华中科技大学

博士学位论文

特征选择及半监督分类方法研究

姓名：黄东山

申请学位级别：博士

专业：计算机应用技术

指导教师：宋恩民

2011-05-23

摘 要

特征选择和半监督分类是缓解“高维小样本”问题的有效方法，在统计学、数据挖掘、机器学习、模式识别、生物信息学等领域取得了丰硕的研究成果，特别是在数据挖掘和机器学习领域，特征选择和半监督分类是近几年来的研究热点，具有十分重要的理论研究价值和实际应用价值。

特征选择和半监督分类主要存在以下几个问题：（1）在选取表征能力强的特征子集时，很多特征选择方法选择类别识别能力靠前的 k 个特征，但这样组成的特征子集并不一定具有很强的类别区分能力；（2）标准协同训练方法要求特征集能构成充分冗余的两个视图，即两个特征子集相互独立且能独自训练分类器。这个条件在很多情况下特别是在面临“高维小样本”时难以满足；（3）集成多分类器类型的半监督分类方法提升“小样本”背景下弱分类器的性能时，对基分类器的要求较高，但基分类器的分类性能因标记样本的不足往往不高，进而造成集成后的分类器总体性能提升不明显；（4）当前已有研究者将流形学习方法与半监督学习结合起来，利用大量未标记样本蕴含的几何结构信息来设计高精度的分类算法。但通常这类半监督方法不仅复杂且参数调节较为繁琐。

针对“高维小样本”中特征选择及半监督分类存在的问题，在如下 4 个方面进行了研究，相关研究及主要成果有：

（1）对单特征类别区分能力的评价扩展到对特征子集的类别区分能力评价，并结合“最好优先”搜索策略，给出了一种能直接选取具有强类别区分能力的特征子集选取方法 FSCRF。实验结果表明，FSCRF 能在大多数情况下有效地选出特征数目更少、分类精度更高的特征子集。在此基础上，将该方法应用在老年痴呆诊断方面，同样取得了令人满意的结果。

（2）分析了现有协同训练方法存在的一些问题，给出了一种新的交叉训练半监督分类方法 NC-T。NC-T 将标记样本划分成三份，并利用三个基分类器对其进行训练。它不需要假设数据特征存在两个或多个独立特征视图，相比标准协同训练，每

个分类器训练的标记样本为 2/3 而不是 1/2，对标记样本利用更充分。实验结果显示，NC-T 方法的分类精度对标准协同方法在多数情况有所提高。

（3）为了能有效降低对基分类器的要求，通过融合大量未标记样本信息，给出了一种多类别多分类器集成半监督分类方法 SSMAB。SSMAB 只需要基分类器的分类精度达到$1/K$（K 为类别数目），就能取得较满意的效果。实验结果表明，在分类精度上，SSMAB 与同类型的方法在多数情况下占有优势。

（4）由于非公度距离度量对数据之间距离关系的度量更加合理，给出了一种非公度的半监督学习方法 NMSNN，NMSNN 定义了一个代价长度函数用以度量数据点间的距离，这种距离度量不仅考虑了数据点间的直接关系，而且考虑了全局关系，有效地利用了未标记样本的几何结构信息和标记样本的类别信息。NMSNN 只需要设置一个参数，使得该方法更加简便实用。实验结果表明，NMSNN 方法具有良好的分类精度和稳定性。

关键词：特征选择，半监督分类，协同训练，集成学习，非公度度量，高维小样本，决策树，神经网络

Abstract

Feature selection and semi-supervised classification are effective methods that can alleviate the problem of low sample data size in high-dimensional space. These methods have achieved fruitful research results in statistics, data mining, machine learning, pattern recognition, bioinformatics and other fields. Especially, feature selection and semi-supervised classification have become a current research focus with important theoretic and practical value in the fields of data mining and machine learning.

Some feature selection and semi-supervised classification problems are as follows: (1) for most feature selection methods, all features are ranked according to certain evaluation scores, then the final feature subset is constructed by selecting top ranked features subject to a empirical threshold; although each selected feature has a better discrimination capability than any of the unselected features, the feature subset constructed in this way may not possess the best discrimination capability; (2) the standard co-training algorithm requires two sufficiently redundant views, i.e., the featrues can be naturally divided into two independent sets; unfortunately, this assumption would fail in most practical learning scenarios, especially for low sample data size in high-dimensional space; (3) the existing semi-supervised classification methods that assemble multiple base classification algorithms are used to improve the performance of weak base learning algorithm in low sample size settings; however, they require base learning classifiers with high accuracy, which is difficult to satisfy in the low sample size settings; therefore, the performances of there methods are usually poor; (4) currently, many researchers combine semi-supervised learning methods with manifold learning, trying to use a lot of geometrical information structural information of unlabeled data to design high-precision classification algorithm; but generally these types of semi-supervised methods are complex, and their parameter adjustment methods are complicated.

In order to solve the existing problems of feature selection and semi-supervised classification for low sample data size in high-dimensional space. The main contributions of this dissertation are summarized as follows.

Firstly, a new feature subset evaluation method which estimates the discrimination

capability of candidate features in low-dimensional feature subspace is proposed. By combining the new evaluation method with the best-first search strategy, a new filter method for feature subset selection FSCRF is developed. Experimental results demonstrate that not only the method is able to select a feature subset with smaller number of features for most data sets, but also the performance of classification is good in most cases. FSCRF also achieves satisfactory results on Alzheimer's disease truth data.

Secondly, the deficiency of standard co-training is discussed, and a new cross-training based learning algorithm NC-T is proposed. This algorithm generates three classifiers based on the three subsets of original labeled and unlabeled training set. The raw data are randomly divided into three subsets with similar size which include both labeled and unlabeled data. NC-T does not need to assume sufficient and redundant views and different supervised learning methods. Compared with the co-training, the base classifier of NC-T is trained on 2/3 labeled samples instead of 1/2. Experimental results show that the NC-T algorithm can improve classification accuracy in most cases.

Thirdly, the existing semi-supervised assemble classification methods require the based classifiers of high accuracy. In order to effectively reduce the requirements to the base classifiers, the SSMAB using multi-class Adaboost is proposed, which can achieve high classification accuracy by exploiting the unlabeled data. The required classification accuracy to each base classifier is only $1/K$ (K is the number of classes).

Fourthly, since the non-metric measure is a more reasonable distance measure between samples, a new cost length of path as a non-metric measure is introduced to measure the affinity between two samples, then the non-metric measure based NMSNN method is proposed. It considers both the direct relationship between two samples and the global information from other samples, to make use of the structure information of unlabeled and labeled samples effectively. The proposed method is simple and practical because there is only one parameter needed to be adjusted. The presented experimental results suggest that the NMSNNis able to use unlabeled data effectively in most cases.

Key words: Feature Selection, Semi-Supervised Classification, Co-training, Ensemble Learning, Non-metric Measure, High-dimensional and Low-sample, Decision Trees, Neural Network

附录 3 英文缩写名词

英文缩写	英文全称	中文译名
BBFS	Branch and Bound Feature Selection	分支界定法特征选择
CA	Cluster Assumption	聚类假设
Co-T	Co-Training	协同训练
FSCRF	Feature Subset Category Resolve Capability based Feature Selection	基于子集类别类别区分能力的特征选择
GMM	Gaussian Mixture Model	混合高斯模型
HMT	Harmonic Mixture Training	调和混合训练
I.I.D	Idependently and Identically Distributed	独立同分布
ICT	Cross-Training	交叉训练
ISOMAP	Isometric Feature Mapping	等距特征映射法
LapSVM	Laplacian Support Vector Machines	拉普拉斯支持向量机
LLE	Local Linear Embedding	局部线性嵌入法
LLGC	Learning with Local and Global Consistency	局部与全局一致性学习
LNP	Linear Neighborhood Propagation	线性近邻传播
LVF	Las Vegas Filter	拉斯维加斯过滤法
MA	Manifold Assumption	流形假设
MissSVM	Multi-Instance Learning by Semi-Supervised Support Vector Machine	基于半监督支持向量机的多实例学习
MMM	Multimodal Mixture Model	多模态混合模型
NC-T	New Cross-Training	新交叉训练
NFL	No Free Lunch	无免费午餐定理
NMSNN	Non-metric based Semi-supervised Nearest Neighbor	非公度半监督近邻分类
PAC	Probably Approximately Correct	概率近似正确
RASCO	Random Sub-space Method for Co-training	随机子空间协同训练
RBFNN	Radial Basis Function Neural Networks	径向基神经网络
S3VM	Semi-supervised Support Vector Machine	半监督支持向量机
SA	Simulated Annealing	模拟退火法
SA	Smoothness Assumption	平滑假设
SBS	Sequential Back Selection	序贯后向选择
SDP	Semi-Definite Programming	半定规划
Se-T	Self-Training	自训练
SFS	Sequential Forward Selection	序贯前向选择
SGT	Spectral Graph Transducer	谱图传导
SSL	Semi-Supervised Learning	半监督学习
SSMAB	Semi-supervised Multi class AdaBoost	半监督多类 AdaBoost
SVM-RFE	Support Vector Machine-Recursive Feature Elimination	支持向量机迭代特征剔除法
TFIDF	Term Frequency Inverse Document Frequency	词频-逆文档频率
TSVM	Transductive Support Vector Machine	直推式支持向量机
USGGM	Unsupervised/Supervised Generalizable Gaussian MixtureModel	无监督/有监督一般化混合高斯模型

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到，本声明的法律结果由本人承担。

学位论文作者签名：

日期：　　年　月　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本论文属于　保　密□，在_____年解密后适用本授权书。
　　不保密□。

（请在以上方框内打“√”）

学位论文作者签名：　　指导教师签名：

日期：　　年　月　日　　日期：　　年　月　日

1 绪论

1.1 研究背景

在许多复杂的数据分析问题中，机器学习和数据挖掘都面临着一个亟需解决的问题——“高维小样本”问题。“高维小样本”即数据样本特征维数很高，而有监督信息（如类别）的数据样本却很少。在数据类别多的情况下，“高维小样本”问题更加突出。因此，在模式分类过程中，为了更加有效地进行特征提取和分类，训练样本的数量必须满足一定的要求才有可能获得较好的结果。假定各类别的训练样本为n，数据样本特征向量的维数为d，则根据 Raudys[1]关于小样本问题的讨论和 Fukunaga[2]关于训练样本数目对分类器影响的论述可知，训练样本数量n应该尽可能地满足$n=\alpha\cdot d$，其中α的典型取值为 2、5、10 等。

其实，高维数据在实际应用中相当普遍，如基于基因的疾病诊断[3, 4]（往往需要对成千上万基因进行并行处理）、医学数据检测[5, 6]、网页分类[7]、图像分类[8, 9]、人脸识别[10]、语音识别[11]等过程中，都涉及到高维数据的处理。随着数据维度的急剧增加，大量的冗余信息和无关信息通常也会随之产生，这些信息可能极大地降低机器学习算法的学习性能。因此在处理高维数据时，特征选择对于机器学习显得尤为必要。

此外，对于数据分析问题，获取有用的监督信息（或称人工标记信息，有时也指类别标记）需要做大量繁琐的工作，耗费大量的时间、人力。如进行网页推荐时，需用户标记哪些网页是感兴趣的，一般很少有用户愿意花大量时间来提供标记；在医学影像处理中，很容易从医院获得大量的医学影像，但对于病灶的标定需耗费医学专家的很多时间和精力，因此只有少量病例得到了医学专家的标定，而大量病例没有标记。在使用传统的机器学习方法来处理这些分类问题时，就会出现因标记样本过少而分类精度不高的情况。

1.2 研究目的及意义

对于“高维小样本”问题，降低特征维度和进行半监督学习是解决这个问题的有效途径之一，特征选择是降低特征维度的一种重要方式，研究特征选择和半监督分类的目的是希望从数据预处理和分类方法两方面解决“高维小样本”问题，即，（1）设计出能直接从高维特征中选择出具有强表征能力的特征子集算法，尽可能的去除数据噪声、降低特征之间的冗余；（2）利用少量标记样本和大量未标记样本，设计具有良好分类性能的分类算法，解决标记样本不足所带来的分类精度不高问题。

降低特征维度的算法通常分成两大类，第一类是特征抽取：通常是对高维特征空间按照某一准则向低维特征空间进行转换。但由于经过特征抽取算法变换后的新特征均不包含对原始特征的具体含义，因此，这类降维算法对需要解释特征意义的问题（如医学诊断问题）没有太多的指导价值。第二类方法是特征选择：从原始的特征空间中剔除没有作用的特征，选择重要的特征。这类方法将保留原始特征的具体含义，容易理解和识别，对实际应用具有很好的指导作用。但无论是特征抽取还是特征选择，在把高维数据降低为低维数据时都会一定程度地丢弃有用信息。过多的选取特征或者保留维数将会带来不必要的计算，甚至会夹杂噪声，但过低的估计又会造成重要信息的损失。因此，如何寻找高维数据的“本征维数”对降低数据的维数及其后续处理起到至关重要的作用。

半监督学习（SSL，Semi-supervised Learning）[12]是缓解“高维小样本”问题中小样本问题的有效方法，也是近年来模式识别和机器学习的重点研究领域。它主要考虑如何同时利用少量的标记样本和大量的未标记样本进行分类学习的问题。与传统的有监督方法相比，它能有效地利用大量未标记样本的信息来改善有监督方法的分类精度，缓解“小样本”存在的问题。与完全没有任何监督知识的无监督学习相比，它能利用少量具有指导价值的信息来引导大量未标记信息进行聚类。因此，半监督学习在基因数据分析[13-16]、图像分割[17-20]（聚类问题）等应用领域已有了广泛的应用。半监督学习对于减少标注代价，提高学习机器性能具有非常重大的实际意义。此外，Zhu 等人[21]研究发现，人的认知方式与半监督学习模型有很多类似之处，

因此研究半监督机器学习理论对探索人的认知模式具有重要的参考价值。

“高维小样本”问题涉及到统计学、模式识别、数据挖掘、机器学习等多个学科，是一个极富挑战性的课题。而特征选择和半监督学习方法是解决这个问题的有效途径之一，并已被广泛用于解决许多具体的“高维小样本”问题，如自然语言处理（文档分类，词义消歧，机器翻译）、计算机视觉（对象预处理、目标识别，图像分割）、生物计算（基因筛选、蛋白质功能预测）、认知心理学等。因此，研究特征选择和半监督分类具有重要的理论价值和现实意义。

1.3 特征选择与半监督分类存在的问题

尽管解决“高维小样本”问题的方法有很多，本文主要研究特征选择和半监督分类方法。随着研究的深入，特征选择和半监督分类方法中需要解决的问题不断增多，本文认为，目前需要解决以下几个问题：

（1）选取表征能力强的特征子集

John[22]从提高预测精度的角度对特征子集的选取作出了定义：特征子集的选取是一个在增加分类器精度，或者在不降低分类器精度的条件下降低特征集维数的过程。目前很多特征选择的方法在选取特征时只考虑对单个特征的类别识别能力进行排序，然后选取得分最高的前 k 个特征。但由前 k 个强分类特征组成的特征子集并不一定具有很强的类别区分能力，因此这类特征选择方法不能直接从大量的特征中选出具有强类别识别能力的特征子集。在面对具体的高维特征选择问题时，现有的特征子集选择方法中，还不存在一种方法能够完全解决这类问题。

（2）协同训练方法的假设前提太强、标记样本利用不充分

在“高维小样本”问题中，利用已有少量标记样本的信息构建多分类器系统协同训练能有效地改善系统的分类性能。但标准的协同训练方法要求数据集有两个充分冗余的视图，即存在两个相互独立的特征集，并且每个特征集都足以训练一个分类器，这个条件在很多情况下很难以满足。因此弱化假设前提，充分利用标记样本数据是一个需要研究的问题。

（3）弱分类方法集成解决半监督分类问题

大量的研究成果表明，多分类器集成能有效地提高系统预测结果的准确性和稳定性，并能有效缓解过拟合问题，因而它被广泛用以提升传统“弱”机器学习方法的性能。应用多分类器集成方法解决“小样本”问题是一条很好的途径，但新背景下的集成学习方法由于缺少足够的标记样本用于训练基分类器，致使基分类器的性能不高，进而导致集成后的分类器总体性能也同样不高。因此如何发挥集成学习在半监督学习中的优势是一个新的研究问题。

（4）充分利用大量未标记样本的几何结构信息

尽管获取标记样本（“小样本”）非常困难，但是大量的未标记样本的获取通常都非常容易。如何利用这些未标记数据蕴含的几何结构信息提高分类性能是半监督学习方法的一个重要分支，也是半监督学习方法相对于传统机器学习方法的一个非常重要的区别与优势，这种半监督学习方法与传统的流形学习具有很多相似之处。目前已有一些基于数据内部几何结构的半监督方法，但这些方法通常设计比较复杂，参数调节较为繁琐。因此设计出简单、实用又能挖掘数据内部几何结构的半监督分类器是一个值得重点研究的问题。

1.4 本文的研究工作及成果

为了解决上一节所提出的 4 个问题，本文对此主要做了 4 个方面的工作，如图 1.1 所示，他们分别对应本文的第三章至第六章。本文的主要工作受到了**国家 863 目标导向项目“基于网格的数字化医疗决策支持系统”**（项目编号：**2006AA02Z347**）和**国家国际科技合作计划(中英联合创新项目 ICUK)“医学图像特征量化分析系统”**（项目编号：**2009DFA12290**）的资助，主要研究成果如下：

（1）在 Relief 方法基础上，将对单特征类别区分能力的评价扩展到对特征子集的类别区分能力评价，并结合贪心搜索策略，提出了一种基于强类别区分能力的特征子集选取方法（FSCRF，Feature Subset Category Resolve Capability Based Feature Selection）。实验结果表明，该方法相比其他常用方法，能在大多数情况下有效地选

出特征数目更少、分类精度更高的特征子集，并在此基础上给出了该方法在老年痴呆诊断方面的实际应用，取得了满意的结果。

（2）分析了协同训练类型方法存在的一些问题，提出了一种新的交叉训练半监督学习分类方法（NC-T，New Cross-Training），该方法将标记样本划分成三份，并利用三个基分类器对其进行训练。不需要假设数据特征集存在两个或多个独立特征视图，且能充分利用标记样本信息。实验结果显示，NC-T 方法的分类精度相比标准协同方法在多数情况有所提高。

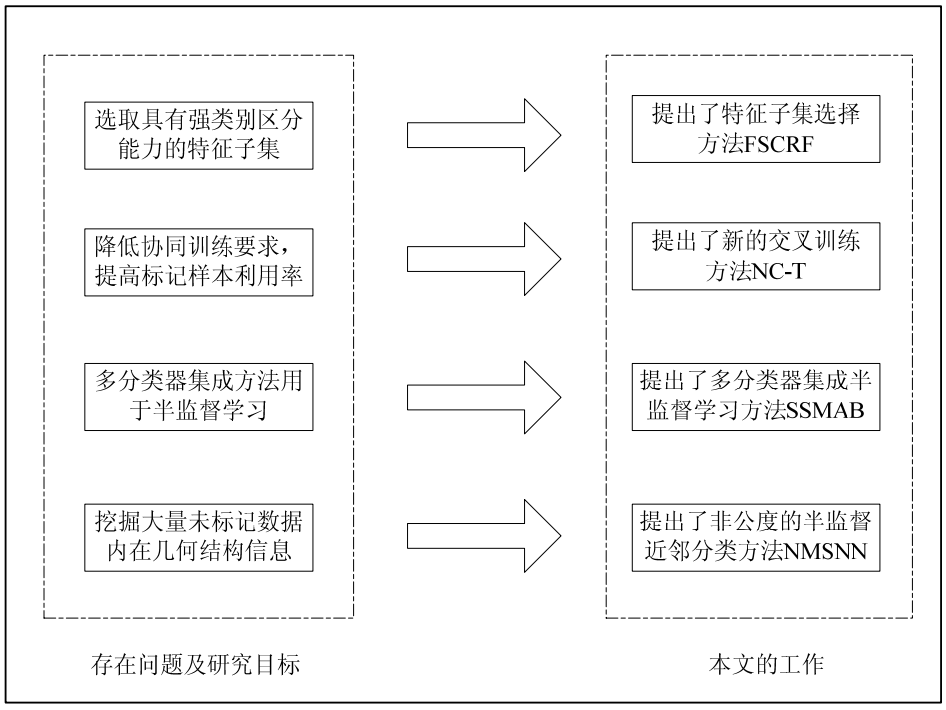

图 1.1 本文的主要工作

（3）集成多分类器方法的确能改善对基分类器的性能，但已有的集成方法对基分类器的分类精度要求较高，而多类 Adaboost 方法能有效降低对基分类器的要求，因此，在融合大量未标记样本信息下，提出了多类 Adaboost 集成半监督分类方法（SSMAB，Semi-supervised Multi-class AdaBoost），该方法只需要基分类器的分类精度达到$1/K$（K 为类别数目），就能取得较满意的效果，实验结果表明，在分类精度上，它与同类型的方法在多数情况下占有一定的优势。

（4）非公度距离度量对数据之间距离关系的度量更加合理，提出了一种非公度的半监督近邻分类方法（NMSNN，Non-metric based Semi-supervised Nearest Neighbor），该方法定义了一个代价长度函数用以度量数据间的距离，有效地利用了未标记样本的几何结构信息和标记样本的类别信息，不仅考虑了数据点之间的直接关系，更借助其它数据点，考虑了全局信息，且只需要设置一个参数，相比同类型的方法，更简单实用。实验结果表明，NMSNN 方法具有良好的分类精度和稳定性。

1.5 论文结构

本文共有 7 章，自第 2 章开始论文的主要内容如下。

第 2 章重点介绍了特征选择和机器学习三种类型（有监督学习、无监督学习、半监督学习）的基本概念，并详细介绍了特征选择和半监督分类的研究现状及进展。

第 3 章提出了一种能直接选取具有强类别区分能力的特征子集方法 FSCRF，给出了该方法在老年痴呆诊断方面的实际应用。实验结果表明，相比其他常用方法，FSCRF 能在大多数情况下选出特征数目更少、分类精度更高的特征子集。

第 4 章指出了协同训练类型方法的相关问题，提出了一种的新的交叉训练的半监督分类方法 NC-T。实验结果表明，在多数情况下，NC-T 方法的分类精度比标准协同方法有所提高，并且不需要假设数据特征存在两个或多个独立特征子集。

第 5 章提出了一种多类别多分类器集成半监督分类方法 SSMAB，该方法只需要基分类器的分类精度达到$1/K$（K 为类别数目），就能取得较满意的效果。实验结果表明，在多数情况下，它比同类型的方法在分类精度上占有优势。

第 6 章提出了一种非公度的半监督学习方法 NMSNN，该方法有效利用未标记样本的几何结构信息和标记样本的类别信息，只需要设置一个参数。实验结果表明，该方法设置简单，且具有良好的分类精度和稳定性。

第 7 章为本文的总结，并展望了特征选择及半监督分类仍需努力的研究方向。

最后，在附录部分给出了本人在攻读博士学位期间发表（录用）论文情况和参加科研项目情况，以及本文的英文缩写名词列表。

2 相关研究进展

自从特征选择和半监督学习方法提出以来，关于它们的研究工作就一直没有停止过。尤其自 20 世纪 90 年代以来，随着数据挖掘技术的迅猛发展，“高维小样本”等问题成为了这一领域的一大障碍。为了解决这些问题，很多学者对此展开了大量卓有成效的研究，并在特征选择和半监督分类方法研究方面提出了许多新的概念和方法，取得了丰硕的研究成果。

2.1 相关重要概念

（1）特征选择（Feature Selection）最主要的任务就是去除不相关或者冗余的特征[23]。不相关特征是指与类别不相关，删除后不影响学习性能的特征。冗余特征通常是单个特征与类别相关，但组成子集时与其他特征高度相关，因而被删除后也不影响学习性能。特征选择一般通过按一定准则对特征进行排序，或者在不降低学习性能的前提下寻找一组最小的特征子集来完成。特征选择方法通常为三类：包装法（Wrapper）、过滤法（Filter）及嵌入法（Embedded）。

包装法[24-26]是一种非常直接的挑选特征子集的方法。它与特定的分类器结合在一起，通过交叉验证在特征子集上对分类的精度进行评价，挑选分类精度最高的一组特征子集作为最优特征子集。包装法的缺点是：计算太复杂，需要调节的参数过多，容易过拟合，与特定分类器绑定，所选特征泛化能力（Generalization Ability）较差。

过滤法主要通过度量特征与类标记的关联程度等进行特征筛选。它独立于分类器，具有良好的泛化能力，该方法计算相对包装法要简单。缺点：不能处理特征冗余问题[27]。

嵌入法本质上是包装法的一个扩展，其特征选择是在对一个特定的分类方法训练过程中进行的。如特征评价过程中嵌入支持向量机学习方法，形成逐步迭代特征剔除法（SVM-RFE，Support Vector Machine- Recursive Feature Elimination），SVM-RFE 方法对高维小样本数据的处理有很好的效果[28]。

（2）PAC 可学习理论是 1984 年 Valiant[29]提出的概率近似正确（PAC，Probably Approximately Correct）机器学习理论，它是机器学习领域中具有里程碑意义的研究成果。正是由于这一杰出的理论成果所带来的深远影响，Valiant 荣获了 2010 年度计算机界最高奖——图灵奖。在 PAC 理论提出之前，传统的模式识别理论都是以概率为 1 成立为基础，学习 $error_D(h)=0$ （$error_D(h)\equiv \Pr_{x\in D}\left[c(x)\neq h(x)\right]$表示随机样本被错分的概率）的假设 h，但这样的要求往往在实际中是不可行的：首先要求对样本集 X 中每个可能的实例都提供训练样例；其次要求训练样例无误导性。而 PAC 理论认为应以概率近似正确为基础。在已有的算法中，弱可学习理论及集成（Ensemble）学习都以 PAC 为理论基础。PAC 可学习理论的定义为：

在规模为 n 的样本集 X 上的一类别 C，学习方法 L 使用假设空间 H。当对所有 $c\in C$，X 上的分布 D，ε 和 δ（满足 $0<\varepsilon$，$\delta<1/2$），学习方法 L 将以至少 $1-\delta$ 的概率输出一般假设 $h\in H$，使得 $error_D(h)\leq\varepsilon$，这时候称 C 使用 H 上的 L 是 PAC 可学习的，所使用的时间为 $1/\varepsilon$，$1/\delta$，n 以及 $size(c)$ 的多项式函数。

其中 $1/\varepsilon$，$1/\delta$ 表示对输出假设要求的强度，n 和 $size(c)$ 分别表示样本空间 X 和类别空间固有的复杂度，$size(c)$ 为 c 的编码长度。该理论表明，学习方法 L 是 PAC 可学习，必须满足两个条件：（1）L 必须以任意高的概率（$1-\delta$）输出一个错误率任意低（ε）的假设；（2）学习过程必须是高效的，时间复杂度最多为多项式时间。

此外 PAC 理论还推导了训练样本数目的一般理论下界，$m\geq\frac{1}{\varepsilon}\left(\ln|H|+\ln(1/\delta)\right)$。它表明了训练样本数目 m 足以保证任意一致假设 h 是可能（可能性为 $1-\delta$）近似（错误率为 ε）正确的。且 m 随着 $1/\varepsilon$ 线性增长，随着 $1/\delta$ 和假设空间 H 的规模对数增长。

（3）无监督学习（Unsupervised Learning）是指不加入人工标记信息的学习方法，在统计学中通常称之为密度估计（Density Estimation）。其数学描述为：令 $X=(x_1,...,x_n)$ 为 n 个没有标记类别的数据样本，其中 $x_i\in\mathcal{X}$，$i\in\{1,...,n\}$，最经典的做法是假定这些数据在 $\mathcal{X}$ 是独立同分布（I.I.D，Idependently and Identically

Distributed）的。X 通常被转换成一个 $n\times d$ 的矩阵 $\mathbf{X}=(x_i^{\mathrm{T}})_{i\in[n]}^{\mathrm{T}}$，其中每行存储一条样本数据。无监督学习方法就是在 $\mathbf{X}$ 中挖掘出有价值的内在几何结构来，它的根本目的就是要凭借 $\mathbf{X}$ 估计出数据后面潜藏的密度分布来。无监督学习的研究通常有以下三种形式：特征降维、聚类、异常点检测。

