声样本比例相对较少，而且可以得到多个具有差异性的训练数据子集，再通过训练得到的基分类器的集成，