（4）有监督学习（Supervised Learning）是指利用一组已知类别的样本数据训练、调整分类器的参数，使其达到所要求性能的过程。有监督学习的目的是学习一个从 x 到 y（类别标记）的一个映射，其中，给定训练集的数据通常为数据对 (x_i,y_i)，$y_i\in\mathcal{Y}$ 为样本 x_i 的标记或者类别。如果标记是数字，则 $y=(y_i)_{i\in[n]}^{\mathrm{T}}$ 表示标记列向量。同样会假定所有这些数据对在（$\mathcal{X}$，$\mathcal{Y}$）为独立同分布。映射结果的好坏常通过测试数据集在其上进行评估。$\mathcal{Y}=R$（一维）或者 $\mathcal{Y}=R^d$（多维），当标记 $\mathcal{Y}$ 为连续值时，叫做回归（Regression）；当标记 $\mathcal{Y}$ 为离散值时，则为分类问题，这时 y 值只从有限集中取值，对于这种有监督的学习通常有生成模型（Generative Model）和判定模型（Discriminative Model）两大类方法。生成模型试图通过无监督过程评估类条件概率密度分布 $p(x\mid y)$，其预测概率密度可以通过 Bayes 理论推导得到：

$$p(y\mid x)=\frac{p(x\mid y)p(y)}{\int_y p(x\mid y)p(y)dy} \tag{2.1}$$

事实上，$p(x\mid y)p(y)=p(x,y)$ 就是数据的联合分布，所有的数据对 (x_i,y_i) 均由其产生；判定模型不评估 x_i 如何产生，而是直接评估 $p(y\mid x)$。判定方法比生成方法要容易学习且更有实际应用价值，但是只能给出判定的类别，不能把整个场景给描述出来。生成模型的主要代表有高斯模型（Gaussians Model）[30]、朴素贝叶斯（Naïve Bayes）[31]、混合高斯模型（Mixtures of Gaussians Model）[32]、马尔科夫随机场（Markov Random Fields）[33]等；判定模型主要有：逻辑回归（Logistic Regression）[34]、支持向量机（Support Vector Machine）[35]、神经网络（Neural Networks）[36]等。

（5）半监督学习（Semi-Supervised Learning）是一种介于有监督学习和无监督

学习之间的学习方法。半监督学习方法通常分为两大类，一类称之为直推式学习方法（Transductive Learning），由统计学习理论与支持向量机创始人 Vapnik[37]提出。在给定标记样本和未标记样本的情形下，直推式方法仅对未标记样本进行标记。而与之相对的归纳学习（Inductive Learning）需学习一个定义在整个数据空间 $\mathcal{X}$ 的预测模型。

此外，根据给定监督信息方式的不同，也有两种分类方式：第一种就是除了未标记数据外，还提供少量的监督信息用于学习。通常这些信息指的是样本的标记信息。在这种情况下，数据 $X=(x_i)_{i\in[n]}$ 被分成两部分：一部分为 $X_l=(x_1,...,x_l)$，其标记为 $Y_l=(y_1,...,y_l)$；另一部分为 $X_u=(x_{l+1},...,x_{l+u})$，其标记未知。绝大多数半监督学习方法均关注这种情况；另一种就是学习前先给定一些经验限制作为监督信息，如规定“有些样本点必须归在一起，有些则不能归为一起”[38, 39]，这类学习方法主要是无监督方法在一定约束条件下的延伸。由于给定先验知识的方式不一样，建模方式也有较大差异。

本文主要重点研究第一类情形下的半监督学习方法。

2.2 特征选择研究进展

研究特征选择方法的一个最根本的原因是维数灾难问题（The Curse of Dimensionality）[40]的存在。该问题早在 1957 年由 Bellman 正式提出，认为在给定逼近精度的条件下，估计一个多元函数所需要的样本点数随着变量个数的增加以指数形式增长。2000 年，统计学家 Donoho[41]在由美国数学学会举办的“21 世纪的数学挑战”大会上又重新提到了数据的维数灾难问题。可以想见，高维数据处理面临着极大挑战与困难。机器学习理论证明，对于特征值和类别为二进制（0-1）类型的数据集，学习的假设空间为 2^{2^d}（d 为样本的特征维数），因此需要 $O(2^d)$ 个样本[42, 43]才能学习一个概率近似正确的 PAC 假设空间[29]。

因而特征选择作为一种削减（按指数级削减）假设空间数量的方法显得尤为重

要。它主要用于以下几个方面：数据可视化显示、数据理解、数据冗余与不相关特征剔除，以及提高分类或聚类性能等[44]。但寻找一个最优特征子集是一个 NP-hard[45] 问题，因此绝大多数情况下只能获得次优特征子集，当特征维数很高时更是如此。特征评价和搜索策略的设计对特征选择非常重要，对此国内外已有不少研究。此外，对特征选择方法稳定性方面也有相关研究[46-50]。图 2.1 给出了特征选择的一般过程。

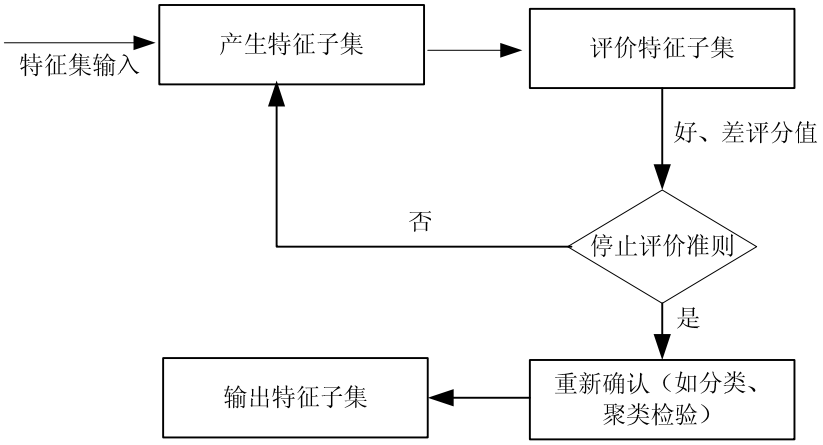

图 2.1 特征子集选择一般过程

更多的内容可以参考：Guyon 与 Elisseeff 整理的特征选择专题报道[51]；Liu 和 Yu 的综述[44]中对分类和聚类中的特征选择方法进行了统一整理；Saeys 等人给出了特征选择在生物信息方面的应用综述[52]；此外还有特征选择方法的相关书籍[23, 53]等。与机器学习方法类似，特征选择也分为三大类：有监督特征选择、无监督特征选择及半监督特征选择三个部分，后续三节将按照这三大类介绍特征选择的研究进展。

2.2.1 有监督特征选择

有监督特征选择通常是指有类别标记参与评估与计算的特征选择，通常的想法是选择一组最能代表原始特征集、类别区分能力强的特征子集。这种特征选择一般要求：所选的特征数目尽可能的少，特征与类别高度相关，单个特征之间尽可能的独立。其方法主要分为包装方法和过滤方法两类，此外也有嵌入方法。

其特征评价中一般采用的度量方式主要有：距离（Distance），如 Kira 和 Rendell

提出的 Relief 方法及改进版本[54-56]；信息增益（Information Gain）[57]，假定两个特征 F_1 和 F_2，计算特征 F_i 的信息增益为 $E_0 - E_i$，特征信息增益越大，表示该特征越好。其中 E 为信息熵。E_0 为除去特征 F_i 之前的信息熵，$E_0 = -\sum_c p_c \log p_c$，$c = 1,2,...,C$，$p$ 为类 c 的估计概率。E_i 为除去 F_i 之后的信息熵；依耐性（Dependency），如 Song 等人提出的特征过滤框架（BAHSIC，Backward elimination Hilbert-Schmidt Independence Criterion）[58]，采用 HSIC（Hilbert-Schmidt Independence Criterion）作为特征与类别标记的依赖性度量；信息度量（Information），Koller 和 Sahami[59]提出了 Markov blanket 特征选择概念 M，对于给定的目标 Y、$M \subset F \& Y \notin M$；如果 $Y \perp (F-M)|M$，则称 Y 条件独立于除 M 外的其他特征变量，一个最小的 Markov blanket 就有可能为所要找的特征子集。寻找最小 Markov blanket 的方法在最近的文献中被加以研究[27, 60, 61]；相关性（Correlation Measures），如 Hall 提出的基于相关性的特征选择方法（CFS，Correlation-based Feature Selection）[62]，该方法具有很好的特征子集选择效果，但是 CFS 作为一种特征子集评价方法，同样没能清晰地处理好冗余问题，并且迭代运算复杂度是特征数的平方与样本数的线性时间的乘积，且 CFS 对数据异常点（Outliers）也比较敏感[27]。Peng 等人提出了最小冗余最大相关法（mRMR，minimal-Redundancy Maximal Relevance）[63]来处理特征冗余问题。mRMR 方法计算效率比较高，并且从理论上证明了为什么选出的特征子集对分类是有效地；Fisher 分数（Fisher Score），对每个特征进行打分，给定数据集 $X = [x_1, x_2, ..., x_m]$，及相应的类别 $Y=[y_1, y_2, ..., y_m]$，该方法能对每个特征按式(2.2)计算出一个 Fisher 分数：

$$F_r = \sum_{i=1}^{c} n_i (\mu_i - \mu)^2 \Big/ \sum_{i=1}^{c} n_i \delta_i^2 \tag{2.2}$$

式(2.2)中，μ_i 和 δ_i 分别代表这一类样本该特征对应数值的均值和方差，n_i 代表该类样本的个数，c 代表类别数目。

2.2.2 无监督特征选择

无监督特征选择没有任何指导信息，其目的是根据一定的准则，从数据中找出一组数量最少、最能发掘数据本质聚类的特征。Dy 和 Brodley[64]给出了无监督特征选择包装方法的框架图。

无监督特征选择方法[64-72]与有监督特征选择方法最主要的区别为：样本没有标记类别信息参与特征选择分析。对此 Handl 和 Knowles[73]给出了详细的阐述。该类方法主要通过统计分析数据本身的信息来评价特征。无监督特征选择方法也主要分两大类：包装法和过滤法。

包装法主要是指用聚类方法引导特征选择，它一般通过聚类质量的高低来评价特征子集的优良。如作者 Dy 和 Brodley[64]将离散矩阵（Sactter Matrices）和可分性准则（Separability Criteria）作为特征选择准则，该准则假定感兴趣的特征能够将数据聚成单峰分布（Unimodal）、可分离（Separable）的多个聚类，求解过程采用最大期望 EM 算法求解。相关定义如下：

$$S_w = \sum_{j=1}^{k} \pi_j E\{(X-\mu_j)^{\mathrm{T}} \mid \omega_j\} = \sum_{j=1}^{k} \pi_j \sum_j \tag{2.3}$$

$$S_b = \sum_{j=1}^{k} \pi_j (\mu_j - M_o)(\mu_j - M_o)^T \tag{2.4}$$

$$M_o = E\{X\} = \sum_{j=1}^{k} \pi_j \mu_j \tag{2.5}$$

式(2.3)−(2.5)中，S_w 为类内离散矩阵，S_b 为类间离散矩阵，π_j 表示样本属于某个聚类 ω_j 的概率，X 为表示数据的 d 维随机向量，k 表示聚类数目，μ_j 为聚类 ω_j 的样本均值，M_o 为总样本均值，$\sum_j$ 为聚类 ω_j 的样本协方差矩阵，$E\{\cdot\}$ 为数学期望操作。

而过滤方法常通过对特征进行评分实现，如拉普拉斯分数（Laplacian Score）。其主要思想是要保持数据的局部几何结构。经常用于流形学习或者降维。Laplacian Score 被 He 等人[74]用于特征选择，就是要寻找那些能够保持局部几何结构的特征。

如(2.6)式所示，L_r 表示第 r 个特征的拉普拉斯分数。该值越小表示对应的特征越好。其中 f_{ri} 表示第 i 个样本第 r 个特征值，**L** 为拉普拉斯矩阵，**D** 为对角矩阵。

$$L_r = \frac{\sum_{ij}(f_{ri}-f_{rj})^2 S_{ij}}{\sum_i (f_{ri}-\mu)^2 D_{ii}} = \frac{\tilde{f}_r^{\mathrm{T}}\mathbf{L}\tilde{f}_r}{\tilde{f}_r^{\mathrm{T}}\mathbf{D}\tilde{f}_r} \tag{2.6}$$

式 (2.6) 中，$\tilde{f}_r = f_r - f_r^{\mathrm{T}}\mathbf{D1}/\mathbf{1}^{\mathrm{T}}\mathbf{D1}$，$f_r = [f_{r1}, f_{r2}, ..., f_{rn}]^{\mathrm{T}}$，$\mathbf{D} = diag(\mathbf{S1})$，$\mathbf{1} = [1,...,1]^{\mathrm{T}}$，**L = D - S**，**S** 为相似性度量矩阵，且：

$$S_{ij} = \begin{cases} e^{-\|x_i - x_j\|^2/t} & \text{如果数据 } i \text{ 和 } j \text{ 是邻居,} \\ 0 & \text{否则} \end{cases} \tag{2.7}$$

无监督特征选择需要解决的主要问题在文献[64]已经提出：（1）特征子集不同会导致聚类数目不同；（2）特征评价准则与特征子集的维数存在偏差（Bias）。

2.2.3 半监督特征选择

半监督特征选择方面的研究目前还比较少，主要通过两种方式给定先验知识，给定一个数据集 $X = [x_1, x_2, ..., x_n]$，一种是利用少量标记样本 $X_L \subset X$ 加上大量未标记样本 $X_U \subset X$；另一种是给定两个成对约束（Pairwise Constraints）集合，分别为必须相连（***Must-link***）集合 M 和不能相连（***Can't-link***）集合 C，其中 $M = \{(x_i, x_j) \mid x_i \text{和} x_j \text{必须属于同一类}\}$，$C = \{(x_i, x_j) \mid x_i \text{和} x_j \text{属于不同类}\}$；然后加上大量未标记样本信息一起对特征或特征子集进行评价和选取。

（1）标记样本与未标记样本结合

Wu 和 Li[75]利用直推式支持向量机[76]（TSVM，Transductive Support Vector Machine）方法进行特征选择，TSVM 利用了未标记样本的信息，但特征选择的基本思想来源于 Guyon 等人[28]：通过逐步剔除权重小的特征达到特征选择的目的；Zhao 和 Liu[77]提出了一种基于谱分析的半监督特征选择方法。该方法利用一个正则化框架来利用标记样本和未标记样本，主要引入了一个聚类指示器（Cluster Indictors）的概

念，然后通过评价聚类可分性和一致性对所选特征进行评分。在 Sun 等人提出的逻辑迭代 Relief 方法（Logistic I-RELIEF）[79]的基础上，Yubo 等人[78]在目标函数中添加未标记样本信息以实现半监督特征选择。Ren[80]等人提出了一种包装法类型的半监督特征选择方法，该方法的基本思想是：选择几个特征作为初始的特征集，然后利用这个特征集在标记样本上训练，训练完后在未标记样本集上进行预测，随机的从未标记样本集中选一定比例的数据加入到标记样本集中，形成新的训练集。不断的随机评价所选特征，特征选择就在随机评价过程中进行，选择出现频率最高的一个特征加入到当前的特征集中，算法不断循环直至结束，最终获得特征子集。Yaslan 等人[81]利用选取两个相关性特征子集对 Co-Training 算法[82]（之前的两特征子集往往是随机划分的）进行改进。

Zhao 等人提出了局部敏感性半监督特征选择[83]，在原有拉普拉斯分数的基础上加入未标记样本的信息。该方法主要通过修改相似性权重矩阵构建一个类内图（Within-class Graph），其类内相似度定义为：

$$S_{w,ij}=\begin{cases} r & \text{若数据}i\text{和}j\text{有相同的标记，} \\ 1 & i\text{若数据}i\text{或}j\text{未标记，但是数据}i\in knn(j)\text{，或者数据}j\in knn(i)\text{，} \\ 0 & \text{否则} \end{cases} \tag{2.8}$$

同时构建一个类间图（Between-class Graph），其类间相似度定义为：

$$S_{b,ij}=\begin{cases} 1 & \text{若数据}i\text{和}j\text{有不同的类别，} \\ 0 & \text{否则} \end{cases} \tag{2.9}$$

计算特征分数 L_r：

$$L_r=\frac{f_r^{\mathrm{T}}\mathbf{D_b}-\mathbf{S_b}f_r}{f_r^{\mathrm{T}}\mathbf{D_w}-\mathbf{S_w}f_r}=\frac{f_r^{\mathrm{T}}\mathbf{L_b}f_r}{f_r^{\mathrm{T}}\mathbf{L_w}f_r} \tag{2.10}$$

式(2.10)中，$\mathbf{D_w}=diag(\mathbf{S_w}\mathbf{1})$，$\mathbf{D_b}=diag(\mathbf{S_b}\mathbf{1})$，$\mathbf{1}=(1,1,...,1)^{\mathrm{T}}$。$L_r$ 表示第 r 个特征的拉普拉斯分数，该值越大对应的特征越好。f_r 表示第 r 个特征值，$\mathbf{L}$ 为拉普拉斯矩阵，$\mathbf{D}$ 为对角矩阵。

（2）成对约束集与未标记样本结合

Zhang 等人[84]提出了一种用于特征选择的新半监督评价方法，CS 方法，该方法通过利用成对约束集来引导特征选择，并给出了一个约束分数（CS，Constraint Score）。该文通过实验对著名的 Fisher Score 和 Laplacian Score 方法进行了比较分析。结果表明 CS 方法只需要少量成对约束就可以获得比较好的结果。

$$C_r^1 = \frac{\sum_{(x_i,x_j)\in M}(f_{ri}-f_{rj})^2}{\sum_{(x_i,x_j)\in C}(f_{ri}-f_{rj})^2} \tag{2.11}$$

或者

$$C_r^2 = \sum_{(x_i,x_j)\in M}(f_{ri}-f_{rj})^2 - \sum_{(x_i,x_j)\in C}(f_{ri}-f_{rj})^2 \tag{2.12}$$

式(2.11)和式(2.12)中，约束集合 M 和 C 分别为 ***Must-link*** 集合和 ***Can't-link*** 集合。

根据以上(2.11)式和(2.12)式两个评价公式（C_r^1 和 C_r^2）之一，可以对所有特征按分数大小进行排序。约束分数 CS 越小，对应的特征越好。该文还利用图谱理论进行了扩展。但该方法只是对每个特征进行了分析，没有对特征子集进行分析。另一个缺点就是，特征的评价不太稳定，与成对约束集的选取有很大关系。

针对这一缺点，Sun 和 Zhang 又提出了 BCS（Bagging Constrained Score）方法 BCS[85]，将 Bagging 方法引入其中，利用多个约束集来代替原有单一约束集的想法，并通过投票表决的方式学习假设空间，成为一种利用原有 CS 评分准则的一种集成方法，但这两篇论文都没有结合未标记样本信息，仅仅在成对约束集 M 集和 C 集上进行特征评价和排序，这样容易造成评价结果的不稳定。

2.2.4 特征子集搜索策略

由于寻找一组最优的特征子集是 NP-hard 问题[45]，因此，当特征数目过大时，穷举所有的特征子集（2^d 个）随着特征维数 d 的不断增加而变成一个让计算机几乎不可能完成的任务，因此搜索局部最优特征子集显得十分必要。用于特征子集搜索的主要方法有：Narendra[86]提出的分支界定特征选择法（B&B，Branch and Bound）以及 Chen 等提出的改进方法（IBB，Improved Branch and Bound）[87]，Whitney 提出

的序贯前向搜索选择（SFS，Sequential Forward Selection）[88]，Marill 和 Green 提出的后向剔除选择法（SBS，Sequential Back Selection）[89]），在此基础上 Stearns[90] 又提出向前加 *l* 个特征和向后减 *r* 个特征前后结合的方法 floating 搜索法（SFFS，Sequential Forward Floating Selection）以及 Pudil 等提出的改进方法（SBFS，Sequential Back Floating Selection）[91]（*r-l* 数目不确定），基于 Tabu 搜索的特征选择[92]，以及其它随机搜索策略，如模拟退火法（SA，Simulated Annealing）[93]，拉斯维加斯特征选择（LV，Las Vegas）[94]，遗传算法（GA，Genetic Algorithm）[95]等等。此外还有各种搜索策略相结合的方法[96]，所有这些方法都是为了缓解 NP-hard 问题所提出的局部最优方法。

表 2.1 主要特征选择方法归类

<table>
<tr><td rowspan="10">评价准则</td><td rowspan="5">过滤法</td><td colspan="6">搜索策略</td></tr>
<tr><td></td><td>穷举法</td><td>随机搜索</td><td colspan="2">序贯搜索</td><td></td></tr>
<tr><td>距离</td><td>B&B[86]
IBB[87]</td><td>Tabu[92]</td><td>Relief[97]
ReliefF[55]
ReliefS[98]
SFS[88]
SBFS[91]</td><td>FSSEM[64]
LS[99]</td><td>Zhao’s[77]
LSDF[83]
Yubo’s[100]
CS[84]
BCS[85]
LS[74]</td></tr>
<tr><td>信息</td><td></td><td>MBEGA[101]</td><td>MBBE[59]
FCBF[60]
mRMR[63]</td><td>Dash's[102]</td><td></td></tr>
<tr><td>依赖性</td><td>Bobrowski's[103]</td><td></td><td>CFS[62]
DVMM[104]</td><td>Mitra's[105]</td><td></td></tr>
<tr><td>一致性</td><td>ABB[106]</td><td>QBB[23]
LVF[94]</td><td>GLSR[107]
BAHSIC[58]</td><td></td><td></td></tr>
<tr><td>包装法</td><td>预测精度
或
聚类效果</td><td>AMB&B[108]
FSBC[109]</td><td>SA[93]
GA[95]
RVE[110]</td><td>SBS-SLASH[111]
RC[112]</td><td>AICC[113]
FSSEM[71]
ELSA[114]
MOEA[73]</td><td>FW-SemiFS[80]</td></tr>
<tr><td>嵌入法</td><td></td><td colspan="3">SVM-RFE[28]</td><td></td><td>Wu’s[75]</td></tr>
<tr><td></td><td></td><td colspan="3">有监督（分类）</td><td>无监督（聚类）</td><td>半监督</td></tr>
<tr><td></td><td></td><td colspan="5">数据挖掘方式</td></tr>
</table>

为了能够清晰的对特征选择方法进行分类，在 Liu 和 Yu[44]的基础上，按照评分准则、搜索策略、数据挖掘方式（有监督、无监督、半监督学习）进行了分类整理，具体见表 2.1。

特征选择中搜索策略采用的停止准则通常有：（1）搜索完成；（2）达到给定的搜索上界（比如最大特征个数，迭代的最大次数等等）；（3）在特征子集搜索中，不能找到比目前更好的子集（比如就分类精度而言）；（4）在子集搜索过程，所选子集的分类精度已经达到了设定期望值。

输出方式：（1）特征排序（按一定的评价准则对单个特征进行分析），如 Relief[54]，基于马尔科夫桶的向后剔除法[59]（MBBE，Markov Blanket based Backward Elimination），SVM-RFE[28]；（2）特征子集（对特征子集按照一定的标准进行统一评价）。

2.3 半监督分类研究进展

最早利用未标记样本进行分类的思想源于自学习（Self-Learning）方法，该方法的其他表述形式还有自训练（Self-Training）、自标记（Self-Labeling）及决策导向学习（Decision-Directed Learning）等。自学习方法是一种包装（Wrapper）算法，它使用一个有监督的学习方法不断的自我训练，通常在算法开始阶段仅仅在有标记的样本上进行训练。在后续的每次迭代过程中，部分未标记样本根据当前的决策函数被标记成有标记样本，然后有监督的学习方法在原有标记样本和新标记的样本上不断的训练。自训练的想法最早出现在 Scudder[115]的文献中。与半监督学习概念联系最为紧密的是直推式学习方法，它由 Vapnik 和 Chervonenkis[116]最先提出。与归纳法相对，该方法不推导出针对整个数据集的一般规则，而是着眼于解决未标记样本的预测标记问题。半监督学习兴起于 20 世纪 70 年代，当时已经有学者[117, 118]开始考虑用未标记样本来估计 Fisher 线性分类器判定准则，使用标记样本和未标记样本并结合最大期望算法（EM，Expectation Maximization）[119]完成混合高斯模型的优化建模。

20 世纪 90 年代开始至今，半监督学习引起了很多研究者人员的广泛兴趣，并开

发设计出了很多的具体学习方法，目前主要的半监督分类方法有四大类：早期方法、协同训练类型、基于图的半监督分类、基于支持向量机的半监督分类。具体分类情况见表 2.2，后续 4 小节将对这 4 种类型的半监督分类方法进行详细论述。

表 2.2 主要半监督分类方法归类

早期方法（生成模型）	协同训练类型	基于图类型	基于支持向量机类型
Nigam's[120]	Co-Training[82]	Graph Mincuts[129]	S3VM[137]
USGGM[121]	Co-EM[124]	HMT[130]	TSVM[76]
Fujino's[122]	Co-Testing[125]	LLGC[131]	Xu's[138]
Self-Training[123]	Co-EMT[126]	Szummer's[132]	LapSVM[139]
	Tri-Training[127]	Belkin's[133]	MissSVM[140]
	RASCO[128]	LNP[134]	
		Ando's[135]	
		SGT[136]	

2.3.1 半监督分类早期方法

自训练方法（Self-training）是早期提出的半监督分类方法中一种最简单的学习方式。该方法首先利用已有标记样本训练一个分类模型，然后预测未标记样本，并将置信度高的样本加入原有的训练集中重新训练，不断迭代执行该过程。

生成模型（Generative Models）可能是最早的一种半监督分类方法，生成模型需要估计输入空间 x 和输出空间 y 的联合概率分布 $p(x,y|\theta)=p(y|\theta)p(x|y,\theta)$。例如，$p(x,y|\theta)$ 可以为一个 $\mathcal{Y}$ 上多项式分布（Multinormial Distribution），而条件概率分布 $p(x|y,\theta)$ 可以假定为 $\mathcal{X}$ 上的多元高斯分布（Multivariate Gaussian Distribution），当然也可以为别的分布。引入未标记样本是为了更好的估计模型参数 θ，其中未标记样本 $\{x_i\}_{i=l+1}^{l+u}$ 在参数 θ 下出现的概率（为便于计算，通常取对数）的计算式为(2.13)式。

$$\log p(\{x_i\}_{i=l+1}^{l+u}|\theta)=\sum_{i=l+1}^{l+u}\log\left(\sum_{y\in\mathcal{Y}}p(x_i,y|\theta)\right) \tag{2.13}$$

生成模型的一般思路是，先建立一个包含标记样本和未标记样本信息的联合概率分布模型，如(2.14)式所示，该问题为一个非凹（Non-concave）问题，然后利用

EM[119]等方法计算(2.14)式局部最大值所对应的参数θ^*：

$$\theta^* = \arg\max_{\theta} p(X_l, Y_l, X_u \mid \theta) = \log p(\{x_i, y_i\}_{i=1}^{l} \mid \theta) + \lambda \log p(\{x_i\}_{i=l+1}^{l+u} \mid \theta) \quad (2.14)$$

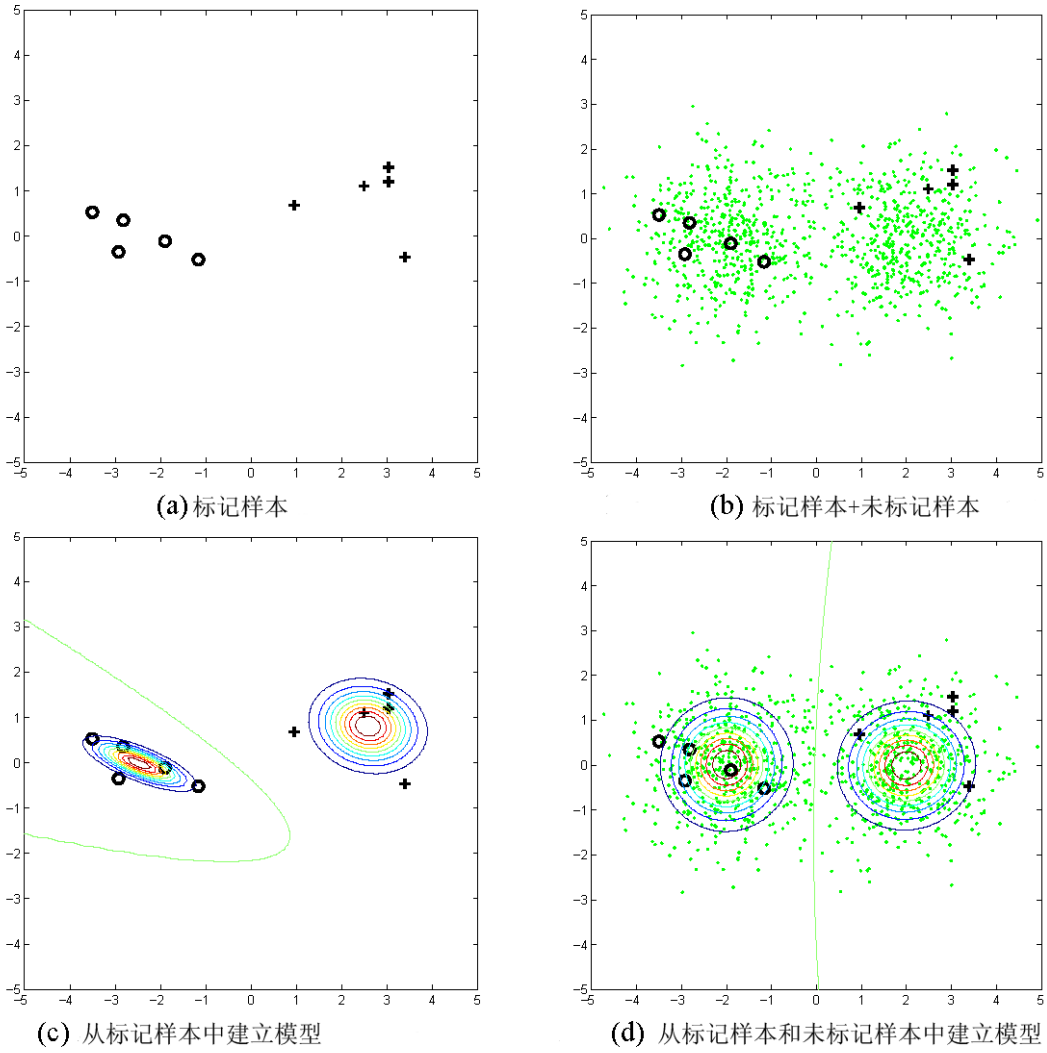

图 2.2 两分类问题中，未标记样本可以改善对混合高斯模型的参数估计

式(2.14)中，X_l，Y_l，X_u分别为标记样本、标记样本类别及未标记样本，λ为标记样本与未标记样本之间的平衡权重。再使用贝叶斯准则(2.15)式计算条件分布，从而得出样本的标记。

$$f_{\theta^*}(x) = \max_{y} p(y \mid x, \theta^*) = \max_{y} p(x, y \mid \theta^*) \Big/ \sum_{y'} p(x, y' \mid \theta^*) \quad (2.15)$$

该方法一般基于以下几个模型建模：混合高斯模型[121]（GMM，Gaussian Mixture Model），Nigam 等人[120]将 EM 算法作用于多模态混合模型（MMM，Multimodal Mixture Model）来完成文档分类任务，实验结果表明，分类的效果比仅仅只用标记样本的效果要好。

以混合高斯模型为例，拥有大量未标记样本、且混合高斯的每个成分是可辨识的，则在理想的情况下，只需要每个成分有一个标记样本，就可以完整的对整个数据分布进行建模。如图 2.2 所示（该示意图由 Xiaojin Zhu 给出）[141]，给出了两分类问题中未标记数据的作用。假定每类为一个高斯分布，其中子图（c）仅通过标记样本拟合数据分布，而子图（d）同时考虑了未标记样本，从而得出的模型更能反映数据的整体分布。

2.3.2 协同训练及多视图模型

在半监督分类研究领域中，有另一类重要的问题需要处理，即，如何利用特征集中存在天然分割的半监督样本集进行分类。如网页的分类问题中，特征集（或简称视图）由两部分组成，出现在网页文本中的词汇集和出现在网页超链接中的词汇集；又如同时利用音频和视频对说话内容进行识别的问题中，数据特征子集就有音频特征子集和视频特征子集。目前有很多学者对此展开研究，比较有代表性的方法有 Co-Training[82]、Co-EM[124]、Co-Testing[125]、Co-EMT[126]以及 Zhou 和 Li 提出的多视图方法 Tri-training[127]。

这一类方法的主要思想均源于 Blum 和 Mitchell 于 1998 年提出的协同训练（Co-Training）这一经典方法，简单的说，Co-Training 策略就是假设数据有两个不同的冗余特征视图（View），而每个视图的特征集合都足以单独训练出一个强分类器。在两个视图的基础上分别利用少量已标记数据训练出两个分类器，再用训练得到的两个分类器分别对未标记样本进行预测，然后从中选出置信度靠前的数据帮助对方重新训练分类器，以改善性能，不断重复这个过程。Co-EM 方法与 Co-Training 方法的不同之处在于其采用迭代算法，每一轮循环中，Co-EM 将一个视图上获得的分类器的概率标注结果给另外一个分类器。从根本上说 Co-EM 只是在各个视图中应用了

最大期望 EM 算法，但整体上而言，又不符合最大期望算法框架。而 Tri-training 方法需要三个分类器来进行协同训练，该方法对特征集合上的 3 个分类器所用监督分类算法都没有约束，且不需要交叉验证，因此效率更高。针对实际应用中很难以找到天然的特征子集分割，Wang 等人提出了随机子空间协同训练方法（RASCO，Random Sub-space Method for Co-training）[128]，Yaslan 等人在此基础上提出了相关随机特征子空间的改进方法 Rel-RASCO[81]。周志华[142]回顾了协同训练框架发展过程、理论分析及意义，并指出多视图是协同训练的一个重要研究方向。

2.3.3 基于图的半监督分类

基于图的半监督模型[129, 130, 139, 143, 144]一般先构建一个邻接图$G(V,E)$，其中V代表标记数据点和未标记数据点，E表示数据点之间的边。该类模型的目的就是要寻找一个函数f，使它既能很好的拟合标记样本又能使得未标记样本在相似区域的f输出变化缓慢。w_{ij}为数据点之间的权重（通常为 kNN 图或者高斯权重图上点间相似关系），这个值越大就暗含数据点之间越相似，因此$f(x_i)$和$f(x_j)$应该越相同越好。经函数f形式化后，图的能量方程为：

$$\sum_{i,j=1}^{l+u} w_{ij}(f(x_i)-f(x_j))^2 \tag{2.16}$$

图的能量可以根据不同的$f\in\mathcal{F}$从小到大排列，越小就表示图越光滑（指的是图中越相似的点，相应的f输出值变化越缓慢）。上述能量图可以理解为一个没有规范化的图拉普拉斯矩阵（GLM，Graph Laplacian Matrix），通常一个基于图的半督分类优化问题可以写成类似(2.17)式的形式，许多变种往往只是代价函数$c(\cdot)$、正则项$\|f\|$、构图方式不一样而已。该问题通常可以为以下凸优化问题，式(2.17)中$\lambda_1,\lambda_2>0$。

$$\arg\min_f \frac{1}{l}\sum_{i=1}^{l} c(f(x_i),y_i)+\lambda_1\|f\|^2+\lambda_2\sum_{i,j=1}^{l+u} w_{ij}(f(x_i)-f(x_j))^2 \tag{2.17}$$

图 2.3 给出了基于图的半监督分类的示意图，先将标记样本与未标记样本一起建立一个图，后通过图将标记样本的类别传播给未标记样本。

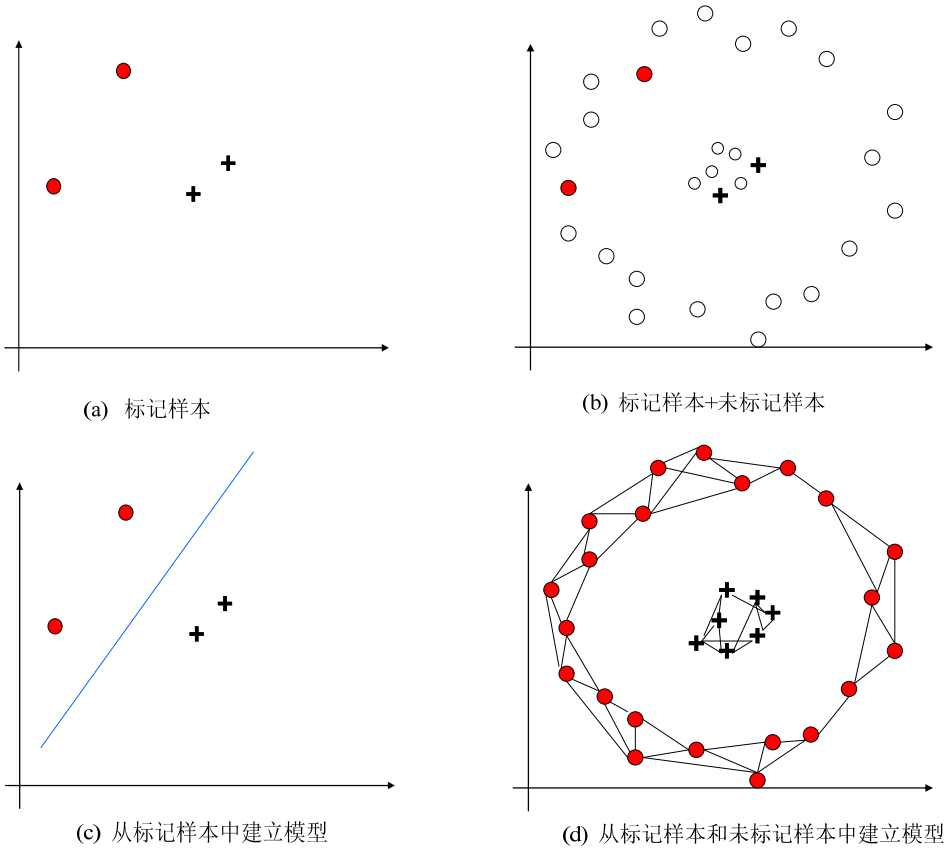

图 2.3 两分类问题中，可以将标记样本的类别通过图传播给未标记样本

基于图的半监督分类方法最大的不足是，计算量比较大，时间复杂度一般为样本数量的立方。因此降低时间复杂度，处理大规模数据集，是这类方法必须要解决的问题。为此，Amarnag 和 Jeff 已设计出相应的快速算法[145]，并取得了满意结果。

2.3.4 基于支持向量机的半监督分类

基于支持向量机（SVM，Support Vector Machine）的半监督分类方法[146-149]通常假定决策边界 $f(x)=0$ 应该在两类 $y=\{1,-1\}$ 之间的低密度区域（密度的评估包含了未标记样本）最为合适。为未标记样本 x 定义一个帽子状（Hat）损失函数式(2.18)。

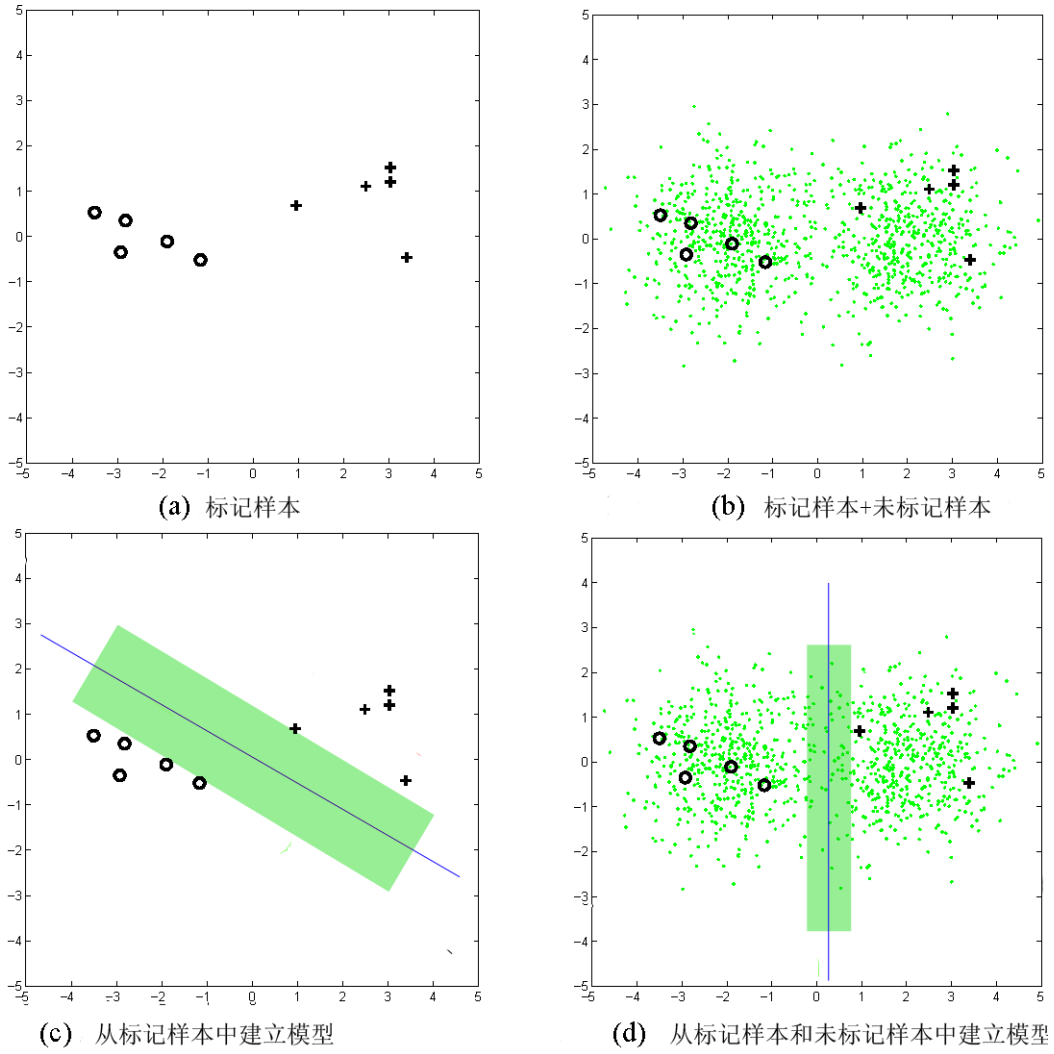

图 2.4 两分类问题中，未标记样本能帮助支持向量机找到更为合理的分界线

$$\max(1-|f(x)|,0) \tag{2.18}$$

式(2.18)中，$-1<f(x)<1$，0 除外。该函数主要是度量未标记样本大间隔的可分性，平均所有未标记样本的损失值，可以得到

$$\frac{1}{u}\sum_{i=l+1}^{l+u}\max(1-|f(x)|,0) \tag{2.19}$$

对任意的函数 $f\in F$ 可以根据该函数从小到大排列，越小表示决策边界通过使用大间隔可分避免了穿越大多数的未标记样本。为了使得分类得到的函数 f 既能与标

记样本的类别保持一致，又能减小在未标记样本中的 hat 函数值。通常优化目标为（非凸优化，由于 hat 函数为非凸优化问题，相关的优化方法可以参考[150]）式(2.20)：

$$\arg\min_{f}\frac{1}{l}\sum_{i=1}^{l}\max(1-y_i f(x_i),0)+\lambda_1\|f\|^2+\lambda_2\frac{1}{u}\sum_{i=l+1}^{l+u}\max(1-|f(x)|,0) \qquad (2.20)$$

Bennett 等人提出的 S^3VM（Semi-supervised SVM）[137]和 Joachims 提出的 TSVM（Transductive SVM）[76]都是 SVM 在半监督分类中的扩展。遗憾的是，为这类问题寻找一个精确解是一个 NP-hard 问题，因此研究者将精力转向求解其近似解，通常将问题转换成半定规划（SDP，Semi-Definite Programming）[138]或者凹凸过程（CCCP，Concave-Convex Procedure）[151]等优化问题求解。基于支持向量机的半监督分类方法最主要的缺点就是计算代价较大。

如图 2.4 所示，给出了支持向量机引入未标记样本所带来的好处。子图（c）仅在标记样本上利用支持向量机寻找最优分界面，而子图（d）利用标记样本和未标记样本一起寻找最优分界面。显然，子图（d）所找的分界面更合理，更能反映数据潜在的整体分布。

2.4 本章小结

本章主要介绍了一些与全文直接相关的基本概念及研究进展，为后续章节的论述提供必要的基础知识。首先给出了特征选择概念的定义，并介绍了有监督特征选择、无监督特征选择、半监督特征选择三种类型，阐述他们之间的区别与联系；其次，介绍了机器学习中 PAC 可学习理论、有监督学习、无监督学习、半监督学习的概念，并重点阐述了半监督分类的主要研究内容及发展历程，详细介绍了半监督学习的自训练方法、生成模型、协同训练方法、基于图的方法以及基于支持向量机的方法等，分析了各方法的主要原理，并指出了存在的主要不足。

3 基于子集类别区分能力的特征选择

进行特征选择时，通常的作法是利用某种评价函数或评价准则独立的对每个原始特征进行评分，然后按照分值从高到底进行排序，从中挑选出若干个分值最高的特征作为特征子集。但这些方法并不能有效地消除所选特征间的相互冗余性，因此，如何从高维数据中选出具有强类别区分能力的特征子集是本章所要研究的问题。

3.1 单特征评价及选择

相对于原始特征集，获取包含特征数目少且表征能力强的特征子集是特征选择的重要目标。一般来说，在给定特征子集的评价标准后，特征选择问题就变成了一个搜索问题。因此，特征子集评价标准的设定显得非常重要。

输入：数据集 $D=\{(x_n,y_n)\}_{n=1}^{N}$

初始化：设置 $w_i:=0,\ 1\le i\le I$，指定抽样次数 M；

Step1. **For** $m:=1:M$ **do**

(1) 随机的从 D 中选出样本 x；

(2) 在 D 中（原始特征空间）找到 x 的 NH(x) 及 NM(x)；

For $i=1:I$ **do**

计算 $w_i:=w_i+\left|x^{(i)}-\mathrm{NM}^{(i)}(x)\right|-\left|x^{(i)}-\mathrm{NH}^{(i)}(x)\right|$；

End For

End For

输出：特征权值 $w_i:=0,\ 1\le i\le I$

图 3.1 Relief 方法的伪代码

在着眼于单个特征评价的算法中，Relief 方法[97]是其中最具代表性的一个算法。

Relief 算法由 Kira 和 Rendell 在 1992 年提出，如图 3.1 所示。其中 NH(x) 及 NM(x) 分别 x 的同类最近邻和异类最近邻。该算法根据特征对其同类与异类近邻的区分能力，能够简单有效地评价每个特征对于分类的贡献程度，是一种广为应用、非常成功的特征选择方法。起初的 Relief 方法仅适用于两分类问题的特征选择，为此，Kononenko 等人对其进行了扩展[152]，扩展后的 Relief 方法可以处理多分类、数据噪声以及特征值缺失等问题；后来，Robnik-Sikonja 等人[153]提出了 Relief 的回归版本——RRelief 方法，将预测变量类型从离散型扩展到连续型；Robnik-ikonja 和 Kononenko 对 Relief 方法从理论和实际应用上进行了充分论述[154]。Yijun 等人[56]提出了 I-Relief 迭代方法，将传统的 Relief 方法转化成有约束的优化问题，并从理论上对算法迭代的收敛性进行了证明。

但是，Relief 及其扩展方法着眼于评价单个特征，选取特征子集时需要设定特征评分阈值，根据评分阈值选取前 k 个特征组成特征子集[155, 156]。因此，子集中每个特征的类别区分能力都强于剩余特征，但其综合区分能力不一定是所有可能子集中最强的。此外，Relief 方法定义的近邻基于原始特征空间，在约减后的特征子空间中，先前的近邻不一定还是近邻[56]。

针对 Relief 方法存在的上述问题，着眼于特征子集的综合区分能力，本章给出了一个能直接评价特征子集的评分方法；结合最好优先特征子集搜索策略，本章提出了一种新的特征子集选取方法，该方法能选出具有强分类能力的特征子集。

3.2 基于子集类别区分能力的特征选择方法FSCRF

本节提出了一种特征子集选取方法 FSCRF（Feature Subset Category Resolve Capability Based Feature Selection Approach），它是一种特征过滤法。该方法的主要目的是在低维特征子空间中寻找具有强类别区分能力的特征子集。其与 Relief 方法不同之处在于能直接选取特征子集而非单特征评价排序；此外对样本的同类 K 近邻和异类 K 近邻度量空间改用特征子集空间而非原始特征空间，其原因是高维数据的近邻判断不准确[157]，相关定义和定理如下。

定义 3.1 设原始特征集为 $F = F_1F_2...F_I$，样本 $x \in F$ 且 $x = f_1 f_2 ... f_I$；设 t 为 x 的子字符串（长度≥ 1），若 t 对 $y \in Y$ 具有表征能力，则称 (t, y) 为模型的一个复合特征，其对应的特征空间称为特征子空间 s。

定义 3.2 设 x_c 是当前类别为 c 的一个样本，则 x_c 在当前特征子空间 s 上与其类别相同，且最近的 K 个近邻称为 $\mathrm{KNH}(x_c, s)$。

定义 3.3 设 x_c 是当前类别为 c 的一个样本，则 x_c 在当前特征子空间 s 上与其类别不同的某一类最近的 K 个近邻称为 $\mathrm{KNM}(x_c, s)$。

定义 3.4 设 d_I 为一个距离测度函数，则当前查询样本 x 与数据集中最远点的距离为 $D_{\max I} = \max\{d_I(D_{In}, x_I) \mid 1 \leq n \leq N\}$。

定义 3.5 设 d_I 为一个距离测度函数，则当前查询样本 x 与数据集中最近点的距离为 $D_{\min I} = \min\{d_I(D_{In}, x_I) \mid 1 \leq n \leq N\}$。

定理 3.1 假设 X 为 I 维随机变量，若 $\lim_{I\to\infty} \mathrm{var}\left(\|X_I\| / E\left[\|X_I\|\right]\right) = 0$ 则 $\frac{(D_{maxI} - D_{minI})}{D_{minI}} \to_p 0$。其中 $E[X]$ 为 X 的期望，$\mathrm{var}(\cdot)$ 为 X 的方差；$\to_p 0$ 表示依概率收敛到常数 0。证明见文献[158]。

定理 3.1 表明，随着特征维数增大，对给定的一个样本，其与最远点和最近点的距离之差的增长速度远远没有其与最近点的距离增长快[157]。该定理意味着，在高维空间中一个样本的最远点和最近点区分不明显，从而导致高维数据近邻判断不准确。

3.3 算法流程

FSCRF 方法主要由特征子集评价和特征搜索两部分组成。在特征子集评价部分，

首先从所有数据中随机抽取 M 个样本，并在特征子空间中对每一个样本 x_c 寻找同类的 K 个近邻 $\mathrm{KNH}(x_c,s)$，同时在与样本 x_c 类别不同的各类数据中也分别寻找 K 个近邻 $\mathrm{KNM}(x_c,s)$；然后按照以下两式计算样本 x_c 与 $\mathrm{KNH}(x_c,s)$ 和 $\mathrm{KNM}(x_c,s)$ 的差异。

$$diff(x_c,\mathrm{KNM}(x_c,s)) := \sum_{c_i=1\&c\neq c_i}^{C}\frac{p(c_i)}{K(1-p(c))}\sum_{k=1}^{K}\left(\frac{1}{N_s}\sum_{j=1}^{N_s}\left(x_c^j(s)-x_{c_i^k}^j(s)\right)^2\right)^{1/2} \tag{3.1}$$

$$diff(x_c,\mathrm{KNH}(x_c,s)) := \sum_{k=1}^{K}\frac{1}{K}\left(\frac{1}{N_s}\sum_{j=1}^{N_s}(x_c^j(s)-x_{c_i^k}^j(s))^2\right)^{1/2} \tag{3.2}$$

1.特征子集选取：

输入：数据集 $D=\{(x_n,y_n)\}_{n=1}^{N}$。

初始化：设最好特征子集 $bestS := \mathrm{NULL}$；

$W_{bestS}=\mathrm{MIN}$； $W_{F_i}:=0,\ 1\le i\le I$，

抽样次数 M；

Step1. $bestS' := bestS$； $W_{bestS'}=W_{bestS}$；

Step2. $\forall\{F_i \mid F_i\notin bestS\ \&\ F_i\in F\}$

(1) $s=bestS'+F_i$；

(2) $W_s := evaluateFsubspace(D,s,M,K)$；

(3) 若 $(W_{bestS'}<W_s)$ 则 $W_{bestS'}:=W_s$；

$bestS' := s$； $bestS := bestS'$；

Step3. 若 $W_{bestS}<W_{bestS'}$ 则转至 *step1*；否则，结束。

输出：最终选择的特征子集 $bestS$。

2.特征子集评价：

$evaluateFsubspace(D,s,M,K)$

按抽样次数 $m:=1:M$ 循环

Step1. 随机的从 D 中选出样本 x_c；

Step2. 寻找 $\mathrm{KNH}(x_c,s)$ 及 $\mathrm{KNM}(x_c,s)$；

Step3. 求 $diff(x_c,\mathrm{KNM}(x_c,s))$、

$diff(x_c,\mathrm{KNH}(x_c,s))$；

Step4. 按公式(3.3)更新特征子集 s 的权重 W_s；

Step5. 若 $m<M$；转向 *step1*。

返回：特征子集的权重 W_s。

图 3.2 FSCRF 方法的计算过程

最后把式(3.1)与式(3.2)的结果代入下式(3.3)，不断更新当前特征子集的评分权重 W_s，每次选取 W_s 值最大的特征子集作为当前的最优特征子集，直到迭代结束，最后选出最大 W_s 所对应的特征子集。

$$W_s := W_s + \frac{diff(x_c, \mathrm{KNM}(x_c, s)) - diff(x_c, \mathrm{KNH}(x_c, s))}{M} \tag{3.3}$$

式(3.3)中，N_s 为特征空间 s 的维数，C 为数据集 D 的类别数，$p(c_i)$ 为第 i 类样本在 D 中所占比例。

在特征搜索部分，由于穷举搜索最优的特征子集是一个 NP-hard 问题，其 I 个特征就有 2^I-1 个非空特征子集，从而使用局部搜索策略显得十分必要。为了能尽可能的在低维特征空间中找到具有强类别区分能力的特征子集，本文均采用最好优先前向引入搜索方法进行特征搜索。FSCRF 方法特征子集选取过程如图 3.2 所示。

3.4 时间复杂度分析

假定有 N 个样本和 I 个特征，当前的特征空间维度为 N_s，数据类别数目、近邻个数、样本数目分别为 C，K，M。则在子空间中归一化所有的样本数据的时间复杂度为 $O(N*N_s)$，计算某一样本数据与所有其他样本数据的欧氏距离时间复杂度为 $O(N*N_s)$。为当前样本寻找每一类的 K 个近邻样本的时间复杂度为 $O(N*K)$，计算当前样本的 $diff(\cdot)$ 时间复杂度为 $O(K*C)$，因此，总的时间复杂度为 $O(N*N_s + M*(N*N_s + N*K + K*C))$。

因此，评价一个特征子集的时间复杂度大约为 $O(M*(N + K*C))$。假定所选的特征个数为 $S(S \le I)$，则搜索时间复杂度为 $O(I*S)$，因此，FSCRF 总的时间复杂度为 $O(I*S*M*(N + K*C))$

3.5 实验及结果分析

3.5.1 公共数据集实验设置及分析

为了检验 FSCRF 方法的特征选择性能，实验采用 12 个 UCI 公共数据集[159]对算

法进行测试，详细描述见表 3.1。

表 3.1 12 个 UCI 数据集特征描述

数据集	特征数	样本数	类别数
breast-cancer	9	286	2
Diabetes	8	768	2
heart-statlog	13	230	2
lung-cancer	56	32	3
Mushroom	22	8124	2
Optdigits	64	5620	10
Promoters	57	106	2
Sonar	60	208	2
Soybean	35	683	19
Splice	60	3190	3
Twonorm	20	7400	2
Waveform	40	5000	3

本节对 FSCRF 方法与 CFS 算法[62]、拉斯维加斯算法[94]（LVF，Las Vegas Filter）及 Relief 的扩展方法[152]在公共数据集上进行比较，所有方法均在 WEKA[160]系统中统一进行，其他 3 种方法均采用 WEKA 系统中已实现的方法。其中 CFS 方法是一种基于特征子集与类别相关性的特征选择方法；LVF 方法是一种检测特征子集不一致性的随机概率特征选择方法[94]；3 个方法中除 Relief 之外均采用最好优先搜索策略进行子集搜索。所有方法的参数都采用系统默认值，FSCRF 的参数 M 和 K 分别进行了调整，调整参数的原因之一是由于各个数据集的大小不一且各个类别的数据分布不均匀。此外，由于 Relief 方法是一种特征排序法[152]，不能直接给出特征子集，因此对 Relief 方法选取前 k 个特征构成特征子集，k 值等于相应数据集上 FSCRF 方法所选的特征个数，其中 Relief 的参数 M 和 K 与 FSCRF 相同。表 3.2 分别记录了 4 种特征选择方法所选特征的个数，并对选取的特征个数按数目越小越有效进行了比较（Relief 方法没有参与特征数目比较），其中 w 表示获胜，t 表示相当，l 表示失利。从表 3.2 中可以看出，FSCRF 方法在选择特征个数方面性能与 LVF 方法基本相当，但要稍好于 CFS 方法。

表 3.2 四种特征选择方法在 12 个 UCI 数据集上所选取的特征个数

数据集	CFS	LVF	Relief	FSCRF(*M*,*K*)
breast-cancer	5	8	Top 3	**3** (20,50)
diabetes	**4**	8	Top 6	6 (100,1)
heart-statlog	7	10	Top 6	**6** (30,8)
lung-cancer	11	4	Top 3	**3** (30,5)
mushroom	**4**	5	Top 11	11(50,65)
optdigits	38	**9**	Top 47	47 (30,100)
promoters	6	4	Top 3	**3** (35,50)
sonar	19	**14**	Top 34	34 (90,30)
soybean	22	**13**	Top 25	25 (85,5)
splice	22	10	Top 6	**6** (100,5)
twonorm	20	**9**	Top 20	20 (100,5)
waveform	15	**11**	Top 14	14 (15,50)
w/t/l	**2/5/5**	**5/4/3**	**-**	**5/3/4**

为了评价 FSCRF 选取特征子集的分类效果及泛化能力（所选特征能适应多个分类器）。本文选用两个非常经典的分类算法（决策树[161]和朴素贝叶斯[162]）用于实验比较。所有方法的参数均为系统默认参数值。

表 3.3 在原始数据集与各种特征选择方法所选特征子集上使用朴素贝叶斯的分类精度

数据集	Original set	CFS	LVF	Relief	FSCRF
breast-cancer	72.70±7.74	73.01±8.08	72.36±7.55	71.20± 7.61	**74.84±7.14**
diabetes	75.75±5.32	**77.06±4.70**	75.75±5.32	76.19± 4.79	76.24±4.96
Heart-statlog	83.59±5.98	**85.56±5.83**	84.07±5.83	82.11± 7.43	85.30±6.67
lung-cancer	65.10±27.79	70.80±23.05	63.08±23.58	69.67± 23.91	**78.75±21.98**
mushroom	95.76±0.73	98.52±0.45	98.52±0.44	96.79± 0.69	**99.65±0.22**
optdigits	91.39±1.04	91.56±1.13	80.94±1.40	91.67± 1.07	91.37±1.07
promoters	90.14±9.59	94.57±6.98	92.44±8.32	90.25± 8.84	94.48±7.25
sonar	67.71±8.66	67.62±9.23	66.09±9.15	69.68± 9.07	**73.20±8.84**
soybean	**92.94±2.92**	92.11±2.90	83.56±3.32	90.07± 3.34	92.45±3.06
splice	95.42±1.14	**96.16±0.95**	94.41±1.27	91.60± 1.65	93.53±1.43
twonorm	97.86±0.49	97.86±0.49	91.06±1.13	97.86± 0.49	97.86±0.49
waveform	80.01±1.45	80.10±1.40	80.75±1.59	80.34± 1.36	80.34±1.36
平均	84.00±6.07	85.41±5.43	81.92±5.74	83.95±5.85	**86.50±5.37**

对表 3.2 中各原始数据集以及各种方法所选特征构成的新数据集采用 10 次交叉验证法进行比较，并对每个数据集随机重复 10 次该过程，故对每个数据集共完成 100 次实验。表 3.3 和表 3.4 记录了平均分类精度和标准差。

从表 3.3 和表 3.4 可以看出，在大多数情况下，FSCRF 方法要比其他方法所选出的特征子集有更高的分类精度，且在决策树和朴素贝叶斯两个分类器下的表现均是如此，展现了 FSCRF 所选特征良好的分类能力和泛化能力。

表 3.4 在原始数据集与各种特征选择方法所选特征子集上使用决策树的分类精度

数据集	Original set	CFS	LVF	Relief	FSCRF
breast-cancer	74.28±6.05	72.94±5.47	72.51±6.32	67.22± 4.19	**75.30±5.39**
diabetes	74.49±5.27	74.38±5.04	74.49±5.27	74.48± 4.90	**75.14±5.11**
heart-statlog	78.15±7.42	81.41±7.04	79.11±7.49	81.33± 7.14	**84.89±6.78**
lung-cancer	47.17±20.35	57.83±24.75	59.42±24.21	55.58± 24.33	**77.50±24.60**
mushroom	100.00±0.00	99.02±0.34	100.00±0.00	100.00± 0.00	100.00±0.00
optdigits	90.52±1.20	90.75±1.13	82.42±1.40	90.55± 1.14	90.57±1.13
promoters	79.04±12.68	80.59±11.43	80.75±11.09	83.15± 12.27	83.15±2.27
sonar	73.61±9.34	**80.06±7.67**	79.55±8.03	76.02± 10.13	77.88±8.92
soybean	91.78±3.19	90.42±3.37	84.15±3.33	88.75± 3.29	91.10±2.67
splice	94.17±1.28	**94.30±1.34**	93.89±1.41	91.39± 1.38	93.76±1.34
twonorm	84.95±1.16	84.95±1.16	83.86±1.43	84.93± 1.25	84.95±1.16
waveform	75.25±1.90	76.97±1.82	76.35±2.01	76.93± 2.00	76.97±1.82
平均	80.28±5.82	81.97±5.88	80.54±6.00	80.86±6.00	**84.27±5.93**

同时从表 3.3 中发现只有 CFS 和 FSCRF 方法的平均分类精度超过了在原始数据集上的分类精度，其余方法的平均分类精度均低于原始特征集的平均分类精度。在表 3.4 中的其他方法所选特征子集的分类精度跟原始特征集的分类精度相当，而 FSCRF 所选特征子集的平均分类精度却比原始数据集的分类精度高近 4 个百分点（84.27 % VS 80.28 %）。由此表明，本章所提出的方法不但能够有效地剔除原始特征中的冗余特征和“脏”特征，而且能够选出具有强分类能力的特征子集。

3.5.2 FSCRF与Relief方法性能比较

为了验证本章开头提到的一个观点——“由前 k 个类别区分能力强的特征组成的

特征子集类别区分能力不一定强”。特意对 FSCRF 方法与 Relief 方法单独进行了比较。

先用 Relief 方法对每个数据集中的各个特征按类别区分能力从大到小排序，然后选取 4 个特征子集构成 4 个新的数据集，其中每个新的数据集的特征由前 k 个特征构成，k 的取值分别为 3、5、8 以及 FSCRF 在该数据集上所选特征个数。测试方法与 3.5.1 节方法相同。其中 FSCRF 和 Relief 参数值均设置为：M=20，K=1。

表 3.5 在 Relief 与 FSCRF 所选特征子集上使用朴素贝叶斯的分类精度

数据集	Relief_Top3	Relief_Top5	Relief_Top8	Relief_Equal	FSCRF
breast-cancer	69.79±6.93	69.38±6.49	**73.12±8.00**	69.38±6.49	66.70±6.36
diabetes	68.83±4.09	77.58±4.47	75.75±5.32	**77.58±4.47**	74.65±5.43
heart-statlog	71.85±8.21	76.07±7.67	78.48±7.70	77.63±8.05	**81.00±7.51**
lung-cancer	**60.83±27.23**	55.00±25.98	52.50±26.89	54.83±27.78	56.92±27.42
mushroom	**98.52±0.45**	97.64±0.47	98.33±0.40	95.80±0.76	96.30±0.70
optdigits	44.09±1.87	60.94±1.61	72.11±1.80	**91.98±1.06**	91.53±1.12
promoters	69.12±14.53	74.83±13.08	83.45±10.37	74.83±13.08	**90.81±8.01**
sonar	70.99±10.04	69.66±9.86	**71.77±9.37**	70.70±9.22	65.27±9.15
soybean	57.95±3.09	59.15±3.31	74.77±4.56	**89.76±3.22**	89.69±3.24
splice	60.68±1.71	72.51±1.92	73.52±1.98	72.65±2.01	**88.93±1.70**
twonorm	78.39±1.43	84.80±1.06	89.62±0.94	**95.13±0.76**	94.79±0.90
waveform	70.33±1.85	76.11±1.84	78.27±1.74	79.81±1.43	**80.26±1.39**
平均	68.45±6.79	72.81±6.48	76.81±6.59	79.17±6.53	**81.40±6.08**

表 3.5 和表 3.6 分别记录了两分类器对各个数据集测试的平均分类精度和标准差。从实验结果可以看出，FSCRF 方法所选特征子集的分类精度在多数情况下要比 Relief 方法所选子集的精度要高。使用朴素贝叶斯分类器的对 12 个 UCI 数据集的平均分类精度达到了 81.4%，比选取相同数目的 Relief 精度高 2 个百分点，而使用决策树的平均分类精度，FSCRF 比 Relief 方法高近 5 个百分点。因而从某种程度上说明了 FSCRF 方法所选的特征子集比 Relief 方法中由前 k 个特征构成的特征子集对原始特征具有更强的表征能力。

表 3.6 在 Relief 与 FSCRF 所选特征子集上使用决策树的分类精度

数据集	Relief_Top3	Relief_Top5	Relief_Top8	Relief_Equal	FSCRF
breast-cancer	68.19±4.92	68.23±5.45	**74.32±6.06**	68.23±5.45	70.37±4.69
diabetes	67.28±3.93	74.06±5.16	74.49±5.27	**74.06±5.16**	73.87±5.25
heart-statlog	73.11±8.55	76.56±8.04	75.11±7.86	75.22±7.13	**81.85±7.03**
lung-cancer	42.42±21.97	39.75±22.81	33.83±19.53	37.17±22.39	**50.42±21.53**
mushroom	99.51±0.23	99.90±0.11	100.00±0.00	100.00±0.00	100.00±0.00
optdigits	50.79±1.98	67.76±1.91	78.73±1.84	**90.63±1.11**	90.56±1.06
promoters	62.02±14.57	64.05±13.69	71.35±12.94	64.05±13.69	**80.65±11.03**
sonar	68.54±8.86	70.33±9.46	72.28±9.72	**75.82±9.83**	72.47±9.10
soybean	60.32±3.05	66.04±3.85	78.91±3.70	**89.98±3.26**	89.03±3.25
splice	60.63±1.70	73.44±2.07	74.29±2.13	73.53±2.06	**89.40±1.58**
twonorm	76.79±1.30	80.81±1.42	83.04±1.43	**84.96±1.10**	84.60±1.36
waveform	68.45±2.12	73.29±1.69	74.57±1.70	76.46±1.86	**76.76±1.93**
平均	66.50±6.10	71.18±6.31	74.24±6.02	75.84±6.09	**80.00±5.65**

3.5.3 实测数据集实验设置及分析

实测数据为华中科技大学附属医院所采集的老年痴呆数据，该数据集共收录样本数据 134 个，每个样本共有 54 项数据，其中检测指标（特征）多达 53 项，第 54 项为痴呆级别，即脑功能严重程度，痴呆级别共分为 4 级（类别数目）。具体见表 3.7。

对 FSCRF 方法与其他 3 种特征选择方法在老年痴呆数据集上进行评价。其中 FSCRF 与 Relief 的参数值均被设为 M=120，K=3，Relief 方法的特征子集由前 10 个特征组成。所有方法选择的特征个数及索引见表 3.8，从结果可以看出，FSCRF 方法最终选出了 10 个有效特征，比 LVF 方法多选两个特征，比其他方法少选两个特征。

为了进一步检验 FSCRF 方法所选特征子集的实际诊断效果，并考虑到医生对诊断模型的可接受性，本文使用文献[163]中提到的逻辑回归模型对老年痴呆数据的原始特征以及 4 种特征选择方法所选特征子集做统一测试。

在实验样本数据集中，80%数据作为训练集，20%作为测试集。图 3.3 给出了 10 次实验的平均诊断精度。

表 3.7 老年痴呆数据特征描述

序号	特征	序号	特征	序号	特征
1	时间定向能力减退	19	地点定向能力减退	37	记忆新知识的能力减退
2	运用知识能力减退	20	判断力减退	38	计算能力
3	逻辑\推理	21	找词困难	39	不能解释成语、谚语
4	命名障碍	22	复述障碍	40	不能临摹图形
5	不能拼图	23	失用	41	懒散、淡漠
6	睡眠障碍	24	情绪障碍-兴奋	42	生活自理能力
7	社交能力	25	工作能力	43	符合体征的改变
8	起病特点	26	人、物定向能力减退	44	回忆远期知识困难
9	虚构	27	注意力分散	45	概括能力
10	左右失认	28	问题解决（计划、组织）	46	主动性
11	各种失语症	29	空洞语言	47	赘述
12	错语	30	语言不连贯	48	阅读理解困难
13	朗读不能	31	重复说话	49	语音减低
14	构音障碍	32	缄默	50	不能做结构性作业
15	不能摆积木	33	失认	51	敌意增加
16	幻觉、妄想	34	情绪障碍-抑郁	52	行为障碍
17	人格障碍	35	局灶体征	53	病理症
18	脑萎缩	36	白质病变	54	**脑功能**

从图 3.3 可以看出，FSCRF 方法所选的特征子集相比其他方法所选特征子集的平均分类精度都高。

表 3.8 每一个特征选择方法在老年痴呆数据集上所选取的特征个数及名称

方法	特征数	各方法所选特征名称
Original Set	53	所有特征
CFS	12	时间定向，地点定向，记忆新知识，运用知识，注意力， 判断力，计算，概括，主动性，找词，语言不连贯，阅读理解
LVF	8	记忆新知识，注意力，计算，概括，成语、谚语，摆积木， 睡眠，电脑操作
Relief	10	运用知识，时间定向，判断力，注意力，地点定向， 计算，找词，记忆新知识，远回忆，阅读理解
FSCRF	10	时间定向，地点定向，运用知识，注意力，判断力， 计算，错语，语言不连贯，重复说话，结构作业

与原始数据相比，FSCRF 选取的特征数目由原来的 54 个减少到 10 个，而平均分类精度不仅没有降低，反而从 64.17%提高到 73.90%。从而，病人的检查种类可以由 53 项缩减为 10，可以大大减少病人许多没有必要的检查，节省人力和物力，这点对痴呆患者尤其重要！

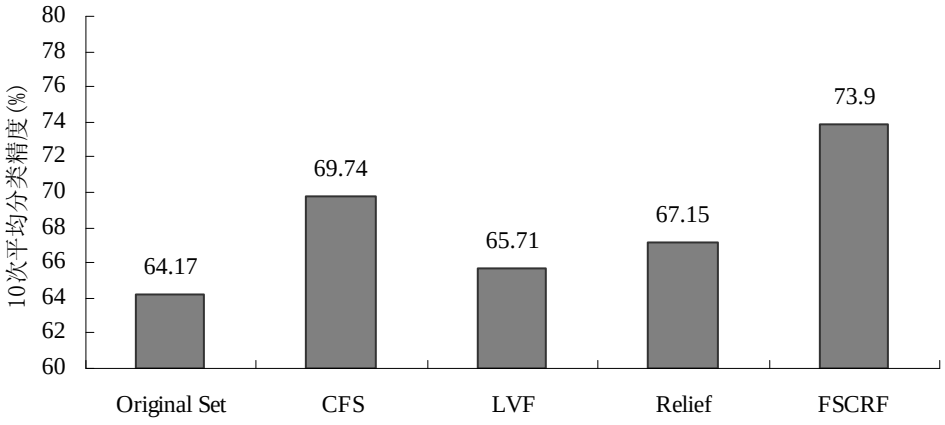

图 3.3 老年痴呆数据集上逻辑回归分类精度

此外，为了方便医生参考使用，本文专门利用 FSCRF 方法所选的 10 个特征和所有数据建立了评价老年痴呆程度的逻辑回归模型，得到 4 个疾病级别的回归表达式 $F_1(x)$、$F_2(x)$、$F_3(x)$、$F_4(x)$分别如下：

$$F_1(x)=2.53-0.95*[\text{时间定向}]-0.86*[\text{地点定向}]-4.32*[\text{运用知识}]-0.39*[\text{判断力}] \\ -1.65*[\text{计算}]-2.75*[\text{重复说话}]-2.54*[\text{结构作业}] \quad (3.4)$$

$$F_2(x)=1.25-0.72*[\text{运用知识}]-2.23*[\text{错语}]-3.16*[\text{语言不连贯}]-11.24*[\text{重复说话}] \\ +0.36*[\text{结构作业}] \quad (3.5)$$

$$F_3(x)=-4.37+1.16*[\text{时间定向}]+0.89*[\text{运用知识}]+0.87*[\text{注意力}]+0.13*[\text{判断力}] \\ -0.89*[\text{错语}]+1.02*[\text{语言不连贯}]+3.78*[\text{结构作业}] \quad (3.6)$$

$$F_4(x)=-9.96+1.35*[\text{时间定向}]+3.68*[\text{地点定向}]+6.58*[\text{注意力}]+3.88*[\text{计算}] \\ +5.03*[\text{语言不连贯}] \quad (3.7)$$

由逻辑回归理论可知，对于待诊病人 x，可以用式(3.8)计算的结果作为该患者老年痴呆程度的评价。

$$c(x) \leftarrow \max_i \left\{ p_i(x) \mid p_i(x) = e^{F_i(x)} \Big/ \sum_{i=1}^{4} e^{F_i(x)},\ i = 1,2,3,4 \right\} \tag{3.8}$$

由于(3.8)式只需比较概率大小，为避免复杂的指数运算，可以将问题转换成求解 4 个回归表达式中的最大值，因而只需按照(3.9)式计算即可。

$$c(x) \leftarrow \max_i F_i(x),\ i = 1,2,3,4 \tag{3.9}$$

然后将结果作为诊断脑功能严重程度的参考意见，该模型因此变得简单实用。

3.6 本章小结

本章提出了一种能有效选出具有强类别区分能力的特征子集新方法 FSCRF 方法。并与其他 3 种有代表性的特征选择方法进行了大量的实验比较与分析。在 12 个 UCI 数据集上就特征约减能力比较的结果表明，FSCRF 方法在大多数情况下能选出更少的特征；在此基础上，使用决策树和朴素贝叶斯分类器就所选特征子集的分类能力进行比较，实验结果显示，FSCRF 所选的特征子集在大多数情况下具有较高的分类精度。此外，在一个老年痴呆实测数据集上使用逻辑回归分类方法对各种特征选择方法进行实验比较表明，本章所提的方法对实测数据在减少特征数目、提高分类精度两方面均有实际效果，并给出了逻辑回归诊断模型的详细参数。

4 协同训练半监督分类

自标准协同训练算法提出以来，许多研究者对其加以改进和扩展，提出了各种改变学习策略、弱化限制条件的算法，推动了协同训练的理论分析和应用研究，使协同训练成为半监督分类中一个非常重要的分支。本章从改变学习策略、充分利用有限标记样本两方面对协同训练半监督分类方法展开研究。

4.1 协同训练

Blum 和 Mitchell 在 1998 年首次提出了协同训练（Co-T，Co-Training）算法[82]。具体算法如图 4.1 所示。它是半监督学习的另一种学习方式，通常被称为标准的协同训练算法。该算法假设数据集有两个充分冗余的特征视图，即特征集有两个冗余的特征子集，每个特征集都足以训练一个学习器。协同训练算法[82]利用标记样本在这两个冗余的视图上对分类器进行训练，并对未标记数据进行预测，然后从每个分类器的预测结果中选取信任度较高的样本添加到另一个分类器所对应的训练样本集中，通过这种方式，使得另一组的训练样本不断增加。

Blum 和 Mitchell[82]已经从理论上对标准协同训练方法进行了证明，当存在两个充分冗余视图这个假设前提成立时，分类器的分类性能可以通过标准的协同训练算法得到有效提高。在上述假设前提成立的情况下，Dasgupta 等人[164]也给出了类似的结论，如果最大化两个分类器对未标记数据预测的一致性，将得到其最小泛化误差。Balcan 等人[165]进一步将标准协同训练的假设条件弱化，认为数据分布只要满足“扩展性”（Expansion）假设，协同训练算法就可以展现其优势，并从理论上分析了协同训练算法的 PAC 假设空间的[29]可学习性。

许多学者对标准协同训练方法提出了改进方法。Abney[167]通过使未标记样本在两视图上预测的不一致性达到最小，来达到改善算法整体性能的目的。Nigam 和 Ghani[124]在两个视图上分别使用 EM 算法进行预测，得到了比标准协同训练方法更好的分类结果。Brefeld 和 Scheffer[166] 提出了 Co-EM 算法，其主要思路是将 EM 算法和 SVM 结合起来在两个视图上进行训练。

算法：[Co-Training(Co-T)]

输入：标记样本训练集 L

未标记样本训练集 U

处理：创建 U'样本池，从 U 中随机选出 u 个样本组成

循环 k 次：

Setp1. 使用 L 训练一个分类器 h_1，仅仅考虑 x 中的部分样本 x_1

Setp2. 使用 L 训练一个分类器 h_2，仅仅考虑 x 中的部分样本 x_2

Setp3. 让 h_1 去标记 U'中的 p 个阳性和 n 个阴性样本

Setp4. 让 h_2 去标记 U'中的 p 个阳性和 n 个阴性样本

Setp5. 把这些标记好的样本加入到原有的标记样本 L 中

Setp6. 再从 U 中随机选出 $2*p+2*n$ 个未标记样本补充 U'

输出：未标记样本训练集 U 的类别

图 4.1 标准协同训练算法（Co-Training）

Goldman 和Zhou[168]用两种不同的有监督学习方法，从同一个特征集上学习两个不同的分类器，而样本空间可以被每个分类器划分成多个等价类，未标记样本的类别标记通过采用交叉验证来完成。最后，通过利用交叉验证方法综合两种学习方法形成最终预测。这种方法由于对每个样本的标记都需要采用交叉验证策略，因此时间复杂度较高；Zhou 和 Li[127]利用重采样技术提出了 Tri-Training 算法，该算法在不同的数据集上协同训练三个分类器，如果其中任意两个分类器的的预测结果一致，就把数据及其类别标记加入到第三组训练数据中。Goldman 和 Li 所提出的这两种算法的一个共同特点是：只用到一个视图，并通过改用不同的分类器或者使用不同的训练数据来实现协同训练。

另一种类协同训练的算法，迭代交叉训练算法[169]（ICT，Iterative Cross-Training）由 Soonthornphisaj 和 Kijsirikul 提出，该算法最终由两个迭代训练的分类器合并而成，但该方法每次训练的标记样本只用了总的标记样本的 1/2，标记样本训练不太充分。

本章将在上述基础上提出一种新的协同训练方法，新交叉训练法（NCT，New

Cross-Training）。它不需要假设存在多个冗余视图，而且除了能充分利用未标记样本外，还能充分利用标记样本的信息。

4.2 新交叉训练半监督分类方法NC-T

本节将给出新的交叉训练方法 NC-T 的具体过程。图 4.2 给出 NC-T 方法的整体框架。假定所有的标记样本记为 L，所有的未标记样本记为 U。并把 L 和 U 这两部分数据分别分成三份，记作 L_1，L_2，L_3 以及 U_1，U_2，U_3。标准的 ICT 方法的做法是将标记训练样本和未标记训练样本分成两份，而不是三份。

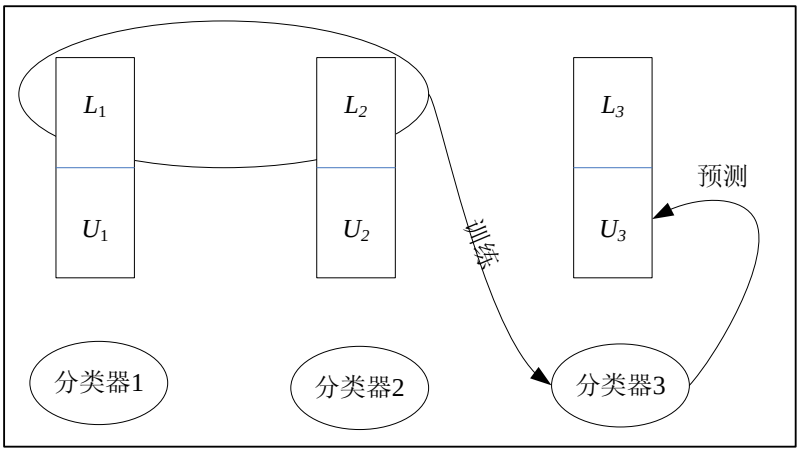

图 4.2 NC-T 方法示意图，其中含有三个分类器，三组训练数据，每组分别有标记样本数据 L_i 和未标记样本 U_i 组成。其中箭头表示训练流程走向。

假定除了两个分类器以外，还有第三个分类器。三个分类器分别记为分类器 1、分了器 2、分类器 3。三个分类器均使用相同的学习算法，比如 RBF 神经网络。对任意一个分类器都使用其中两组标记数据进行训练，假定当前选中 L_1 和 L_2 进行训练，如图 4.2 所示。然后用该训练好的分类器 3 去预测未标记数据 U_3 的数据，并选中其中信任度最高的前 k 个样本连同预测值（类别）加入到 L_3 中，从而 L_3 的训练样本逐渐增加。类似地，替换其他两组训练数据开始新一轮的训练，再完成上述类似工作，直到所有的未标记样本标记完毕。可以看出 NC-T 方法在每一轮训练中用到了标记样本的 2/3 左右。

与其他协同训练方法相比，NC-T 方法最重要的一点是，能做到在每一轮的训练中有更多的标记样本被用于训练，从而可以缓解基分类器在每轮训练过程中标记样本不足所带来的问题，此外也无需假设存在多个独立的特征子集。

算法：[New Cross-Training(NC-T)]

输入：

— L　原始标记样本

— U　未标记样本

— *Classifier*　学习算法

初始化：

(1) N_p=3　数据划分数

(2) k=10　每轮训练信任度最高的数目

Step 1. 将标记样本 L 和未标记样本 U 分别近似均分三份 L_1、L_2、L_3 以及 U_1 、U_2、U_3，并保证每份标记样本按类别近似均分。

Step 2. 重复执行，直到 U_1、U_2、U_3 中所有的数据分类标记完毕。

(1) 按顺序合并其中的两组标记标记样本 L_i 和 L_j

(2) 在 L_i 和 L_j 上训练一个分类器

(3) 预测第三组数据中未标记样本 U_k

(4) 从 U_k 中选出信任度最高的 k 个训练样本连同类别加入 L_k 中

(5) 从该未标记样本中删除 k 个已标记样本

Step 3. 构建一个新的分类器，利用所有的标记样本 L，包括新标记的样本。

(1) 合并所有的训练样本，组成一个新的训练集

(2) 新的训练集上训练一个分类器

输出：分类器

图 4.3 一种新的交叉训练法（New Cross-Training）

4.3 算法流程

NC-T 可以看做是交叉训练 ICT 方法的一种新的扩展。算法的具体流程见图 4.3。值得注意的是，在 NC-T 中，所训练的三个分类器是不同的，原因是由于他们的训练集不同。这种策略能有效避免 NC-T 方法退化为自训练（Se-T，Self-Training 方法（仅仅使用一个分类器）[124]。而 Co-training 方法则在两个足够冗余的特征视图上分别训练两个分类器，它对网页分类问题确实提供了有效地解决方案，因为基于网页本身的视图和网页超级链接的视图是相对独立的，即单个视图都能有效地描述一个网页的相关性质。但是对于很多实际的数据挖掘问题却很少存在这样自然的独立视图。Goldman 和 Zhou[168]对此进行了一些扩展，不需要假定特征视图存在足够冗余，且不同的分类器由不同类型的学习方法训练完成。而 NC-T 方法既不需要假定存在独立视图，也不需要使用不同的学习方法，三个分类器使用相同类型的学习方法（如 RBF 神经网络）在不同的数据集上训练，就可以得到参数完全不同的三个分类器。而且能比 ICT 类型的方法更能充分的利用标记样本的信息。

4.4 时间复杂度分析

假定以决策树作为基分类器，通常这是一个很普遍的选择，假定树的深度为 d，则建立每棵决策树的计算代价为 $O\left(d \cdot I \cdot n_l \log n_l\right)$，$I$ 为输入数据 x 的维数，将未标记样本标记完得计算代价为 $O\left(n_u / k \cdot d \cdot I \cdot n_l \log n_l\right)$，

因此，NC-T 半监督分类方法在伪标记样本与标记样本上训练的总的时间复杂度为 $O\left(n_u / k \cdot d \cdot I \cdot n_l \log n_l + d \cdot I \cdot (n_l + n_u) \log(n_l + n_u)\right)$。

4.5 实验及结果分析

本节使用 12 个 UCI[159]数据集对 NC-T 方法与其他 3 种半监督学习方法进行实验比较，包括交叉训练法 ICT，协同训练法 Co-T，自训练法 Se-T。

4.5.1 实验设置

用于实验分析的 12 个数据集详细描述见表 4.1。对每个数据集，25%的数据用作测试集，其余的用作训练集，训练集按照一定比例划分成两部分：标记样本 L 和未标记样本 U。实验中标记比例有四种，20%、40%、60%、80%。对某个特定的训练集，L 又被分成三部分 L_1、L_2、L_3。类似地，U 被分成 U_1、U_2、U_3。比如，一个数据含有 1000 个样本，则 250 个样本用于测试，750 个样本用于训练，如果标记比例为 20%，则用于训练的标记样本 L 的数目为 150，未标记样本 U 的数目为 600，进一步细分，L_1、L_2、L_3 的数目都是 50，U_1、U_2、U_3 的数目均为 200。实验中分类方法采用 J4.8 决策树[170]和径向基神经网络（RBFNN，Radial Basis Function Neural Networks）[36]，两方法在实验中被独立运行，参数均采用 WEKA[160]系统中默认参数。

表 4.1 12 个 UCI 实验数据特征描述

编号	数据集	样本数目	特征数目	类别数目	各类百分比%
1	bupa	345	6	2	42:58
2	colic	368	22	2	63:37
3	diabetes	768	8	2	65:35
4	german	1000	20	2	70:30
5	hypothyroid	3163	25	2	5 :95
6	ionosphere	351	34	2	36:64
7	iris	150	5	3	33:33:33
8	kr-vs-kp	3196	36	2	52:48
9	sick	3772	29	2	6 :94
10	tic-tac-toe	958	9	2	65:35
11	vote	435	16	2	61:39
12	wdbc	569	30	2	37:63

NC-T 和其他三个用于比较的半监督方法均在 WEKA 中实现。实验中迭代交叉训练的 ICT 方法与原有的 ICT 方法稍微有些不一样，实验中使用两个相同类型的分类器，代替原来不同类型的分类器。Co-T 方法将数据特征随机的分成两个数量相当的视图。Se-T 方法中三个分类器均在三个不同的标记样本集上训练和更新标记样本，并最终训练出三个分类器，最终在测试集上采用选举法进行预测。

4.5.2 实验结果及分析

为了能够较充分地验证 NC-T 方法的有效性，实验对 12 个训练数据集均使用 4 种不同的标记比例对样本进行标识，分别为 20%、40%、60%、80%。表 4.2 与表 4.3 记录了这四种情况中使用 RBF 神经网络作为分类器的平均分类精度。

从整体上来看，NC-T 方法的分类精度要高于 Co-T 方法。其原因应该是：对 Co-T 方法而言，12 个数据集中每个数据集很有可能不存在两个自然独立的特征视图，从而导致分类精度普遍相对较低。与 ICT 和 Se-T 相比，NC-T 方法分类精度相当。原因有可能是 NC-T 方法中三个分类器在不同（类型一样，样本不一样）的训练集上进行训练，但三个分类器在数据类分布相当的情况下，有可能训练出来的模型非常相似，从而退化到 Se-T 方法。

表 4.2 使用 RBF 神经网络作为分类器训练后测试的平均精度，其中测试集占整个数据集的 25%，剩余 75%用于做训练集，记录了训练集的标记比例为 20%和 40%两种情况下的平均精度

Data sets	20% label rate				40% label rate			
	NC-T	ICT	Co-T	Se-T	NC-T	ICT	Co-T	Se-T
bupa	62.1	56.3	60.1	**65.4**	49.4	56.3	56.7	**62.4**
colic	**83.7**	79.4	74.1	71.9	76.1	**77.2**	74.3	76.8
diabetes	**72.4**	68.0	71.8	71.8	69.8	69.6	71.7	**72.2**
german	59.6	69.6	**70.5**	69.6	70.8	**74.0**	67.2	66.8
hypothyroid	**95.7**	94.9	89.0	89.6	**95.3**	94.5	93.6	93.3
ionosphere	**88.9**	83.9	83.9	80.2	87.8	**90.8**	83.1	78.6
iris	90.5	**94.7**	88.2	88.3	**95.2**	92.1	90.3	88.1
kr-vs-kp	72.3	70.1	**88.2**	75.5	78.0	74.7	85.2	**94.4**
sick	**96.4**	94.1	93.4	94.6	96.2	96.3	92.0	92.7
tic-tac-toe	66.0	67.7	72.4	**75.4**	73.1	68.5	71.6	73.5
vote	89.8	85.3	92.2	**93.7**	91.7	89.0	91.4	**92.3**
wdbc	91.6	**95.0**	89.2	90.2	92.3	92.9	91.5	92.5

表 4.4 与表 4.5 给出了 J48 决策树作为分类器在 4 种不同的标记比例下的平均分类精度，数据集仍然为 12 个 UCI 数据。从表 4.4 与表 4.5 可以看出，4 种方法的分类精度仍然是 NC-T 方法的分类精度要高于 Co-T 方法，且与 ICT 和 Se-T 方法相当。

说明了包括 NC-T 方法在内的四种半监督方法对不同基分类器具有良好泛化能力。

表 4.3 使用 RBF 神经网络作为分类器训练后测试的平均精度，其中测试集占整个数据集的 25%，剩余 75%用于做训练集，记录了训练集的标记比例为 60%和 80%两种情况下的平均精度

Data sets	60% label rate				80% label rate			
	NC-T	ICT	Co-T	Se-T	NC-T	ICT	Co-T	Se-T
bupa	59.8	50.6	55.4	**61.5**	60.9	59.8	61.2	**62.3**
colic	82.6	82.6	72.1	79.0	79.4	**84.8**	78.2	74.3
diabetes	75.0	**76.8**	67.7	71.5	**77.6**	76.8	70.8	72.4
german	68.8	**72.4**	70.1	65.9	70.8	**76.4**	70.3	70.4
hypothyroid	94.8	94.4	93.1	93.9	**96.7**	94.8	93.8	94.0
ionosphere	90.0	**93.1**	89.5	83.5	88.9	**92.0**	91.1	86.9
iris	**97.6**	97.4	91.5	88.3	**97.6**	94.7	88.5	84.3
kr-vs-kp	84.9	80.5	89.1	**92.6**	85.7	79.6	89.1	**95.1**
sick	96.5	96.7	92.1	92.9	**96.8**	96.5	91.1	92.2
tic-tac-toe	69.3	71.0	73.6	**93.8**	71.0	69.3	76.4	**93.8**
vote	**91.7**	90.8	90.8	90.8	92.6	91.7	90.5	**93.6**
wdbc	93.0	**95.0**	88.2	89.2	**95.8**	95.0	92.3	90.8

表 4.4 使用 J48 决策树作为分类器训练后测试的平均精度，其中测试集占整个数据集的 25%，剩余 75%用于做训练集，记录了训练集的标记比例为 20%和 40%两种情况下的平均精度

Data sets	20% label rate				40% label rate			
	NC-T	ICT	Co-T	Se-T	NC-T	ICT	Co-T	Se-T
bupa	**70.1**	57.5	56.7	62.0	62.1	**67.8**	60.2	62.4
colic	88.0	88.0	81.2	82.6	85.9	**88.0**	80.1	83.3
diabetes	70.3	69.6	**73.4**	70.0	68.2	73.2	73.3	71.2
german	61.6	66.0	66.3	66.3	67.6	**71.2**	68.7	66.7
hypothyroid	98.0	**99.3**	98.4	98.8	99.3	99.5	96.7	99.0
ionosphere	**87.8**	83.9	85.6	85.6	85.6	70.1	89.7	**90.9**
iris	**88.1**	68.4	86.8	84.3	**92.9**	86.8	87.5	85.3
kr-vs-kp	95.5	92.9	87.7	**97.8**	**96.9**	96.1	92.3	96.0
sick	**97.9**	96.1	94.9	97.7	97.9	**98.2**	95.8	96.2
tic-tac-toe	**70.6**	69.2	70.5	73.1	74.4	67.7	74.8	**75.2**
vote	91.7	91.8	92.6	**93.0**	91.7	91.7	**94.5**	94.1
wdbc	88.1	**93.6**	91.8	90.8	89.5	90.8	92.5	**92.7**

表 4.5 使用 J48 决策树作为分类器训练后测试的平均精度，其中测试集占整个数据集的 25%，剩余 75%用于做训练集，记录了训练集的标记比例为 60%和 80%两种情况下的平均精度

Data sets	60% label rate				80% label rate			
	NC-T	ICT	Co-T	Se-T	NC-T	ICT	Co-T	Se-T
bupa	56.3	52.9	**64.2**	60.5	58.6	54.0	64.8	64.7
colic	**88.0**	84.8	79.4	79.7	84.8	84.8	81.9	82.2
diabetes	76.0	76.8	72.3	71.8	**75.5**	73.7	71.0	73.3
german	**71.6**	70.4	67.7	68.4	**76.0**	72.8	68.0	69.5
hypothyroid	99.5	99.6	97.6	97.4	99.3	99.6	98.5	98.8
ionosphere	86.7	90.8	87.2	90.5	87.8	**92.0**	89.7	88.3
iris	**95.2**	81.6	91.4	90.2	**95.2**	93.5	93.1	92.5
kr-vs-kp	**99.0**	98.8	91.7	97.7	98.8	99.3	90.6	99.1
sick	98.3	98.6	94.9	96.8	98.6	98.7	96.0	98.7
tic-tac-toe	74.8	72.1	77.1	**84.2**	81.9	77.7	75.1	**84.2**
vote	91.9	92.7	91.3	**94.5**	92.0	92.9	93.6	**96.3**
wdbc	90.2	91.5	**94.3**	91.2	94.4	92.9	**95.8**	93.9

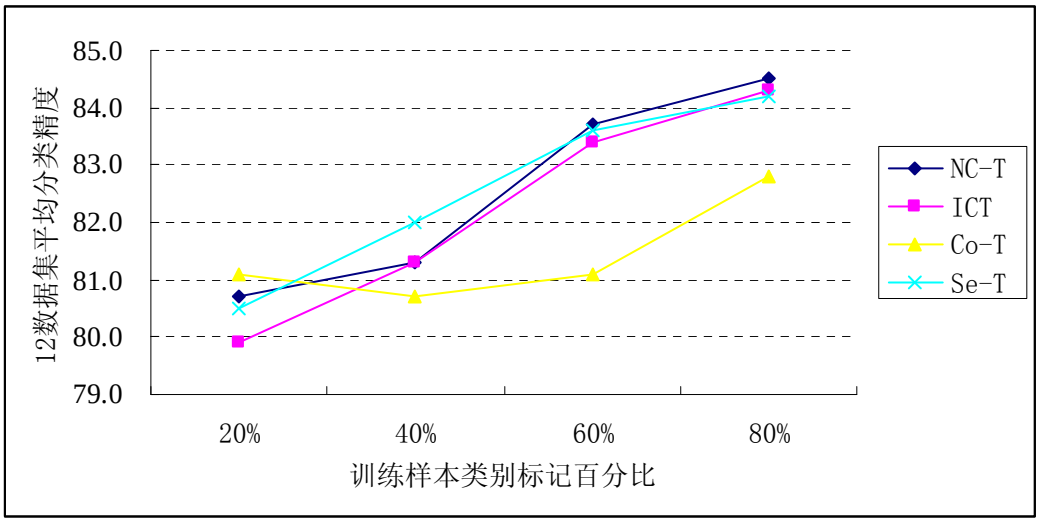

图 4.4 四种方法（每种方法采用 RBF 神经网络作为分类器）对 12 个 UCI 数据所获得的平均分类精度（%），其中每个数据集均用 4 种不同标记比例来划分有标记样本和未标记样本

但是，从表 4.2 可以看出，采用 J48 决策树作为分类器的分类精度要普遍高于采用 RBF 神经网络作为分类器的分类精度，这得益于 J48 决策树良好的泛化能力，而且相比 RBF 神经网络其速度要快很多（文中没有单独记录运行时间）。由此可以得出启示，在设计这类型半监督分类方法时，要想获得更高的分类精度，基分类方法的选取至关重要。

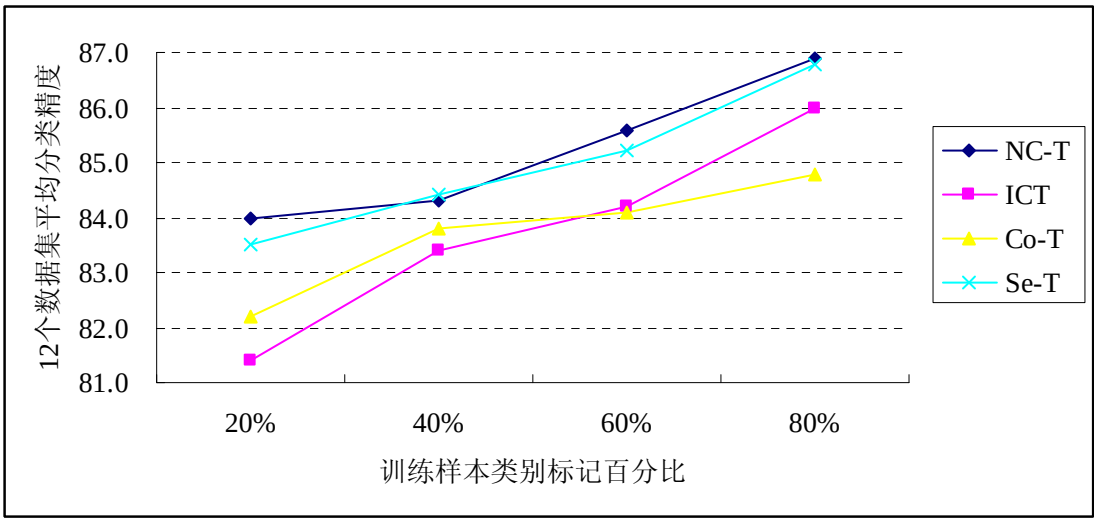

图 4.5 四种方法（每种方法采用 J48 决策树作为分类器）对 12 个 UCI 数据所获得的平均分类精度（%），其中每个数据集均用 4 种不同标记比例来划分有标记样本和未标记样本

为了更加清晰地反映出 NC-T，ICT，Co-T，Se-T 方法的整体性能，本节用图 4.4 和图 4.5 展示四个方法在 12 个 UCI 数据集上的平均分类精度。图 4.4 记录了采用 RBF 神经网络作为分类器的平均分类精度，图 4.5 则是采用 J48 决策树作为分类器的平均分类精度。

整体上来说，随着标记样本的增多，4 种方法的分类精度都有较大提升。但是在标记样本较少的情况下，比如 20%的标记比例，4 种方法也可以获得较为满意的分类精度（相比 80%的标记比例所获得分类精度）。因此，在标记样本很难以获取时候，采用所列举的 4 种方法也能获得较好的效果。当然，能够获得更多的标记样本，对改善分类性能确有帮助，但这并不是说标记样本越多就分类性能就会越好。

4.6 本章小结

本章提出了一种新交叉训练分类方法 NC-T，该方法不需要额外的假设前提，具有较好的实用性。并与其他三种类似方，包括交叉训练法 ICT，协同训练法 Co-T，自训练法 Se-T。在 12 个 UCI 数据集上进行了比较测试。实验结果表明，利用 NC-T 方法可以有效地利用未标记样本数据信息获得较高的分类精度，当然该方法也存在一些不足，比如方法的稳定性不是太好，对于有些数据集分类效果要高于其他方法，有的则要低于其他方法。如何使得该方法对不同分类问题都具有良好分类能力，还有待进一步完善。

5 多分类器集成半监督分类

对于模式分类，往往存在这样一种现象，不同的分类方法常常得到不同的分类效果。目前还不存在一种方法能够对所有的分类问题都有良好的分类表现，甚至同一个方法对同类识别问题的对象也不一定都有好的分类结果，半监督分类同样如此。因此，为了提高分类方法的稳定性、可扩展性，多分类器集成学习成为了半监督分类方法又一个重要研究内容。

5.1 多分类器集成学习

自 20 世纪 90 年代以来，多分类器集成学习（Ensemble Learning）[171, 172]就引起了众多研究机器学习的专家学者的高度重视，并迅速成为了研究的热点方向之一。它被国际著名学者 T.G.Dietterich[173]列为机器学习四大研究方向之首。大量的研究成果表明，集成多个学习器的分类学习性能，形成一个综合系统，能显著的改善系统的整体性能，并在实际应用发挥着越来越重要的作用。

最早的集成学习方法应该是 BMA（Bayesian Model Averaging）[174]，它由 Bartels 在 1997 年提出，从这以后，集成学习方法就逐步引起了相关学者的关注[175]，并提出了许多集成学习的方法，其中研究最多的是 Boosting[176]、Bagging[177]方法。

Bagging 方法[177]是 Breiman 提出的一种机器学习方法，它采用有放回随机取样技术（Bootstrap）获得训练集，并利用这些数据生成集成学习的个体。其中集成个体间的差异是通过有放回随机采样技术来实现的。该方法对提高系统的稳定性有很好的帮助，不会因为训练集较小改变而导致模型有很大变化，该方法对决策树等不太稳定的学习方法有很好的改进效果。

自 Kearns 和 Valiant[178]于 1994 年提出了强学习与弱学习的概念后，Schapire 提出了 Boosting 方法[176]，它是一类集成方法的统称。1997 年，Freund 和 Schapire 提出的 AdaBoost 方法[179]就是 Boosting 方法中最为典型的方法。AdaBoost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的弱分类器，不断集成所训练的弱分类器，并最终构成一个分类能力强的强分类器。具体见图 5.1。

算法：[Adaptive Boosting Algorithm (AdaBoost)]

输入： N 个标记的样本 $<(x_1,y_1),...,(x_N,y_N)>$

弱学习方法 ***WeakLearn***

迭代次数 T

初始化：样本权重向量： $w_i^1 = D(i)$， i=1,…N。

For $t = 1,2,...,T$

Step1. 设置 $p^t = \dfrac{w^t}{\sum_{i=1}^{N} w_i^t}$

Step2. 给数据以权值 p^t，调用 ***WeakLearn***，返回一个预测模型 h_t： $X \to [0,1]$。

Step3. 计算 h_t 的误差： $\varepsilon_t = \sum_{i=1}^{N} p_i^t \left| h_t(x_i) - y_i \right|$。

Step4. 设置 $\beta_t = \varepsilon_t / (1-\varepsilon_t)$。

Step5. 更新权重向量 $w_i^{t+1} = w_i^t \beta_t^{1-|h_t(x)-y_i|}$。

输出： $h_f(x) = \begin{cases} 1 & \text{若 } \sum_{t=1}^{T} (\log 1/\beta_t) h_t(x) \ge \frac{1}{2} \sum_{t=1}^{T} \log 1/\beta_t, \\ 0 & \text{否则.} \end{cases}$

图 5.1 AdaBoost 算法

AdaBoost 方法作为一种非常重要的有监督学习方法而被广泛应用于人脸识别和文本分类。近年来，一些基于 Boosting 的半监督学习方法也引起了一些研究人员的关注。如 Buc 等人提出了 SSMBoost 方法[180]，该方法将边界（Margin）定义扩展到未标记样本数据中，并设计了一个与边界代价函数相对应的梯度下降算法。这个方法虽然能有效地利用未标记样本的信息，但是它要求基分类器是半监督类型的，因此适用范围有一定的局限性；Bennett 等人提出了 ASSEMBLE 方法[181]，它是一个基于 AdaBoost 框架的半监督学习方法，但不需要基分类器是半监督类型的分类方法。该方法通过指派给未标记样本一个伪标记类别（Pseudo-classes），然后利用标记样本和伪标记样本构建一个集成学习方法。它要求基分类器的分类精度要好于 1/2。

Mallapragada 等人[182]提出了另外一种用于半监督学习的 boosting 框架。但由于它是一个针对二分类问题设计的解决方案，因此在面对多分类问题时，就会遇到目前很多半监督学习方法（绝大多数原始设计都是解决二分类问题）遇到的问题，他们通常采用的解决办法是将一个多分类问题转换成多个独立的二分类问题。但是这种独立假设在很多情况下是不合理的，特别是当每一个样本只能属于唯一一类时，这种假设处理会导致类不平衡问题。

为了有效的利用大量未标记样本的信息并缓解上述方法中存在的相关问题，本文提出一种新的半监督分类方法，简称 SSMAB（Semi-supervised Multi-class Adaboost），详细描述见图 5.2。它是一种归纳方法（Inductive），即不仅能解决未标记样本的类别标记问题，而且训练的模型能预测未参与建模的测试数据。受 Zhu 等人提出的 Multi-class Adaboost 的启发[183]，SSMAB 使用一个多类指数代价函数，能有效地解决半监督学习中的多分类问题。Muti-class Adaboost 没有利用未标记样本的信息，因此在标记样本非常少的情况下，其优势难以发挥，而 SSMAB 能有效地缓解这个问题。SSMAB 方法与 ASSEMBLE[181]方法有类似之处，但是存在一些不同，前者对基分类器的分类精度只要求等于或大于 $1/K$（K 为类别，通常 $K \geq 2$），而后者要求不低于 1/2；其次，SSMAB 方法没有要求重采样所有的标记样本和伪标记样本，而是使用所有的数据；再次，在迭代更新中，对于错分类的数据，SSMAB 方法分别给予原始标记样本和伪标记样本不同的权重，以强调原始标记样本在学习过程中的重要性。

5.2 集成半监督分类方法SSMAB

5.2.1 相关定义

记整个训练数据集为 $D=\{x_1, x_2, ..., x_n\}$，由标记数据和未标记数据两部分组成。其中标记数据的数目为 n_l，未标记数据的数目为 n_u，且 $n=n_u+n_l$。标记样本的标记

记作 $Y^l=\{Y_1,Y_2,...,Y_{n_l}\}$，其中每一个输出 $Y=(y_1,y_2,...,y_K)^T$，由一个 K 维向量组成。且 y_k 由式(5.1)表示。

$$y_k=\begin{cases}1 & 若 c=k,\\ -1/(K-1) & 若 c\neq k.\end{cases} \tag{5.1}$$

式(5.1)中，$k\in(1,2,...,K)$，K 为数据类别数目。

为了提高基分类器的分类精度，SSMAB 方法沿用了 boosting 方法的思想，通过在每次迭代过程中使用基分类器形成一个新的分类模型。

在算法开始迭代之前，首先使用 K 近邻（K-NN，K-Nearest Neighbor）方法为每个未标记样本指定一个伪标记 Y^u。受 Muti-class AdaBoost 的启发，为标记样本和未标记样本定义了一个指数代价函数。SSMAB 的目的是想通过最小化代价函数 $L(Y^l,Y^u,H)$ 学习一个 H 集成分类器。

$$\begin{aligned}L(Y^l,Y^u,H)&=\sum_{i=1}^{n_l}\exp\left(-\frac{1}{K}Y_i^{l\mathrm{T}}H(x_i)\right)+\sum_{i=n_l+1}^{n_l+n_u}\exp\left(-\frac{1}{K}Y_i^{u\mathrm{T}}H(x_i)\right)\\&=\sum_{i=1}^{n_l}\exp\left(-\frac{1}{K}(y_1^iH_1(x_i)+y_2^iH_2(x_i)+...+y_K^iH_K(x_i))\right)+\\&\quad\sum_{i=n_l+1}^{n_l+n_u}\exp\left(-\frac{1}{K}(y_1^iH_1(x_i)+y_2^iH_2(x_i)+...+y_K^iH_K(x_i)\right)\\&=\sum_{i=1}^{n_l+n_u}\exp\left(-\frac{1}{K}Y_i^{\mathrm{T}}H(x_i)\right)\end{aligned} \tag{5.2}$$

Y_i^{T} 为 Y_i 的转置，$H(x):x\to R^K$ 表示前 M 次迭代后的集成分类器，其表达式为式(5.3)。

$$H^{(m)}(x)=H^{(m-1)}(x)+\sum_{m=1}^{M}\beta^{(m)}h^{(m)}(x) \tag{5.3}$$

式(5.3)中，$\beta^{(m)}\in R$ 为基分类器的权重系数，$h(x)$ 由 $P(x)$ 决定，它是一个 K 维向量。$P(x)$ 为诸如决策树的多类分类器。

$$h_k(x)=\begin{cases}1 & 若P(x)=k,\\ -1/(K-1) & 否则.\end{cases} \tag{5.4}$$

最小化目标函数(5.2)可以写成(5.5)。

$$\begin{aligned}&\min L(Y^l,Y^u,H^{(m)})\\ &s.t.H_1^{(m)}(x)+H_2^{(m)}(x)+...+H_K^{(m)}(x)=0\end{aligned} \tag{5.5}$$

将式(5.3)代入式(5.5)得式(5.6)。

$$\begin{aligned}&\arg\min_{\beta^{(m)},h^{(m)}}\sum_{i=1}^{n_l+n_u}\exp\left(-\frac{1}{K}Y_i^{\mathrm{T}}\left(H^{(m-1)}(x_i)+\beta^{(m)}h^{(m)}(x_i)\right)\right)\\ &=\arg\min_{\beta^{(m)},h^{(m)}}\sum_{i=1}^{n_l+n_u}w_i^{(m)}\exp\left(-\frac{1}{K}\beta^{(m)}Y_i^{\mathrm{T}}h^{(m)}(x_i)\right)\end{aligned} \tag{5.6}$$

式(5.6)中，$w_i^{(m)}=\exp\left(-\frac{1}{K}Y_i^{\mathrm{T}}H^{(m-1)}(x_i)\right)$。

为强调标记样本的重要性，给出了一个样本权重系数$\alpha>1.0$。标记样本和未标记样本在下一次迭代过程中的权重为式(5.7)。

$$w_i^{(m)}=\begin{cases}\alpha.w_i^{(m-1)}.\exp\left(\dfrac{-\beta^{(m)}}{(K-1)^2}\right) & if\ i\in\{i\mid c_i\neq P^{(m)}(x_i)\ i=1,2,...,n_l\}\\ w_i^{(m-1)}.\exp\left(\dfrac{-\beta^{(m)}}{(K-1)^2}\right) & if\ i\in\{i\mid c_i\neq P^{(m)}(x_i)\ i=n_{l+1},n_{l+2},...,n_l+n_u\}\end{cases} \tag{5.7}$$

式(5.6)求解$\beta^{(m)}$得到：

$$\beta^{(m)}=(K-1)^2/K\cdot\left(\log\left(\left(1-error^{(m)}\right)/error^{(m)}\right)+\log(K-1)\right) \tag{5.8}$$

式(5.8)中，$error^{(m)}$定义如式(5.9)。

$$error^{(m)}=\sum_{j=1}^{n_l+n_u}w_j^{(m-1)}\Bigg/\sum_{i=1}^{n_l+n_u}w_i^{(m-1)}\ j\in\{j\mid c_j\neq P^{(m)}(x_j),j=1...n_l+n_u\} \tag{5.9}$$

求解$\beta^{(m)}$：令$\beta^{(m)}>0$。则

$$
\begin{aligned}
L(Y^l, Y^u, H^{(m)}) &= \sum_{i=1}^{n_l} \exp\left(-\frac{1}{K} {Y_i^l}^{\mathrm{T}} H^{(m)}(x_i)\right) + \sum_{i=n_l+1}^{n_l+n_u} \exp\left(-\frac{1}{K} Y_i^{u\mathrm{T}} H^{(m)}(x_i)\right) \\
&= \sum_{i=1}^{n_l+n_u} w_i^{(m)} \exp\left(-\frac{1}{K} \beta^{(m)} Y_i^{\mathrm{T}} h^{(m)}(x_i)\right) \\
&= \sum_{i=1\&c_i=P(x_i)}^{n_l+n_u} w_i \exp\left(-\frac{\beta^{(m)}}{K-1}\right) + \sum_{i=1\&c_i\neq P(x_i)}^{n_l+n_u} w_i \exp\left(\frac{\beta^{(m)}}{(K-1)^2}\right) \\
&= \exp\left(-\frac{\beta^{(m)}}{K-1}\right)\left(\sum_{i=1}^{n_l+n_u} w_i - \sum_{i=1\&c_i\neq P(x_i)}^{n_l+n_u} w_i\right) + \exp\left(\frac{\beta^{(m)}}{(K-1)^2}\right) \sum_{i=1\&c_i\neq P(x_i)}^{n_l+n_u} w_i \\
&= \sum_{i=1}^{n_l+n_u} w_i \left\{\exp\left(-\frac{\beta^{(m)}}{K-1}\right)\left(1-error^{(m)}\right) + \exp\left(\frac{\beta^{(m)}}{(K-1)^2}\right)\left(error^{(m)}\right)\right\}
\end{aligned}
$$

令 $\frac{\partial L(Y^l, Y^u, H^{(m)})}{\partial \beta^{(m)}} = 0$，则有

$$
-\frac{1}{K-1}\exp\left(-\frac{\beta^{(m)}}{K-1}\right)\left(1-error^{(m)}\right) + \frac{1}{(K-1)^2}\exp\left(\frac{\beta^{(m)}}{(K-1)^2}\right)\left(error^{(m)}\right) = 0 \quad (5.10)
$$

由(5.10)式求解式 $\beta^{(m)}$，即可得到(5.8)式。

5.3 算法流程

基于前节的讨论，本节给出 SSMAB 方法的具体细节。该方法要求预先指定相关参数以及给定标记样本和未标记样本的权重系数，在后续迭代过程中为强调标记样本的重要性，需给定一个比未标记样本稍大的权重系数。算法的流程如图 5.2 所示。

SSMAB 算法继承了多类 AdaBoost 的优点，它对分类器的分类精度只需要$1/K$（K 为类别数目），这点对 SSMAB 非常重要，因为基分类器的精度在 SSMAB 中相对比较低，其原因是给定的伪标记含有错误的标记。但如果仅仅使用少量的标记样本训练，如 Muti-class AdaBoost 则容易陷入过拟合。为了缓解这个问题。在 SSMAB 算法中，所有的未标记样本被指定一个伪标记，同样被用来建立分类器。其目的是使用尽可能多的数据获得更合理的数据类分布，以期更充分反映数据的整体分布。

算法：[Semi-supervised Multi class AdaBoost (SSMAB)]

输入：

— D 训练数据集，包括标记样本和未标记样本

— n_l 标记样本数目

— n_u 未标记样本数目

— M 最大迭代次数

初始化：

(1) 标记样本与未标记样本权重 $w_i^{(0)} = \begin{cases} w_l & i = 1,2,...,n_l, \\ w_u & i = n_{l+1}, n_{l+2},...,n_l + n_u. \end{cases}$

(2) 利用KNN（K =1）方法给每个未标记样本指派一个伪标记

(3) 令 $H^{(0)}(x) = 0$

For $m = 1,2,...,M$ do

Step 1. 使用加权后的标记样本和伪标记样本训练一个多类基分类器 $P^{(m)}(x)$。

Step 2. 评价每个标记样本和未标记样本 $P^{(m)}(x_i)$， $i = 1...n_l + n_u$。

Step 3. 计算误差： $error^{(m)} = \sum_{j=1}^{n_l+n_u} w_j^{(m-1)} \Big/ \sum_{i=1}^{n_l+n_u} w_i^{(m-1)} \; j \in \{ j \mid c_j \neq P^{(m)}(x_j), j = 1...n_l + n_u \}$。

Step 4. 若 $error^{(m)} > (K-1)/K$ 则终止循环。

Step 5. 设定 $\beta^{(m)} = (K-1)^2 / K \cdot \left(\log\left((1 - error^{(m)}) / error^{(m)} \right) + \log(K-1) \right)$。

Step 6. 更新样本权重：

$$w_i^{(m)} = \begin{cases} \alpha . w_i^{(m-1)} . \exp\left(\dfrac{-\beta^{(m)}}{(K-1)^2} \right) & if \; i \in \{ i \mid c_i \neq P^{(m)}(x_i) \;\; i = 1,2,...,n_l \} \\ w_i^{(m-1)} . \exp\left(\dfrac{-\beta^{(m)}}{(K-1)^2} \right) & if \; i \in \{ i \mid c_i \neq P^{(m)}(x_i) \;\; i = n_{l+1}, n_{l+2},...,n_l + n_u \} \end{cases}$$

Step 7. 归一化 $w_i^{(m)} \leftarrow w_i^{(m)} . \sum_{i=1}^{n_l+n_u} w_i^{(m-1)} \Big/ \sum_{i=1}^{n_l+n_u} w_i^{(m)} \quad i = 1,2,...,n_l + n_u$。

Step 8. 更新 $H^{(m)}(x) = H^{(m-1)}(x) + \beta^{(m)} h^{(m)}(x)$。

输出： $H^{(m)}(x)$。

图 5.2 SSMAB（Semi-supervised Multi-class Adaboost）算法

5.4 时间复杂度分析

首先，利用 K 近邻标记未标记样本过程的时间复杂度为 $O(n_l \cdot n_u)$，假定以决策

树作为基分类器，通常这是一个很普遍的选择，假定树的深度为 d，则建立每棵决策树的计算代价为 $O\left(d \cdot I \cdot (n_l + n_u)\log(n_l + n_u)\right)$，$I$ 为输入数据 x 的维数，由于迭代次数为 M 次，所以计算代价为 $O\left(M \cdot d \cdot I \cdot (n_l + n_u)\log(n_l + n_u)\right)$。

因此 SSMAB 总的训练时间复杂度为 $O\left(n_l \cdot n_u + M \cdot d \cdot I \cdot (n_l + n_u)\log(n_l + n_u)\right)$。

5.5 实验及结果分析

本节将在 21 个 UCI[159]数据集上对 SSMAB 方法与其他相关方法进行比较，数据集描述见表 5.1。

表 5.1 21 个 UCI 数据集特征描述

数据集	样本数目	特征数目	类别数目
anneal	898	39	5
audiology	226	70	24
autos	205	26	6
breast-cancer	699	10	2
car	1728	7	4
horse-colic	368	23	2
dermatology	366	35	6
pima_diabetes	768	9	2
ecoli	336	8	8
Glass	214	10	6
ionosphere	351	35	2
letter	20000	17	26
optdigits	5620	65	10
pendigits	10992	17	10
segment	2310	20	7
sick	3772	30	2
sonar	208	61	2
soybean	683	36	19
vehicle	846	19	4
vote	435	17	2
wine	178	14	3

5.5.1 公共数据集实验设置

为了便于评价测试，对表 5.1 中的每个数据集进行划分。其中每个数据集的 25% 用作测试集，剩余的用作训练集。训练集又按照一定的比例划分成两部分，一部分为标记样本，另一部分为未标记样本（由于所有数据都有标记，实验中，本文隐去未标记样本的标记）。比如，数据集 *wine* 的样本数目为 178，如果训练集的划分比例为 10%，则标记样本和未标记样本的数目分别为 13 和 121，测试集的大小为 44。

设定迭代次数 M 最多为 30 次。设定参数 w_l 和 w_u 分别为 8.0，2.0。为强调原始标记样本的重要性，设定其权重系数 α 的值为 2.0。

本文在 WEKA[160]系统中实现了 ASSEMBE 和 SSMAB。并采用决策树（J48 Decision Trees）[161]和朴素贝叶斯（Naïve Bayes）[162]作为基分类器。选择这两个分类器做基分类器的主要原因是很多文献中提到它们是两个非常成功的分类器。Naïve Bayes 是贝叶斯网络（Bayes Networks）方法中最简单的一种。它是一个仅有一个父节点的有向图，该父节点表示未能观测到的节点（预测值），图中还有几个子节点，这些节点是可以观测的，表示特征，Naïve Bayes 最大的特点就是假定特征之间相对类别独立，简化了贝叶斯网络的复杂网络关系；J48 Decision Trees 也有与 Naïve Bayes 类似的树状结构，但是其解释意义却完全不同。J48 Decision Trees 的每个内部节点代表一个输入变量，叶子节点代表分类类别。虽然 J48 Decision Trees 与 Naïve Bayes 之间有许多差别，但由它们所建立的分类模型不仅快速而且易于解释。因此本文采用这两个分类器作为基分类器。在不同的标记比率下，两个基分类器被独立用于实验分析。为了使得决策树基分类器分类能力比较弱，设定其叶子节点的最小样本数目为 20，每个训练集中标记样本所占比例设为 10%，其余参数均为 WEKA 默认参数。

后续的实验安排主要分两部分，首先，将本章所提方法 SSMAB 与 ASSEMBLE，MultiAdaboost 及 J48 Decision Trees 三种方法进行比较。其中前三个方法采用 J48 Decision Trees 作为基分类器。其次，采用 Naïve Bayes 作为基分类器再次进行实验性能比较。该实验的主要目的是为了验证以下三点：

（1）SSMAB 能够使用少量的标记样本改善基分类器的分类性能；

（2）在同样的基分类器下，SSMAB 要比其他所比较的方法更有效；

（3）在 SSMAB 方法中，未标记样本对于分类是有帮助的。

四种方法 SSMAB、ASSEMBLE、MultiAdaboost 及 J48 Decision Trees 均在同样的训练集（后两种只在标记样本上进行训练）和测试集上进行训练和评价，对每个数据集，平均分类精度和方差采用 10 次交叉验证法（10-fold Cross Validation）验证。

5.5.2 实验结果及分析

表 5.2 给出了各种方法在标记比例 10%下的实验结果。其中 J48 Decision Trees 仅用标记样本进行训练。

表 5.2 四种方法的分类精度，其中数据的 75%用作训练集（训练集的 10%为标记样本），数据的 25%用作测试集，且 J48 Decision Trees 作为基分类器。

Data Set	J48 Decision Trees	MultiAdaBoost	ASSEMBLE	SSMAB
anneal	75.17±0.16	76.04±0.84	76.17±0.18	**89.98±4.22**
audiology	24.96±1.24	20.96±1.08	39.11±7.49	**48.36±7.09**
autos	32.69±0.69	32.78±0.68	37.95±9.40	**41.48±4.78**
breast-cancer	89.34±3.78	89.11±3.75	94.47±1.91	**96.47±1.36**
car	69.50±1.65	70.84±3.45	70.01±0.15	**76.23±1.34**
horse-colic	62.04±0.00	63.04±0.02	72.17±7.21	**76.52±5.78**
dermatology	30.80±0.18	30.60±0.20	79.78±5.21	**89.41±2.44**
pima_diabetes	65.16±5.60	65.99±6.64	68.59±5.49	68.70±3.01
ecoli	42.27±0.42	43.27±0.41	75.25±4.84	**79.68±5.30**
Glass	33.69±0.66	35.49±0.66	51.59±8.07	**58.21±6.25**
ionosphere	63.05±0.41	64.08±0.42	71.83±8.94	**80.15±7.86**
letter	55.20±1.07	56.3±1.45	73.71±0.63	**81.99±0.65**
optdigits	68.10±2.66	72.44±3.41	87.11±2.41	**95.73±0.55**
pendigits	79.93±1.75	82.24±1.15	88.88±1.53	**97.51±0.33**
segment	83.19±2.73	81.91±3.59	87.36±1.59	**90.99±1.23**
sick	95.06±1.62	**97.37±0.70**	93.87±0.04	95.34±1.02
sonar	53.55±0.46	55.65±0.41	56.33±5.78	**60.93±7.31**
soybean	26.16±0.64	26.16±3.61	65.42±2.36	**78.58±3.90**
vehicle	39.26±3.20	45.07±8.77	58.80±2.95	**61.17±1.66**
vote	60.40±0.28	61.40±0.15	**91.26±6.23**	88.79±3.51
wine	37.21±3.61	39.73±0.86	84.39±6.70	**91.20±3.62**

从实验结果可以看出，首先，SSMAB 对大部分数据集能显著的提高基分类器 J48 Decision Trees 的分类精度。有些改进特别明显，比如，数据集 *dermatology* 的精度从 30.80%提高到 89.41%。数据集 *optdigit* 的精度从 68.10%提升到 95.73%；其次，与 MultiAdaBoost（没有训练伪标记数据）相比，SSMAB 分类精度也有较明显的改进，比如，对数据集 *vote*，MultiAdaBoost 的精度只有 61.40%，而 SSMAB 的分类精度却有 88.79%。对有些数据集的分类精度，两者之间的差距超过了 30%，比如，*ecoli*，*soybean* 及 *wine*。这个结果表明，未标记数据在提高分类性能上确实起到了积极的作用；此外，我们还注意到，MultiAdaBoost 与 J48 Decision Trees 之间的分类精度几乎没有太大的区别。

表 5.3 四种方法的分类精度，其中数据的 75%用作训练集（训练集的 10%为标记样本），数据的 25%用作测试集，且 Naive Bayes 作为基分类器。

Data Set	NaiveBayes	MultiAdaBoost	ASSEMBLE	SSMAB
anneal	85.74±1.53	**88.87±2.35**	85.57±1.52	86.01±5.35
audiology	39.39±5.88	41.82±6.79	43.37±5.60	**46.89±6.78**
autos	36.25±7.43	36.44±7.54	34.26±9.33	**39.51±5.84**
breast-cancer	94.21±1.84	95.21±1.92	95.38±1.59	95.44±2.19
car	79.8±2.17	**83.55±2.87**	77.92±1.72	77.5±3.28
horse-colic	74.02±5.96	73.70±6.08	**75.76±5.73**	73.37±5.94
dermatology	85.71±4.37	85.81±4.97	**90.72±4.64**	85.59±4.29
pima_diabetes	70.36±4.99	**71.77±4.34**	70.26±4.11	69.74±4.21
ecoli	75.70±4.45	75.10±4.94	78.48±3.31	**79.79±5.66**
Glass	50.24±9.17	52.12±9.58	49.38±5.16	**55.70±4.98**
ionosphere	84.60±5.19	**87.93±4.43**	85.40±4.43	82.44±6.78
letter	61.93±1.54	62.41±1.48	**63.21±1.32**	57.88±1.37
optdigits	88.21±0.97	88.26±0.77	88.57±0.94	**91.21±0.79**
pendigits	84.58±1.65	84.78±1.23	84.66±1.62	**90.57±1.60**
segment	80.66±2.16	82.60±2.56	82.04±2.12	80.63±4.01
sick	92.45±2.92	91.68±2.93	**94.00±1.79**	93.70±1.82
sonar	60.38±7.93	60.19±7.98	59.21±7.50	60.56±7.04
soybean	77.83±3.56	77.65±3.46	**79.7±3.15**	74.36±3.93
vehicle	46.83±5.01	48.91±7.49	43.71±3.86	**54.87±5.29**
vote	89.71±2.41	89.98±2.28	89.52±2.57	87.95±3.04
wine	88.42±4.85	89.42±4.95	**92.34±3.70**	90.07±3.99

这个现象表明，当标记样本很少时，MultiAdaBoost 不能很好的改善决策树分类性能；再次，与 ASSEMBLE 方法相比，SSMAB 在 21 个数据集中也有显著的改进。比如，*dermatology*（89.41% vs 79.78%），*pendigits*（97.51% vs 88.88%），及 *letter*（81.99% vs 73.71%）。但是也存在比 ASSEMBLE 分类精度要差的情况，比如，对数据集 *vote*（88.79% vs 91.26%）。

表 5.3 给出了第二个实验结果，其中 Naïve Bayes 用作基分类器。与前一个实验一样，标记的样本比例仍为训练集的 10%。用于比较的 Naïve Bayes 分类器，只用标记的样本进行训练。从结果可以看出，SSMAB 能较好的提高基分类器 Naïve Bayes 的分类精度。如对数据集 *pendigits*，平均分类精度从 84.58%提高到了 90.57%。与 ASSEMBLE 方法相比，SSMAB 的分类精度也有较大的提高，比如数据集 *Glass*（55.7% vs 49.38%）和 *vehicle*（54.87% vs 43.71%），但是分类精度的改进效果没有表 5.2 中采用 J48 Decision Trees 作为基分类器明显。为什么用 Naïve Bayes 作为基分类器时 SSMAB 的改进效果不是很显著？原因是，Naïve Bayes 方法本身就是一个强分类方法，即便是在标记样本数目较少的情况下，也有很好的表现，该方法只需要少量的标记训练样本就可以评估出用于分类的参数[184]。因此，对于这样一个强方法，实验中其他三种 Boost-based 方法要想提高基分类器的精度就非常困难。表 5.3 的实验结果同时也展示了未标记样本对分类不总是有效。相关的理论已经在先前的有关文献[185]中予以充分讨论。

本文认为，SSMAB 良好的分类性能主要与下三个原因有关：第一，SSMAB 方法只需基分类器的分类精度好于$1/K$（K 为类别数），而 ASSEMLE 方法需要 1/2。因此基分类器越弱，优势越明显。第二，在样本权重迭代更新过程中，对标记样本使用更大的权重以强调标记样本被分错的代价更大。具体原因为，错误样本的权值越大，错误评价值$error^{(m)}$越大，表示当次循环中基分类器的权重$\beta^{(m)}$越小。第三，与有监督方法相比，包括 J48 Decision Trees、Naïve Bayes、MultiAdaBoost，SSMAB 能够利用未标记样本改善数据类分布而提高分类性能。

5.5.3 对标记样本比例敏感性分析

本节研究 SSMAB 对标记样本数量的敏感性。SSMAB 与其他三个方法在不同标记（3%~29%）比例的训练集下分别统一测试。图 5.3 记录了在 6 个数据集上采用 J48 Decision Trees 作为基分类器的实验结果。

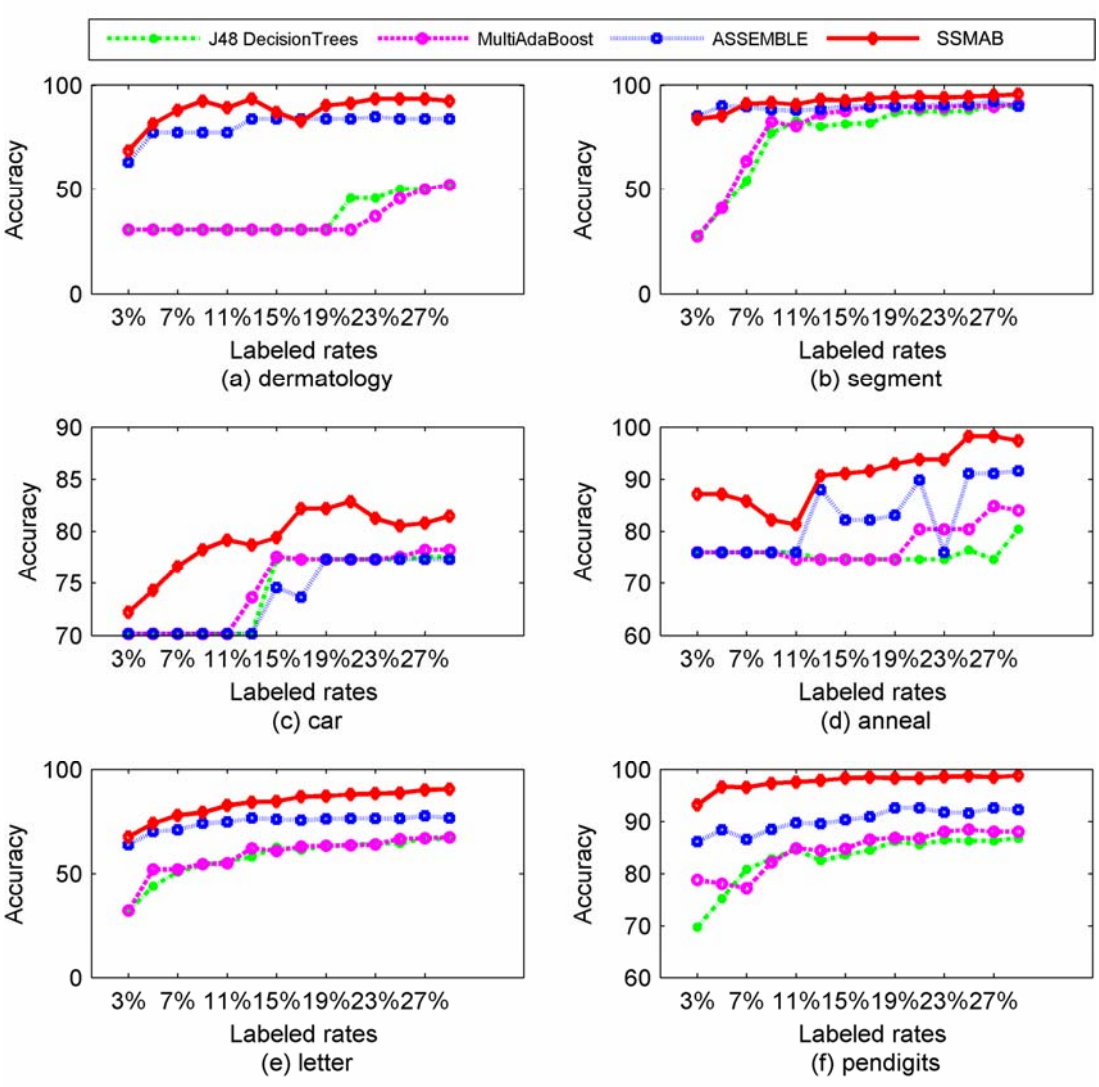

图 5.3 SSMAB 对标记样本数目敏感性分析。横坐标表示标记样本所占训练样本的比例，范围从 3%到 29%。纵坐标表示分类精度。图显示了包括 SSMAB 在内的四种方法的分类性能。

采用这 6 个数据集作为实验数据，是因为这 6 个数据集在 21 个数据集中数目较

大且类别都超过 2 类，其中 *letter* 和 *pendigits* 的类别数目分别为 26 和 10。从图 5.3 可以看出，整体来上，随着标记样本数目的增加，四种方法的分类精度都有所提高。但 SSMAB 对数据集 *pendigits*，*segment*，*car*，*letters* 在不同的标记样本下分类精度都普遍要比其他三种方法的分类精度要高。

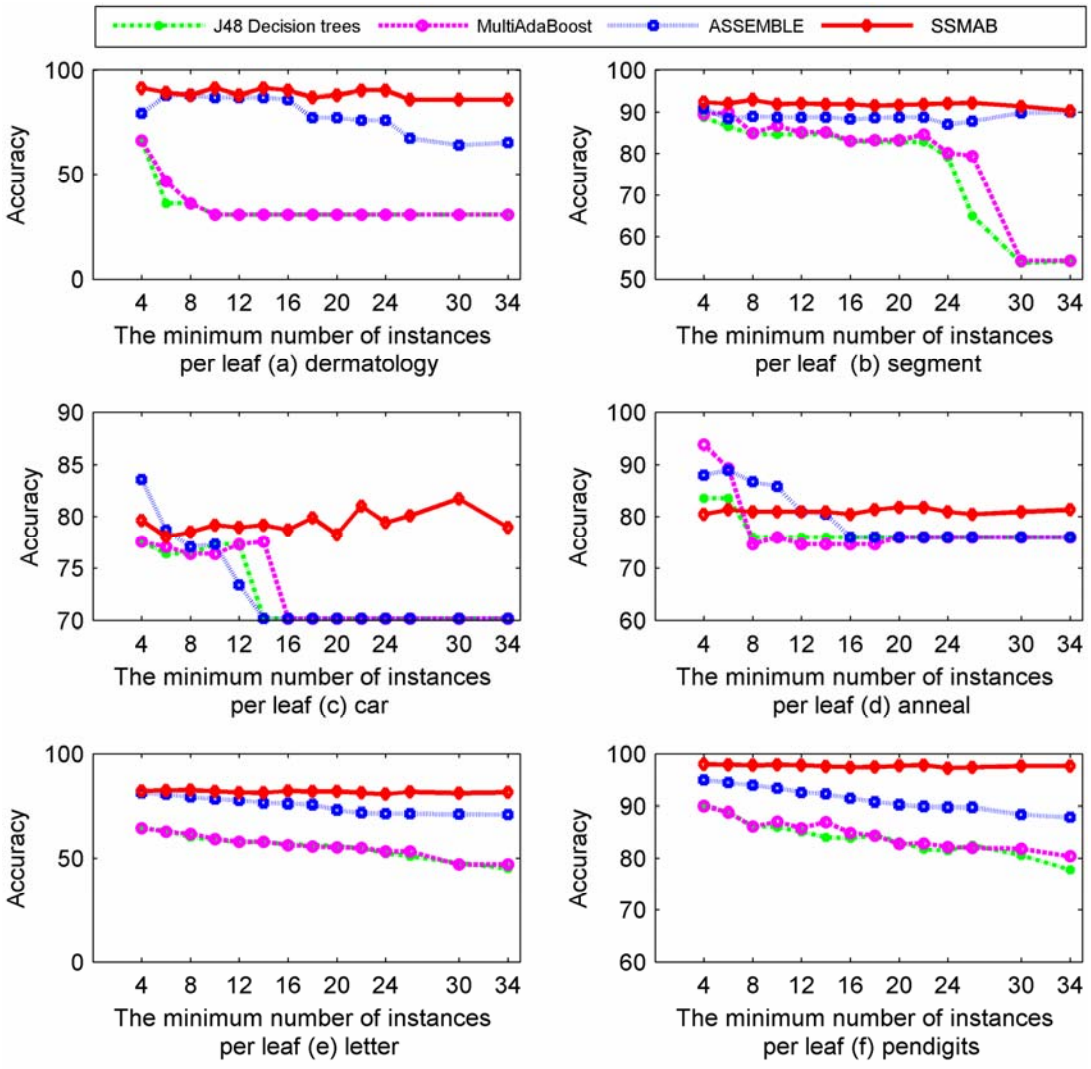

图 5.4 SSMAB 对基分类器敏感性分析。横坐标表示 J48 Decision Trees 中每个叶子节点最少样本数目，范围从 4 到 34。纵坐标表示分类精度。图显示了包括 SSMAB 在内四种方法的分类性能。

SSMAB 整体上要好于 MultiAdaBoost 且能有效地提高基分类器 J48 Decision Trees 的分类精度，这种现象在标记样本数目越小的情况下越明显。比如，对数据集 *pendigits*，SSMAB 在 3%标记比例下显著好于其他三种方法。而 ASSEMBLE 在这 6

个数据中对基分类器的改善效果要比 SSMAB 稍差一些。此外，从 6 条学习曲线上看，随着标记样本的增多，SSMAB 要比其他三个方法更稳定。

5.5.4 对基分类器敏感性分析

本节研究 SSMAB 方法对基分类器的敏感性。同样地，采用 J48 Decision Trees 作基分类器，通过调节 J48 Decision Trees 任意叶子节点最少样本数目（different minimum number of instances per leaf）来获得不同的基分类器。

众所周知，J48 Decision Trees 叶子节点上样本数目越多，它的树形结构越简单，所建立的分类模型泛化能力通常也越强。实验证实了每个叶子节点最少样本数目取不同值时，SSMAB 能同样有效地改进基分类器 J48 Decision Trees 的分类精度。

图 5.4 展示了 SSMAB 以及其他三种方法在 6 个数据集上基分类器 J48 Decision Trees 变化时，分类精度的变化情况，其中，基分类器的变化通过改变决策树中任意叶子节点上最少样本数目来实现。从实验结果可以看出，SSMAB 能显著的提高各种复杂程度的 J48 Decision Trees 的分类精度，与 ASSEMBLE 相比，SSMAB 也同样在 6 个数据集上显出了一定的优势。特别值得注意的是，与其他算法相比，SSMAB 的分类精度对 J48 Decision Trees 的变化表现得非常稳定，学习性能曲线变化平缓。比如对数据集 *pendigits* 与 *letter*，MultiAdaBoost 和 J48 Decision Trees 的分类精度迅速下降，而 SSMAB 的分类曲线却几乎没有变化。

5.6 本章小结

本章提出了一种分类器集成类型的半监分类方法 SSMAB 方法，它只需基分类器的分类精度好于$1/K$（K 为类别数）就可以获得良好的分类效果。通过大量的实验表明，该方法能够利用未标记样本改进基分类器的分类性能，且在大多数情况下具有良好的鲁棒性和分类精度；SSMAB 方法通过为每个未标记样本指派一个初始伪标记而获得合理的数据类分布，从而获得比其他三个方法更好的分类效果。虽然未标记样本在绝大多数情况下对半监督分类模型有很好的促进作用，但并不总是有效。

6 非公度的半监督近邻分类

半监督分类中，分类对象之间的差异性度量对分类效果影响非常大。目前已有的方法绝大多数都是基于可公度距离的方法，即数据样本之间相似性度量是基于可公度空间的，并且只考虑对象之间的直接差异，未考虑数据整体分布状况。但实际应用中，存在许多不满足可公度条件的问题。因此，不可公度度量的设计对解决这类问题具有很大的必要性。在本章，将就这个问题作重点研究，利用大量未标记样本的内在分布几何结构，构造一个反映数据整体分布信息、评价数据之间差异性的非公度度量。

6.1 基于非公度度量的机器学习

半监督学习方法研究至今，绝大多数方法都基于对数据的一些先验假设，这些假设通常有平滑假设（SA，Smoothness Assumption）、聚类假设（CA，Cluster Assumption）、流形假设（MA，Manifold Assumption）。平滑假设认为在高密度区域的数据点应该有相同或相似的类别标记，而在低密度区域发生改变；聚类假设认为，在相同的聚类应共享相同的类别标记；而流形假设则认为数据应嵌入在低维流形（Low-dimensional Manifold）中，并且数据的类别标记在同一流形中变化很平滑。

在这些假设下，影响分类性能的一个很重要的因素就是如何去评价数据间的相似与差异程度。怎样选取一个好的相似性度量往往是一个算法成功的关键因素。欧氏距离（Euclidean Distance）是目前用得最普遍的相似性度量方式，它是一种公度距离（Metric Distance）。但是许多应用往往都是非公度（Non-Metric）的，如文档聚类中对词频逆文档频率（TFIDF，Term Frequency Inverse Document Frequency）所采用的余弦相似性（Cosine-Similarity）[186]度量，生物领域中的 Pearson’s 协相关以及统计学中用到的相对熵（Kullback-Leibler Divergence）[187]。更有意思的是，人的判断方式也常常有类似的表现，比如，A 像 B，B 像 C，则不一定 A 像 C，换言之，对象之间相似性度量缺乏传递性。为了更形象的说明这一点，特举出由 Jacobs 等[188]给出的经典示意图来加以诠释。如图 6.1 所示。绝大多数观察者认为，人首马身怪兽

与人和马都很像，而人和马却有很大差异。且马和人之间的差异要远大于 2 倍怪兽与人或者与马之间的差异，这与通常相似性度量是不吻合的。这种非公度相似性度量的好处在文献[188]中进行了深入讨论。近年来，也有些研究人员对非公度距离做了相关工作[189, 190]。但这些工作的主要精力都是集中在有监督的学习问题。

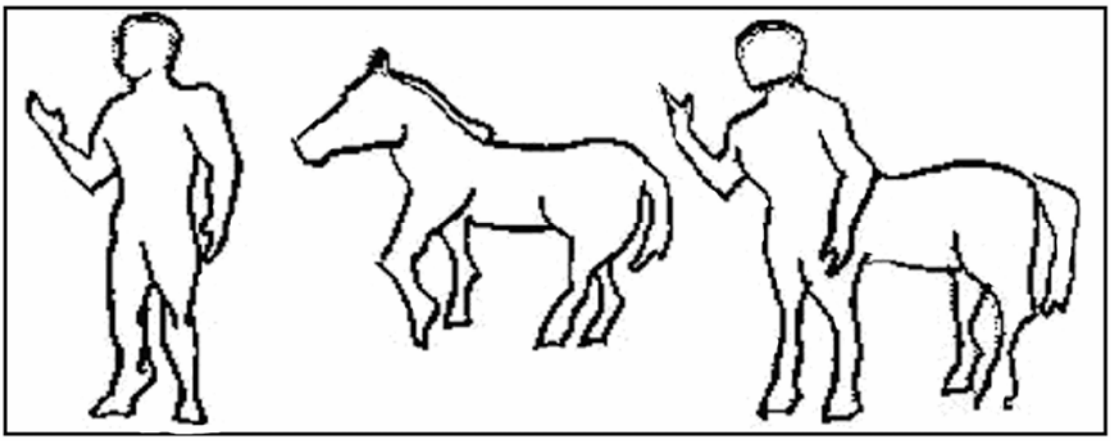

图 6.1 三幅图像之间的视觉差异不满足三角不等式

对于距离度量，测地距离（Geodesic Distance）是其中一种非常重要的距离度量，它经常用于发掘数据的流形结构。如 Tenenbaum 等人[191]于 2000 年在 Science 上发表的用于非线性降维的等距特征映射（ISOMAP，Isometric Feature Mapping）方法，ISOMAP 方法的主要思想是使得变换后的空间能尽可能保持数据之间的测地距离；Varini 等人[192]利用测地距离改进局部线性嵌入法（LLE，Local Linear Embedding）[193]，提出了 ISO-LLE 方法；Cruz-Barbosa 和 Vellido[194]提出了一个半监督学习模型 SS-Geo-GTM，它是一种基于测地距离拓扑图映射的扩展方法。Lebanon[195]提出了一个基于黎曼公度（Riemannian Metrics）目标函数的分类方法，该方法能使高密度区域中两点的最短路径距离更短。其采用的连通核（Connectivity Kernel）[196]被用于半监督分类，其中数据点对之间的距离采用连通两点的路径中最大的一条边长来度量。但这个方法需要调节较多的参数才能获得较好的分类效果，Chang 和 Yeung[197]也用了类似的度量用于改进聚类的鲁棒性。

本章将提出一种新的半监督近邻分类方法 NMSNN，其中所有的标记样本与未标

记样本之间的距离度量将建立在一个非公度空间上，并给出一个新的非公度距离度量——最短路径的代价长度，以衡量在 K 近邻图中两数据点间的亲疏（Affinity）关系。NMSNN 方法在 SA 与 MA 假设前提下，能有效地发掘数据蕴含的内在几何结构并改善分类效果。该方法认为数据在同一流形或高密度区域的类别应该相同。也就是说，数据间亲疏关系不以欧氏空间的距离来度量。这样做的最大好处是建立分类模型时不需要假定数据的每一类满足高斯分布（Gaussian Distribution），在实际应用中也确实存在很多数据分布不满足高斯分布。此外，NMSNN 只需要处理两个参数，尺度参数 δ 和近邻数目参数 K。但通过进一步理论分析，尺度参数不需要调节，因此 NMSNN 只需要调节一个参数，从而更加实用，这减少了因参数设置不当带来的问题，使得该方法更具有实用价值。

6.2 非公度的半监督近邻分类方法NMSNN

6.2.1 相关定义

给定数据集 $X=\{x_1,x_2,...,x_{n_l},x_{n_l+1},...,x_n\}\subset R^d$，其中 x_i 属于（类别数目 $C\geq 2$）其中的某一类，d 表示数据的维数。$X_L=\{x_1,x_2,...,x_{n_l}\}\subset X$ 和 $X_U=\{x_{n_l+1},x_{n_l+2},...,x_n\}\subset X$ 分别表示标记样本数据集和未标记样本数据集，且 $X=X_U\cup X_L$。其中 X_L 中样本的相应类别标记为 $Y_L=\{y_1,y_2,...,y_{n_l}\}$，$n_l$ 和 n_u 分别表示标记样本数目和未标记样本数目，$n=n_l+n_u$，且 n_l 远远小于 n_u。

定义 6.1 令 $K\in R$，$x_i\in X$ 在 X 中的 K 个欧氏距离最相近的近邻成为 K 近邻 $N_K^X(x_i)$。

定义 6.2 K 近邻图是一个连接每个样本到其 K 个近邻的有向图。记为 G，在 G 中任意两对 x_i 和 x_j 的距离定义如式(6.1)，其中 $x_i,x_j\in X$；$i,j=1,2,...,n$，$\|.\|^2$ 表示向量在 R^d 中的 2 阶范数。

$$d(i,j)=\begin{cases}\infty & x_j \notin N_K^X(x_i), \\ \left\|x_i - x_j\right\|^2 & \text{否则.}\end{cases} \tag{6.1}$$

定义 6.3 在 *K*-近邻图 G 中连接 x_m 和 x_n 路径 P 的长度为路径上所有边的长度和。其中 P 由 $e_1, e_2, ..., e_k$ 组成。其长度为 $L(P)=\sum_{i=1}^{k} d_i$ ， d_i 为边 e_i 的长度。

定义 6.4 公度距离，给定一个 X ，若实值度量函数 $d(x_i, x_j)$ 在笛卡尔积 $X \times X$ 下对任意的 $x_i, x_j, x_k \in X$ 都满足下面四个条件，则 $d(x_i, x_j)$ 是一个公度距离。

(1) $d(x_i, x_j) \geq 0$ （非负性）；

(2) $d(x_i, x_j) = d(x_j, x_i)$ （对称性）；

(3) $d(x_i, x_j) + d(x_j, x_k) \geq d(x_i, x_k)$ （三角不等式）；

(4) $d(x_i, x_j) = 0$ ， 当且仅当 $x_i = x_j$ （不可区分者的同一性）。

定义 6.5 在图 G 中连接 x_m 与 x_n 路径的代价长度为 $CL(m,n)=\sum_{i=1}^{k}\exp(d_i)$ ，其中 $\exp(\cdot)$ 为指数函数，且 $CL(m,n)$ 为非公度距离，证明细节可参考**定理 6.2**。

6.2.2 理论分析

定理 6.1 假定 x_m 与 x_n 之间存在许多路径，其代价长度为 $CL(m,n)$ 记为 $CL(k)$ ，则有

$$CL(k) \in \left[\left(\lfloor L \rfloor - 1\right)\exp(1) + \exp(L - \lfloor L \rfloor + 1),\ \exp(L)\right],\quad k = 1, 2, ..., \lfloor L \rfloor 。$$

式中， $L=\sum_{i=1}^{k} d_i$ ， $d_i \geq 1$ ， $\lfloor x \rfloor = \max\{n \in \mathbb{Z} \mid n \leq x\}$ ， x 为实数， n 为整数， $\mathbb{Z}$ 为正整数集合。

证明：分两步证明如下：

第 1 步：假定 $k \geq 1$ 为一个整数，求 $CL(k)$ 的最小值。

$$\begin{aligned} &\min CL(k) = \sum_{i=1}^{k} \exp(d_i) \\ &s.t.\ L = \sum_{i=1}^{k} d_i; \qquad d_i \geq 1. \end{aligned} \tag{6.2}$$

(6.2)式转换成拉格朗日表达形式 $CL(d_i, \lambda)$，具体定义如(6.3)式所示：

$$CL(d_i, \lambda) = CL(k) + \lambda\left(L - \sum_{i=1}^{k} d_i\right) = \sum_{i=1}^{k} \exp(d_i) + \lambda\left(L - \sum_{i=1}^{k} d_i\right) \tag{6.3}$$

式(6.3)中，λ 为拉格朗日系数。

为了求 $\arg\min_{d_i, \lambda} \sum_{i=1}^{k} \exp(d_i)$，对(6.3)式求分别对 d_i 与 λ 导数，令

$$\frac{\partial CL(d_i, \lambda)}{\partial \lambda} = L - \sum_{i=1}^{k} d_i = 0. \tag{6.4}$$

$$\frac{\partial CL(d_i, \lambda)}{\partial d_i} = \exp(d_i) + \lambda = 0. \tag{6.5}$$

从(6.4)式和(6.5)式可得 $d_1 = d_2 = ...d_i... = d_k$，$\lambda = \exp(k/L)$，且 $CL(k)$ 的最小值表达式(6.6)如下：

$$\min CL(k) = \min \sum_{i=1}^{k} \exp(d_i) = k \exp(L/k) \tag{6.6}$$

第 2 步：假定(6.6)式中 $k \geq 1$ 是一个实数变量（由 x 表示），问题转换成求解(6.7)式的最小值与最大值。

$$CL(x) = x \exp(L/x),\ x = [1, \infty) \tag{6.7}$$

为寻找(6.7)式的最小与最大值，令

$$\frac{\partial CL(x)}{\partial x} = (1 - L/x) \exp(L/x). \tag{6.8}$$

从(6.8)式可得，

$$\frac{\partial CL(x)}{\partial x}=\begin{cases}\le 0 & \text{若} L/x\ge 1,\\ >0 & \text{否则.}\end{cases} \tag{6.9}$$

由(6.9)可知，若$L/x\ge 1$，则有(6.10)成立。

$$CL(x=L)\le CL(x)\le CL(x=1) \tag{6.10}$$

综合**第 1 步**与**第 2 步**的分析，有(6.11)式成立。

$$\left(\lfloor L\rfloor-1\right)\exp(1)+\exp(L-\lfloor L\rfloor+1)\le CL(k)\le\exp(L) \tag{6.11}$$

式(6.11)中，$k=1,2,...,\lfloor L\rfloor$，从而**定理 6.1** 得证。

注解 6.1 假定d_i的方差为$\mathrm{var}(d_i)$，$d_i\ge 1$。根据**定理 6.1** 有：

(a) 若$CL(k)=CL(\lfloor \mathrm{L}\rfloor)$，则$\mathrm{var}(d_i)\to 0$，其相应的路径最光滑且通过的点最多，为$\lfloor \mathrm{L}\rfloor$+1（包括起点和终点）。

(b) 若$CL(k)=CL(2)$，则$\mathrm{var}(d_i)=2(\lfloor L\rfloor-1-L/2)^2$，若$d_1=\lfloor L\rfloor-1,\ d_2=L-\lfloor L\rfloor+1$

在这种情况下，该路径仅仅经过 3 个数据点（包括起点和终点）。因此，对于给定长度为L的所有路径，代价长度越小，则路径越光滑，且通过的点越多（越密集）；反之则路径越不光滑，通过的数据点越少（越稀疏）

定理 6.2 图G中路径$P(m,n)$的代价长度为$CL(m,n)$是非公度度量，因为$CL(m,n)$不满足三角不等式和对称性。

证明：令$CL(m,k)$，$CL(k,n)$，$CL(m,n)$分别代表x_m，x_k，x_n之间在图G中的最短路径的代价长度。

(1) 对称性不满足，即$CL(m,n)=CL(n,m)$不成立

由于K近邻图G为方向图，因此最短路径同样是有方向，即$P(m,n)=P(n,m)$不一定成立，因此$CL(m,n)=CL(n,m)$不一定成立。

(2) 三角不等式不满足，即$CL(m,k)+CL(k,n)>CL(m,n)$不成立

根据**定理 6.1**，有(6.12)——(6.14)三式成立。

$$\left(\lfloor L(m,k)\rfloor-1\right)\exp(1)+\exp(L(m,k)-\lfloor L(m,k)\rfloor+1)\le CL(m,k)\le\exp(L(m,k)) \quad (6.12)$$

$$\left(\lfloor L(k,n)\rfloor-1\right)\exp(1)+\exp(L(k,n)-\lfloor L(k,n)\rfloor+1)\le CL(k,n)\le\exp(L(k,n)) \quad (6.13)$$

$$\left(\lfloor L(m,n)\rfloor-1\right)\exp(1)+\exp(L(m,n)-\lfloor L(m,n)\rfloor+1)\le CL(m,n)\le\exp(L(m,n)) \quad (6.14)$$

从(6.12)——(6.14)三式可以推出不等式(6.15)和(6.16)成立

$$\left(\lfloor L(m,k)\rfloor+\lfloor L(k,n)\rfloor-2\right)\exp(1)<CL(m,k)+CL(k,n)\le\exp L(m,k)+\exp L(k,n) \quad (6.15)$$

$$\left(\lfloor L(m,n)\rfloor-1\right)\exp(1)<CL(m,n)\le\exp(L(m,n)) \quad (6.16)$$

注意到$L(m,k)+L(k,n)>L(m,n)$一定成立，因为$L(m,n)$是x_m与x_n在图中G的最短路径。然而从不等式(6.15)和(6.16)可知，$CL(m,k)+CL(k,n)>CL(m,n)$不总是成立。

推论 6.1 不失一般性，假定$L_1\ge L_2$，则：$\exists P_1, P_2$：$CL_1\le CL_2$，若$L_1\in\left[L_2,\ \exp(L_2-1)\right]$（非常弱的限制条件）

其中，假设在x_m与x_n之间的存在许多长度为L_1的路径，记为$Pset_1$；类似的，假定x_m与x_l之间存在许多长度为L_2的路径，记为$Pset_2$，并假定$P_1\in Pset_1$与$P_2\in Pset_2$分别表示两个路径集合中的两条路径。相关定义如下：

$$L_1=\sum_{i=1}^{k}d_i,\ CL_1=\sum_{i=1}^{k}\exp(d_i),\ \ L_2=\sum_{j=1}^{s}d_j, CL_2=\sum_{j=1}^{s}\exp(d_j),\ \ d_i,d_j\ge1,$$

k与s分别为P_1与P_2边数，CL_1与CL_2分别表示路径P_1与P_2的代价长度。

从**推论 6.1**可以得出，在欧氏空间（公度空间）中相对较长的路径有可能在非公度空间中变成相对短的路径。

证明：从**定理 6.1**可得不等式(6.17)和(6.18)成立。

$$\left(\lfloor L_1\rfloor-1\right)\exp(1)+\exp\left(L_1-\lfloor L_1\rfloor+1\right)\le CL_1\le\exp(L_1) \quad (6.17)$$

$$\left(\lfloor L_2\rfloor-1\right)\exp(1)+\exp\left(L_2-\lfloor L_2\rfloor+1\right)\le CL_2\le\exp(L_2) \quad (6.18)$$

将(6.17)和(6.18)化简可得

$$\lfloor L_1 \rfloor \exp(1) \le CL_1 \le \exp(L_1) \tag{6.19}$$

$$\lfloor L_2 \rfloor \exp(1) \le CL_2 \le \exp(L_2) \tag{6.20}$$

由不等式(6.19)和(6.20)根据以下两种情况来证明**推论 6.1**。

(1) 若 $L_1 \in [L_2,\ \exp(L_2 - 1)]$（很弱的限制）

(a) 若 $CL_2 \in (L_1 \exp(1),\ \exp(L_2)]$，则存在路径 P_1 和 P_2 使得 $CL_1 \le CL_2$ 成立。

(b) 若 $CL_2 \in [L_2 \exp(1),\ L_1 \exp(1)]$，则 $CL_1 \le CL_2$ 一定对任何长度为 L_2 的路径都不成立，但是 $[L_2 \exp(1),\ L_1 \exp(1)]$ 相比 $(L_1 \exp(1),\ \exp(L_2)]$ 是一个很窄的范围，因此是一个非常强的限制。

(2) 若 $L_1 \in [\exp(L_2 - 1),\ \infty)$（很强的限制）

对长度为 L_2 的任意路径，$CL_1 \le CL_2$ 都不成立，但是 $L_1 \in [\exp(L_2 - 1),\ \infty)$ 在本文所考虑的问题中，它是一个很难满足的条件。因为 L_1 与 L_2 作为图 G 中的两个最短路径的长度，在绝大多数情况下不会有很显著的差异。

综合（1）和（2）的讨论可得，若 $L_1 \in [L_2,\ \exp(L_2 - 1)]$，则存在路径 P_1 和 P_2 使得 $CL_1 \le CL_2$ 成立，且 P_1 比 P_2 更光滑。

6.3 算法流程

NMSNN 主要由三个部分组成。首先，利用标记样本与未标记样本构建一个 *K*-近邻图，样本与样本之间的距离为欧氏距离。每一个样本与其他样本均有一个有向权重连接。其次，使用 Dijkstra 算法为每一个未标记样本找到其到每一个标记样本的最短路径，因此，对于一个特定未标记样本，有 $l \le n_l$ 条最短路径连接所有的标记样本。

算法：[Non-metric based Semi-supervised Nearest Neighbor (NMSNN)]

输入：

— X_L 标记样本

— X_U 未标记样本

— K 近邻数目

初始化：将标记数据和未标记数据的特征归一化到$[0,1]$。

$\widetilde{Y_U} := \varnothing$ //未标记样本集的伪标记。

处理过程：

Step 1. $G := kNNGraph(X_U, X_L, K)$ //构建一个 K 近邻图

Step 2. 令 $\delta := \{\min d(i,j) \mid d(i,j) > 0; i, j = 1,2,...,n\}$ //设置 δ 尺度因子

Step 3. 给每个未标记样本 $x_i \in X_U$ 赋予一个类别标记 y_i

(1) **For** $x_i \in X_U, i++$ **do** $minCostLength := \infty$；

(2) **For** $x_j \in X_L, j++$ **do**

(3) $P(x_i, x_j) := getShortestPath(x_i, x_j, \delta, G)$；

(4) $CL(x_i, x_j) := getCostLength(G, P(x_i, x_j))$； //获得路径的代价长度

(5) 若 $CL(x_i, x_j) < \min CostLength$ 则

(6) $minCostLength := CL(x_i, x_j)$； $y_i := y_j$； //将 x_j 的标记赋给 x_i

(7) **End** For

(8) **End** For

输出：$\widetilde{Y_U} = \{y_{n_l+1}, y_{n_l+2}, ..., y_{n_l+n_u}\}$

图 6.2 非公度的半监督近邻分类算法（NMSNN）

对于所有这些路径，NMSNN 利用一个路径代价函数（Cost Length Function of the Path）重新评估其距离。为保证 $d_i \geq 1$，设置所有这些路径上边的距离为 $d_i = d_i / \delta$，

原因可以参考**定理 6.1** 其中 $\delta := \{\min d(i,j) \mid d(i,j) > 0;\ i,j = 1,2,...,n_l + n_u\}$。最后，根据**定理 6.2** 可知，代价长度越小，路径越光滑。因此，对一个未标记样本来说，一旦与某一标记样本路径的代价长度确定为最小代价长度，则未标记样本的类别标记为相应标记样本的的类别。详细描述见图 6.2。

6.4 时间复杂度分析

给定 d 维空间数据集 $X = X_U \cup X_L$，则构建 K 近邻图 G 的时间复杂度为 $O\left((n_l + n_u)^2\right) = (n_l + n_u)(n_l + n_u - 1)/2$；计算 δ 的时间复杂度为 $O\left((n_l + n_u)\log(n_l + n_u)\right)$。由于图 G 边数为 $K(n_l + n_u)$，而 $K(n_l + n_u) \ll O\left((n_l + n_u)^2\right)$，因此该图的存储矩阵通常是一个稀疏矩阵，可以用邻接表来更有效地实现迪科斯彻算法。同时需要将一个二叉堆用作优先队列来寻找最小的顶点。优化后的算法所需的时间复杂度为从 $O\left((n_l + n_u)^3\right)$ 下降为 $O\left((n_l + n_u) + K(n_l + n_u)\right)\log\left(K(n_l + n_u)\right)$，且复杂度以该式的后半部分 $O\left(K(n_l + n_u)\right)\log\left(K(n_l + n_u)\right)$ 为主。因此，为所有未标记样本找最小代价长度路径并标记相应归属类别的时间复杂度为 $O\left(K(n_l + n_u)\log\left(K(n_l + n_u)\right)\right)$。

累计各个步骤的时间复杂度，最终可以得到 NMSNN 的总的时间复杂度为 $O\left((n_l + n_u)^2\right) = O\left((n_l + n_u)^2 + (n_l + n_u)\log(n_l + n_u) + K(n_l + n_u)\log\left(K(n_l + n_u)\right)\right)$。

6.5 实验及结果分析

本节对 NMSNN 方法与其他三种方法在合成数据集和实际分类数据集上进行了实验比较，并给出相关实验分析结果。

6.5.1 合成数据集实验结果

合成数据集为 *Two-Moons* 与 *Three-Circle* 数据集。*Two-Moons* 含有 500 个数据

样本，共两类，分别含有 254 和 246 个数据样本。其中每类仅 3 个样本有类别标记。数据可视化的结果见图 6.3。

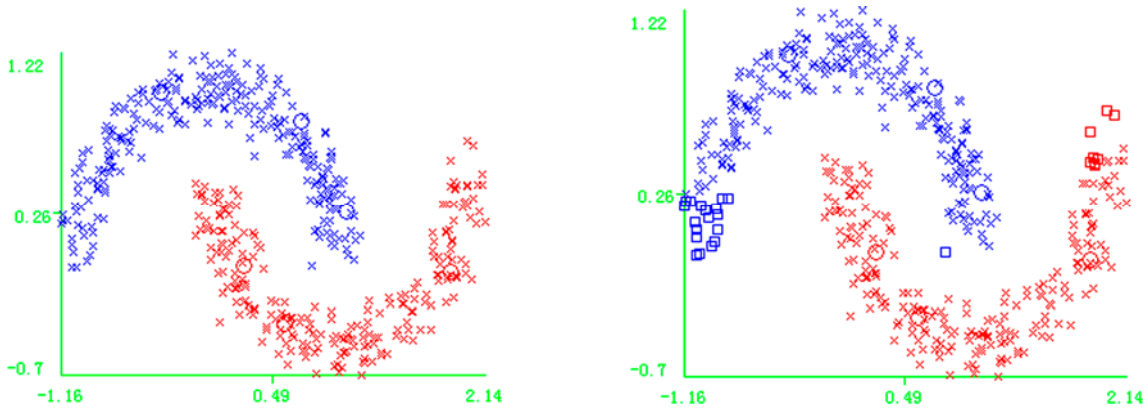

图 6.3 Two-Moons 数据集：左图显示了 NMSNN 方法的分类结果（K =10）。右图是最近邻的分类结果。‘〇’表示标记的样本（6），‘×’表示未标记样本（494），而‘□’则表示分类错误的数据。不同的颜色代表不同的类别，共 2 种。

图 6.4 给出了 *Three-Circle* 的的分类结果。数据集共由 1203 个数据样本组成。每类有 401 个数据样本，所有样本都分布在同一中心、3 个不同半径的圆上（带有高斯扰动），每一类样本仅有 4 个样本被标记。

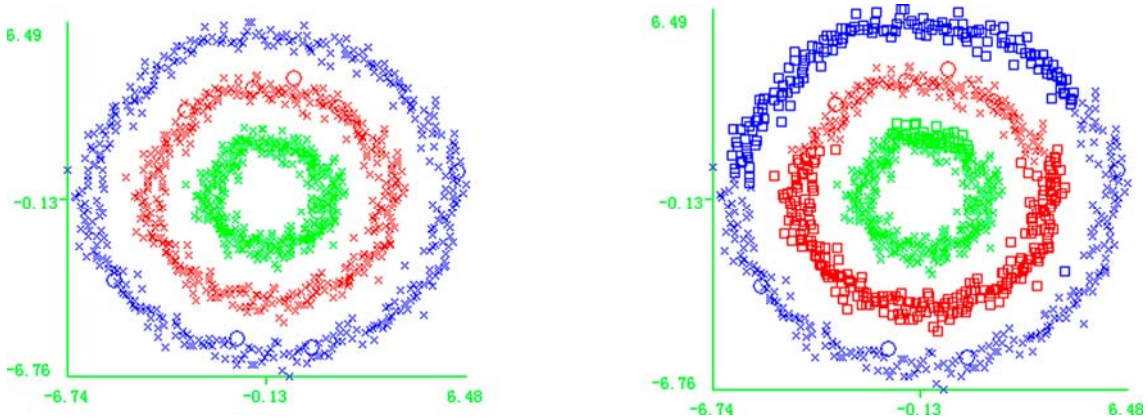

图 6.4 *Three-Circle* 数据集：左图显示了 NMSNN 方法的分类结果（K =6）。右图是最近邻的分类结果。‘〇’表示标记的样本（12），‘×’表示未标记样本（1191），而‘□’则表示分类错误的数据。不同的颜色代表不同的类别，共 3 种。

从图 6.3 和图 6.4 两图左图中可以看出，在标记样本很少的两个数据集中，

NMSNN 方法的分类效果要显著好于 NN 方法。NMSNN 方法在选择合适的参数 *K*（近邻个数）情况下对合成数据的分类没有一个数据被错分。

6.5.2 标准数据集实验结果

表 6.1 描述了实验数据集的相关信息，这些公共的数据集都是用于研究半监督学习的专用测试数据集，文献[198]给出了其下载地址。这些数据被分为聚类型（*cluster-like*）和流形型（*manifold-like*）两大类。*manifold-like* 数据集包括 *Digit1*，*USPS*，*COIL* 及 *BCI*，而 *cluster-like* 则包括 *g241c*，*g241d*，*Text*。

表 6.1 标准数据集数据描述

数据集	样本数目	特征个数	类别数目	类别分布比	数据类型
g241c	1500	241	2	1: 1	*cluster-like*
g241d	1500	241	2	1: 1	*cluster-like*
Digit1	1500	241	2	1: 1	*manifold-like*
USPS	1500	241	2	1: 4	*manifold-like*
COIL	1500	241	6	balanced	*manifold-like*
BCI	400	117	2	1: 1	*manifold-like*
Text	1500	11960	2	1: 1	*cluster-like*

为了评价 NMSNN 方法的有效性，本文比较了非常经典的半监督分类方法——全局与局部一致性学习方法（LLGC，Learning with Local and Global Consistency）[131]和有监督的方法（1-NN，1-Nearest Neighbor）。除此之外，还比较了 MSNN（Metric based Semi-supervised Nearest Neighbor）方法。MSNN 方法同样采用 Dijkstra 算法在图中为每一个未标记样本寻找到标记样本的最短路径。与 NMSNN 方法不同的是，MSNN 方法没有采用代价函数。所有这些方法均在 WEKA[160]环境中实现并统一比较。LLGC 方法的参数 *alpha* 和 *sig* 取值分别设置为 0.99 和 0.15[131]，MSNN 和 NMSSN 中的参数 *K* 设置为 40。对每个数据集，采用保留（Hold Out）验证法进行验证，即数据被分成没有重叠的两部分，一部分作为有标记的训练集，另一部分为未标记数据集，并随机重复 10 次。所有比较的方法均在相同的训练集和测试集上进行，测试集就是对应的未标记数据集。表 6.2——表 6.4 记录了 *manifold-like* 类型数据的平均

测试精度和方差。

表 6.2 在 Digit1 数据集上用 10-100 个标记训练样本的分类精度（%）

Labeled#	10	15	30	45	75	100
1-NN	75.84±3.00	82.03±2.33	86.56±3.25	89.53±2.17	91.99±1.45	93.16±1.21
MSNN	61.03±3.25	64.03±3.56	70.7±2.79	74.85±1.88	81.24±1.31	84.64±0.73
LLGC	80.76±3.91	83.33±2.67	85.07±0.02	87.07±0.00	91.07±1.02	92.76±1.72
NMSNN	**81.34±3.96**	**85.99±1.96**	**88.67±2.73**	**90.49±1.86**	**92.22±1.41**	93.59±1.28

表 6.3 在 USPS 数据集上用 10-100 个标记训练样本的分类精度（%）

Labeled#	10	15	30	45	75	100
1-NN	80.64±3.94	82.53±4.53	86.66±2.68	88.39±2.37	90.52±1.56	91.59±1.08
MSNN	79.95±1.80	80.77±1.88	83.45±1.15	85.15±1.21	86.86±0.88	88.30±1.08
LLGC	82.28±1.96	83.21±3.09	86.00±1.01	88.00±1.34	91.00±0.75	**94.34±1.77**
NMSNN	**83.92±2.65**	**84.84±4.42**	**88.67±1.71**	**89.78±2.09**	91.19±1.48	92.14±1.00

从表 6.2——表 6.4 三个表中的结果可以看出，NMSNN 在 3 个 *manifold-like* 数据集上的分类精度明显好于 MSNN 上的分类精度。这表明非公度代价函数在这种情况下要比公度函数要更有效。但是与 LLGC 相比，NMSNN 的分类优势却不是太明显。对于这个结果，其实并不意外，这与两个方法的类标记过程有关。LLGC 通过一个迭代等式将标记样本的标记传播给未标记样本来完成类别标记过程，而 NMSNN 则通过使用非公度的代价函数让未标记样本在图中主动的寻找合适的标记样本完成类别标记过程。因此，NMSNN 的标记过程从某种意义上来说是 LLGC 的逆过程。

表 6.4 在 COIL 数据集上用 10-100 个标记训练样本的分类精度（%）

Labeled#	10	15	30	45	75	100
1-NN	33.33±4.50	39.07±3.87	57.07±5.39	64.9±4.36	75.09±2.04	79.36±2.04
MSNN	35.56±4.81	40.36±3.30	57.05±4.59	64.94±3.23	74.57±2.29	78.33±2.35
LLGC	33.67±4.36	41.18±3.55	58.63±3.05	62.48±0.52	74.65±0.03	78.89±2.61
NMSNN	**37.32±5.28**	41.76±3.41	58.59±4.74	**66.30±3.08**	75.57±2.56	79.20±2.43

与 1-NN 相比，NMSNN 在 3 个 *manifold-like* 数据集的分类效果要好于 1-NN 上

的分类精度，并且标记样本数目越少越明显。然而，随着标记样本的增加，两者的性能越来越接近。那到底多少标记样本才是合适的呢？在回答这个问题之前，有必要回顾半监督分类方法的初衷——用少量的标记样本和大量的未标记样本得到一个很好的分类结果。因此如果标记样本太多则与其初衷相违背，从这点来说，NMSNN方法有一定的优势。在同样性能的情况下，标记样本越少越好，但对于一个具体的半监督方法，有没有一个具体的定量标准给出多少个标记样本合适？尽管也有相关的理论研究[199]，但仍然是一个有待解决的开放性问题。目前对半监督学习的研究绝大多数都是在理论分析的基础上进行实验验证。

表 6.5 在 g241c 数据集上用 10-100 个标记训练样本的分类精度（%）

Labeled#	10	15	30	45	75	100
1-NN	**56.60±2.37**	**58.50±2.65**	**59.56±2.80**	**59.13±2.84**	**60.24±1.47**	60.85±1.06
MSNN	49.58±1.29	49.80±0.94	50.26±1.33	50.77±1.54	51.32±1.79	51.76±1.77
LLGC	51.44±1.31	52.63±1.76	50.00±0.00	49.97±0.01	49.97±0.01	54.26±1.44
NMSNN	53.64±2.65	55.12±2.62	57.56±2.69	58.05±3.16	59.33±1.43	60.08±1.04

表 6.5 和表 6.6 记录了在 *cluster-like* 类型数据集上 10 次测试的平均分类精度及方差结果。

表 6.6 在 g241d 数据集上用 10-100 个标记训练样本的分类精度（%）

Labeled#	10	15	30	45	75	100
1-NN	**56.95±2.78**	**58.11±1.94**	**59.83±1.88**	**60.98±1.58**	61.82±0.91	61.96±1.02
MSNN	50.30±1.35	49.91±1.02	50.87±1.15	51.01±1.36	51.80±1.61	51.96±1.73
LLGC	51.01±2.17	52.43±1.89	50.16±0.67	50.03±0.14	51.40±4.47	55.83±1.86
NMSNN	53.84±2.34	54.88±2.58	57.88±2.12	59.16±1.24	61.30±0.97	**62.14±1.31**

从表中可以看出，在 *cluster-like* 数据中，半监督分类方法的精度与有监督分类方法相比精度不一定有提高。更有甚者，在有些情况下，分类效果比有监督的分类器还要差。比如，NMSNN 与 1-NN 在 *cluster-like* 数据集上的分类精度差别很小，而 LLGC 的分类精度还不如 1-NN 的分类精度。类似的现象在先前的有关文献中也被提到过[198]，Ben-david[199]对此进行了相关的理论研究，并指出了未标记数据在有些情

况下有用，有些情况下没用。这种现象从某种程度上验证了著名的无免费午餐定理（NFL，No Free Lunch）[200]，即不存在一个学习方法在所有数据集上均能达到最好的学习效果。

6.5.3 “*K*”值敏感性分析

本节主要分析 NMSNN 方法对“*K*”近邻参数的敏感性，以便更加深入地了解 NMSNN 方法的稳定性。实验中 6 个数据集均设置 100 个标记样本。为了便于实验设置，其中 BCI 数据集由于数据样本数目与这 6 个数据集不一样而未用于实验比较。图 6.5 和图 6.7 给出了 *manifold-like* 与 *cluster-like* 数据集上“*K*”值敏感性分析的结果。

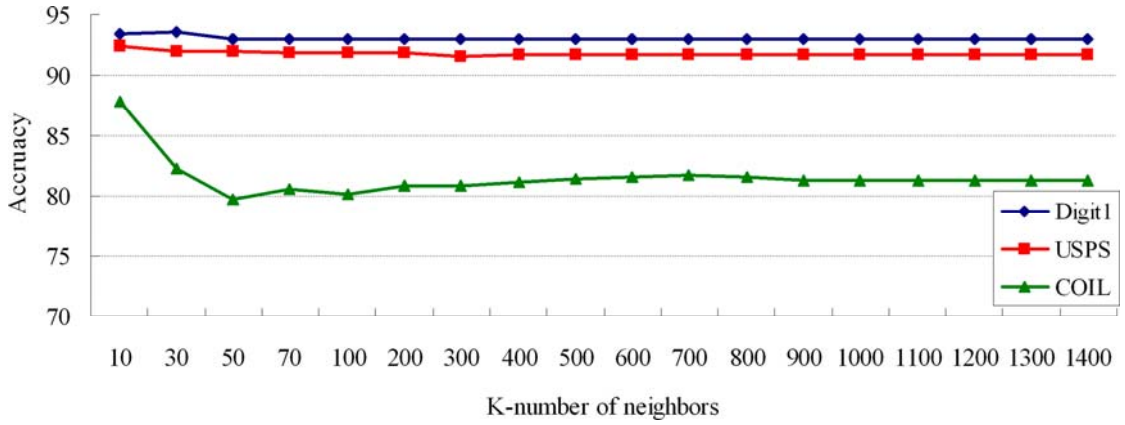

图 6.5 NMSNN 在 *manifold-like* 数据集上的“*K*”值敏感性分析（100 个标记样本）

从图 6.5 中可以看出，在近邻数目较少的情况下，随着近邻的增加，NMSNN 方法的性能有所下降。然而，当 *K* 值达到一定程度后，性能几乎没有变化。这个实验结果表明，对于流形数据，在保证图是连通的前提下，*K* 值的设置越小越好。

这个现象可以用示意图 6.6 来解释。给定某一个 *K* 值，现假定未标记样本***‘u’***到 A 和 B 两类数据集中的某两个标记样本之间分别有两条最短路径 *Path* A_2 和 *Path* B_2。

由于路径 *Path* A_2 的代价长度要小于 *Path* B_2 的代价长度，未标记样本***‘u’***能被 NMSNN 方法正确的分到 A 类数据集。若 *K* 不断增大，性能将会逐渐降低。比如，当 *K* 为最大值 $n-1$ 时，其中 *n* 为数据样本的总数，这时所有的数据都直接连接，则

Path A_1 与 *Path* B_1 被 NMSNN 方法找到，这种情形中**‘*u*’**将被错误的分到数据集 B 中。

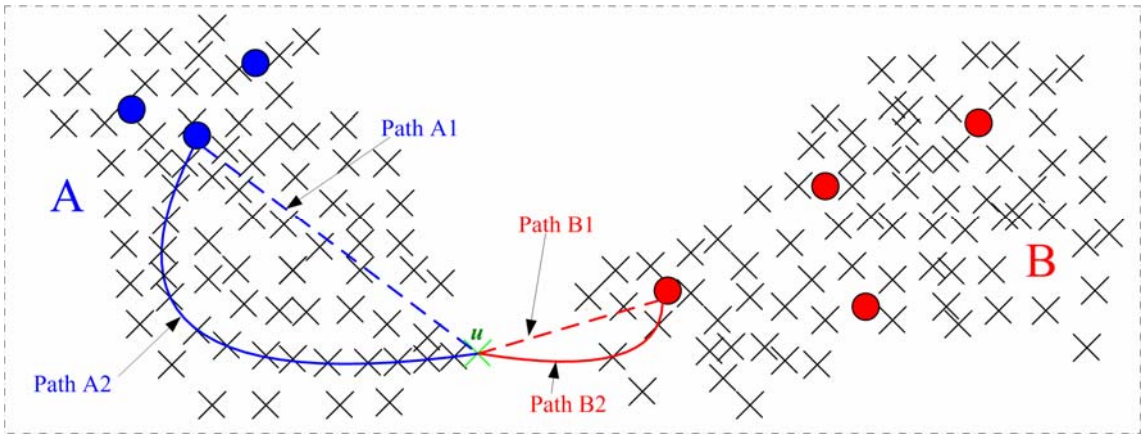

图 6.6 NMSNN 方法的示意图。数据集包含两类数据，A 与 B 来表示。‘○’与‘×’分别表示标记样本和未标记样本。‘*u*’（绿色）表示一个待分类的未标记样本，A 与 B 中分别有 3 个标记样本和（蓝色‘○’）和 4 个标记样本（红色‘○’）。

图 6.7 给出了 NMSNN 方法在 *cluster-like* 数据集上对 K 良好的稳定性。比如，对数据集 *g241c* 和 *g241d*，当 $K>50$ 时，NMSNN 方法的分类性能几乎没有变化。这个现象与在 *manifold-like* 数据集上的表现很相似。但对于 *Text* 数据，随着近邻 K 值的增加分类性能会有增加。

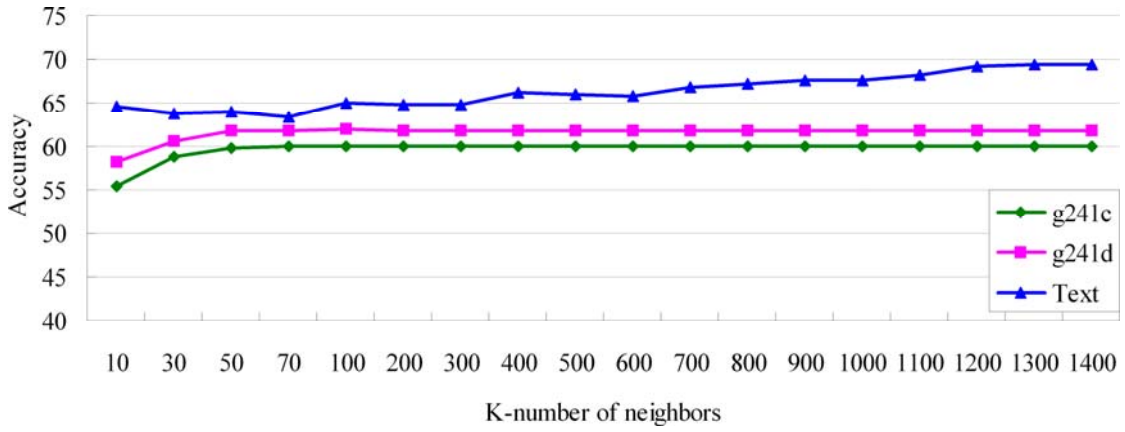

图 6.7 NMSNN 在 *cluster-like* 数据集上的“K”值敏感性分析（100 个标记样本）

作者认为这与 *Text* 的维数有关，*Text* 数据维数为 11960，属于超大维数据集。

因此，采用欧氏距离在高维空间中度量 K 近邻并不合适，其原因是受到维数灾难（Curses of Dimensionality）[41]的制约。

6.6 本章小结

本章提出了一种新的非公度近邻半监督分类方法 NMSNN，该方法利用最短路径的代价长度整合了大量未标记样本的内在几何结构信息和少量标记样本的标记信息，并给出了具体的理论分析。NMSNN 方法的使用很简单，只需调节一个近邻参数 K，不需要进行很多繁琐的参数调节工作；将 NMSNN 方法与经典方法半监督学习方法 LLGC 和有监督分类方法 1-NN，以及 MSNN 方法在 2 个合成数据集和 6 个标准数据集（*manifold-like* 和 *cluster-like*）上进行了比较，实验结果表明 NMSNN 具有良好的分类效果。最后分析了参数 K 以及标记样本数目对 NMSNN 方法的影响，实验结果表明 NMSNN 方法具有良好的稳定性，特别是对 *manifold-like* 类型的数据集。

7 总结与展望

7.1 主要工作总结

如何发掘反映数据本质的特征以及尽可能的利用少量人工干预信息进行分类决策一直是模式分类、机器学习和数据挖掘的重点研究内容。特征选择和半监督分类研究作为其有效地解决方法而被广泛研究。本文针对特征选择和半监督分类需解决的几个问题，即：选取表征能力强的特征子集，协同训练标记样本不充分问题，多个弱分类器集成解决半监督分类问题，充分利用未标记样本的几何结构信息改善分类器精度问题，作了几个方面的研究，总结如下：

（1）在第 2 章，本文详细介绍了特征选择和机器学习中三种学习类型（有监督学习、无监督学习、半监督学习）的基本概念，并系统地综述了特征选择和半监督分类的研究现状及进展，分类整理了国内外众多学者在此领域所取得的研究成果。

（2）在第 3 章，本文分析了现有特征选择方法的不足，提出了一种能直接选取具有强类别区分能力的特征子集选取方法 FSCRF 方法。实验结果表明，该方法相比其他常用方法，能在大多数情况下有效地选出特征数目更少、分类精度更高的特征子集，并将该方法应用于老年痴呆诊断，给出了老年痴呆疾病的逻辑回归计算模型，取得了比同类方法更令人满意的结果。

（3）在第 4 章，本文指出了协同训练方法存在的相关问题，在此基础上，提出了一种新的交叉训练半监督分类方法 NC-T。实验结果表明，NC-T 方法的分类精度比标准协同方法在多数情况有所提高，并且不需要假设数据特征存在两个或多个独立特征视图，弱化了假设前提，实用性得到了增强。

（4）在第 5 章，本文分析了多分类器集成策略的优势，指出了半监督分类学习中多分类器集成的必要性，并提出了一种多类别多分类器集成半监督分类方法 SSMAB，并给出了充分的理论分析及求解证明。该方法只需要基分类器的分类精度达到$1/K$（K为类别数目），就能取得较满意的效果，实验结果表明，它比同类型的方法在多数情况下占有优势。

（5）在第 6 章，本文提出了一种非公度的半监督学习方法 NMSNN，该方法能

有效利用未标记样本的几何结构信息和标记样本的类别信息，且只需要设置一个参数，简单有效。实验结果表明，NMSNN 方法具有良好的分类精度和稳定性。

7.2 研究展望

本文所提出的思想、方法和模型均已在相应的数据集上经过了实验验证，有的已经开始了实际应用。除了前一节所研究的几个问题之外，本文认为，特征选择和半监督分类的理论和应用还存在许多局限性，还有很多方面需要完善和发展。本文认为，以下几个方面就是其中可以进一步展开研究的内容：

（1）关于特征选择的方法虽然已有了许多研究，但由于选取最优特征子集在理论上是 NP-hard 问题，因此特征选择的近似算法还有许多工作可以开展，特别是针对高维、高相关、高冗余、小样本的数据分析。此外，如何增强特征选择算法的稳定性，减少时间复杂度也有待进一步研究。

（2）协同训练类型的半监督方法随着训练的不断进行，自动标记的样本中噪声会逐渐积累，所产生的负面作用不断增加。因此，找到未标记样本导致性能下降的真正原因，有利于对半监督学习方法的进一步改进；此外对这类半监督学习方法的理论分析也是一个需要努力的方向。

（3）当前的半监督分类方法都是批量模式的方法，即，需要一次性获得所有的标记样本和未标记样本，但有些应用在学习之前是很难全部获取这些数据的，这些应用要求分类学习方法是实时的，能持续更新分类方法。因此在线模式的半监督学习方法是一个值得研究的问题。

（4）半监督分类方法通常是在标记样本较少的情况下从未标记样本中获得好处，这种好处在标记样本很少的情况下非常明显，但随着标记样本增加，这种优势越来越小。对于一个具体的半监督方法，需要的最少标记样本数目是具体是多少？这个问题虽然目前已经有相关学者开展了这方面的研究，但目前还没有非常明确的理论支持，有待进一步的研究与探讨。

致 谢

本文是对我攻读博士学位期间所作研究工作的一个总结，在这毕业之际，谨向关心支持我的老师、亲戚、朋友表示最衷心的感谢和诚挚的敬意！

首先我要衷心感谢我的导师宋恩民教授，本文研究方向的确定、论文的修改、定稿等大部分工作正是在宋老师的指导下完成的。宋老师从美国回国后一手创立了医学图像信息研究中心（CBIB，Center for Biomedical Imaging and Bioinformatics）。作为宋老师的第一届硕士生，亲眼见证了实验室的创建、发展和壮大。宋老师为实验室的发展付出了艰辛的努力，同时在我攻读硕士、博士学位的 6 年里，宋老师一直为我提供了很好的学习环境和研究机会，对我的成长、成才倾注了大量心血，所有这一切我将终生难忘。此外，宋老师严谨的治学态度、勤勉的工作作风以及宽广的国际视野，值得我终生学习。在此，对宋老师的精心栽培再次表示衷心的感谢！

特别感谢马光志副教授，本文从选题、查阅资料、论文构思、论文修改、定稿审阅，马老师都付出了大量的心血和劳动，并对本人公开发表的论文也都进行了反复细致的指导和修改。在平时的项目开发和科研讨论中，马老师总是平易近人，为人谦逊。本人在攻读博士学位期间遇到麻烦时，马老师也总是细心开导，排忧解惑。马老师丰富的科研项目经历和高度负责的工作精神是鞭策我未来不断进步的动力。

非常感谢南方理工州立大学（Southern Polytechnic State University）*Chih-Cheng Hung* 教授，*Hung* 教授对本人公开发表的英文学术论文进行了多次深入细致的修改，正是有了这些修改，论文才得以非常顺利发表。*Hung* 教授严谨的科学精神、认真负责的学术态度以及无私的奉献精神，使我深受感动、受益终身。

感谢华中科技大学校医院主任医生肖强为本文所提供的老年痴呆实验数据以及相关医学知识的讨论。

感谢在学习、生活中关心我、帮助我的 CBIB 实验室的老师们，他们是金人超教授、许向阳副教授、刘宏副教授、金良海副教授。

感谢美国斯坦福大学（Stanford University）*Yinyu Ye* 教授，加州大学伯克利分校（University of California Berkeley）*Shaofan Li* 教授，德州泛美大学（University of

Texas-Pan American）*Zhixiang Chen* 教授，杜克大学（Duke University）*Qiang Li* 教授，匹兹堡大学（University of Pittsburgh）*Bin Zheng* 教授，以及卡耐基梅隆大学（Carnegie Mellon University）的 *Yanxi Liu* 教授，英国伦敦大学（University of London）*Wen Wang* 教授，加拿大纽布郎斯威克大学（University of New Brunswick）*Huajie Zhang* 副教授，新西兰坎特伯雷大学（University of Canterbury）*Tim Bell* 教授，天津大学（Tianjin University）李刚教授。他们开阔的研究思路、敏锐的学术眼光，极大的开拓了我的研究视野，并为我的博士论文提供了前沿的研究方向，使我博士期间的研究少走了许多弯路。

感谢先后进入 CBIB 实验室的各位博士、硕士同学，他们是杨词慧、邹耀斌、徐胜舟、张国军、刘晶晶、兰义华、李向、胡怀飞、余玛俐、潘宁、王杰、李磊，还有王倩、姜娈、陈晓林、李明、杨艳屏、闵志方；特别感谢代大攀、蔡冲、方红霞、占慧融、刘新鸣、周渊、张魁、汤海先、伍胜、周涛、郦琪君等在课题研究中给予的具体支持。感谢他们给我带来的欢乐与鼓励，难忘的缘聚，永恒的回忆。

感谢我的父母在漫长求学生涯中给予的许多关心与爱护，他们供我从小学念到博士，对我的成长付出太多太多。感谢我的弟弟黄红山、弟妹肖利对我一直以来学业上的支持和鼓励。

最后，我要非常感谢我的妻子李春花对我的理解、支持与爱。博士期间，当我遇到挫折时，她总是给予我鼓励与安慰，陪我共度每一个难关。默默的支持是我顺利完成博士学业的有力保障。

参考文献

[1] Raudys S J, Jain A K. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991, 13 (3): 252-264

[2] Fukunaga K. *Introduction to statistical pattern recognition*. New York: Academic Press, 1990.

[3] Wu P-S, Müller H-G. Functional embedding for the classification of gene expression profiles. *Bioinformatics*, 2010, 26 (4): 509-517

[4] Liu H, Liu L, Zhang H. Ensemble gene selection for cancer classification. *Pattern Recognition*, 2010, 43 (8): 2763-2772

[5] Antoniadis A, Lambert-Lacroix S, Leblanc F. Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, 2003, 19 (5): 563-570

[6] Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 2005, 21 (13): 3001-3008

[7] Qi X, Davison B D. Web page classification: Features and algorithms. *ACM Computing Surveys (CSUR)*, 2009, 41 (2): 1-31

[8] Crosier M, Griffin L. Using Basic Image Features for Texture Classification. *International Journal of Computer Vision*, 2010, 88 (3): 447-460

[9] Moosmann F, Nowak E, Jurie F. Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, 30 (9): 1632-1646

[10] Zhao W, Chellappa R, Phillips P J, et al. Face recognition: A literature survey. *Acm*

Computing Surveys (CSUR), 2003, 35 (4): 399-458

[11] O'Shaughnessy D. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 2008, 41 (10): 2965-2979

[12] Zhu X, Goldberg A B. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2009, 3 (1): 1-130

[13] Bourgon R, Mancera E, Brozzi A, et al. Array-based genotyping in S.cerevisiae using semi-supervised clustering. *Bioinformatics*, 2009, 25 (8): 1056-1062

[14] Kang B-Y, Ko S, Kim D-W. SICAGO: Semi-supervised cluster analysis using semantic distance between gene pairs in Gene Ontology. *Bioinformatics*, 2010, 26 (10): 1384-1385

[15] Koestler D C, Marsit C J, Christensen B C, et al. Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. *Bioinformatics*, 2010, 26 (20): 2578-2585

[16] Qi Y, Tastan O, Carbonell J G, et al. Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics*, 2010, 26 (18): i645-i652

[17] Socher R, Fei-Fei L. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. San Francisco, USA: IEEE, 2010: 966-973

[18] Wu W, Yang J. Semi-automatically labeling objects in images. *IEEE Transactions on Image Processing*, 2009, 18 (6): 1340-1349

[19] Lehmann F. Turbo Segmentation of Textured Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 33 (1): 16-29

[20] Filipovych R, Resnick S M, Davatzikos C. Semi-supervised cluster analysis of imaging data. *NeuroImage*, 2011, 54 (3): 2185-2197

[21] Zhu X, Rogers T, Qian R, et al. Humans perform semi-supervised classification too.

In: *Proceedings: 22nd AAAI Conference on Artificial Intelligence and the 19th Innovative Applications of Artificial Intelligence Conference*. Vancouver, BC, Canada: AAAI, 2007: 864-869

[22] John G H, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. In: *Proceedings of the Eleventh International Conference on Machine Learning*. Rutgers University, New Brunswick, NJ, USA: Morgan Kaufmann, 1994: 121-129

[23] Liu H, Motoda H. *Computational methods of feature selection*: Chapman & Hall/CRC Boca Raton, 2008.

[24] Kohavi R, John G H. Wrappers for feature subset selection. *Artificial intelligence*, 1997, 97 (1-2): 273-324

[25] Nguyen M H, de la Torre F. Optimal feature selection for support vector machines. *Pattern Recognition*, 2010, 43 (3): 584-591

[26] Monirul Kabir M, Monirul Islam M, Murase K. A new wrapper feature selection approach using neural network. *Neurocomputing*, 2010, 73 (16-18): 3273-3283

[27] Tuv E, Borisov A, Runger G, et al. Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research*, 2009, 10: 1341-1366

[28] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. *Machine learning*, 2002, 46 (1): 389-422

[29] Valiant L G. A theory of the learnable. *Communications of the ACM*, 1984, 27 (11): 1134-1142

[30] Cover T M, Thomas J A, Wiley J. *Elements of information theory(Second Edition)*: John Wiley and Sons, 2006.

[31] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 1997, 29: 103-130

[32] Ghahramani Z, Jordan M I. Supervised learning from incomplete data via an EM

approach. In: *Advances in Neural Information Processing Systems*. San Francisco, CA: Morgan Kaufmann, 1994: 120-127

[33] Rue H, Held L. *Gaussian Markov random fields: theory and applications*: Chapman & Hall, 2005.

[34] Hilbe J. *Logistic regression models*: Chapman & Hall/CRC, 2009.

[35] Vapnik V N. *The nature of statistical learning theory*: Springer Verlag, 2000.

[36] Haykin S. *Neural networks: a comprehensive foundation (2nd Edition)* Prentice Hall PTR Upper Saddle River, NJ, USA, 1998.

[37] Alexander G, Katy S A, Vladimir V. Learning by Transduction. In: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Madison, Wisconsin, USA: Morgan Kaufmann, 1998: 148-155

[38] Kulis B, Basu S, Dhillon I, et al. Semi-supervised graph clustering: a kernel approach. *Machine Learning*, 2009, 74 (1): 1-22

[39] Baghshah M S, Shouraki S B. Semi-supervised metric learning using pairwise constraints. In: *The 21st International Joint Conference on Artificial Intelligence*. Pasadena, CA, United states: IJCAI, 2009: 1217-1222

[40] Bellman R E, Corporation R. *Dynamic programming*: Princeton University Press, 1957.

[41] Donoho D L. high-dimensional data analysis: The curses and blessings of dimensionality. In: *American Math Society Lecture "Math Challenges of the 21st Century"*. University of California Los Angeles, Los Angeles, 2000

[42] Mitchell T M. *Machine learning*: WCB/Mac Graw Hill, 1997.

[43] Silverman B W. *Density estimation for statistics and data analysis*: Chapman & Hall/CRC, 1998.

[44] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2005: 491-502

[45] Amaldi E, Kann V. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 1998, 209 (1-2): 237-260

[46] Loscalzo S, Yu L, Ding C. Consensus group stable feature selection. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009: 567-576

[47] Yu L, Ding C, Loscalzo S. Stable feature selection via dense feature groups. In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008: 803-811

[48] Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 2007, 12 (1): 95-116

[49] Davis C A, Gerick F, Hintermair V, et al. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 2006, 22 (19): 2356-2363

[50] Boulesteix A L, Slawski M. Stability and aggregation of ranked gene lists. *Briefings in bioinformatics*, 2009, 10 (5): 556-568

[51] Guyon I, Elisseeff A. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 2003, 3: 1157-1182

[52] Saeys Y, Inza I, Larra aga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 2007, 23 (19): 2507-2517

[53] Guyon I, Gunn S, Nikravesh M, et al. *Feature Extraction: Foundations and Applications*: Springer, 2006.

[54] Kira K, Rendell L A. The feature selection problem: Traditional methods and a new algorithm. In: JOHN WILEY & SONS LTD, 1992: 129-129

[55] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. In: *The Proceedings of the European Conference on Machine Learning* Springer, 1994:

171-182

[56] Sun Y. Iterative RELIEF for feature weighting: Algorithms, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29 (6): 1035-1051

[57] Kullback S, Leibler R A. On information and sufficiency. *The Annals of Mathematical Statistics*, 1951, 22 (1): 79-86

[58] Song L, Smola A, Gretton A, et al. Supervised feature selection via dependence estimation. In: *Proceedings of the 24th international conference on Machine learning*. ACM, 2007: 823-830

[59] Koller D, Sahami M. Toward optimal feature selection. In: *The Thirteenth International Conference in Machine Learning*. Bari, Italy: Morgan Kaufmann, 1996: 284-292

[60] Yu L, Liu H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In: *Proceedings, Twentieth International Conference on Machine Learning*. Washington, DC, United States: AAAI, 2003: 856-863

[61] 崔自峰, 徐宝文, 张卫丰等. 一种近似 Markov Blanket 最优特征选择算法. *计算机学报*, 2007, (12): 2074-2081

[62] Hall M A. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In: *The 17th International Conference on Machine Learning*. Stanford, CA, USA: Morgan Kaufmann, 2000: 359-366

[63] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27 (8): 1226-1238

[64] Dy J G, Brodley C E. Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, 2004, 5: 845-889

[65] Talavera L. Feature selection as a preprocessing step for hierarchical clustering. In:

Proceedings of the Sixteenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA: Citeseer, 1999: 389-397

[66] Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24 (3): 301-312

[67] Butterworth R, Piatetsky-Shapiro G, Simovici D A. On feature selection through clustering. In: *5th IEEE International Conference on Data Mining*. Houston, TX, United states: IEEE, 2005: 581-584

[68] Law M H C, Figueiredo M A T, Jain A K. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26 (9): 1154-1166

[69] Yuanhong L, Ming D, Jing H. Simultaneous Localized Feature Selection and Model Detection for Gaussian Mixtures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, 31 (5): 953-960

[70] Hou Y X, Zhang P, Yan T X, et al. Beyond Redundancies: A Metric-Invariant Method for Unsupervised Feature Selection. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22 (3): 348-364

[71] Dy J G, Brodley C E. Feature subset selection and order identification for unsupervised learning. In: *Proc. 17th International Conf. on Machine Learning*. Stanford University, USA: Morgan Kaufmann, 2000: 247-254

[72] Zeng H, Cheung Y M. A new feature selection method for Gaussian mixture clustering. *Pattern Recognition*, 2009, 42 (2): 243-250

[73] Handl J, Knowles J. Feature subset selection in unsupervised learning via multiobjective optimization. *International Journal of Computational Intelligence Research*, 2006, 2 (3): 217-238

[74] He X, Cai D, Niyogi P. Laplacian score for feature selection. *Advances in Neural Information Processing Systems*, 2006, 18: 507-514

[75] Wu Z, Li C. Feature Selection using Transductive Support Vector Machine. In: *Proc. NIPS Workshop Feature Selection*. 2003:

[76] Joachims T. Transductive Inference for Text Classification using Support Vector Machines. In: *Proceedings of the Sixteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 1999: 209

[77] Zhao Z, Liu H. Semi-supervised feature selection via spectral analysis. In: *Proceedings of the 7th SIAM International Conference on Data Mining, Minneapolis, MN*. 2007: 1151–1158

[78] Yubo C, Yunpeng C, Yijun S, et al. Semi-supervised feature selection under logistic I-RELIEF framework. In: *The 19th Conference of the International Association for Pattern Recognition*. Orlando, Florida, USA: IEEE, 2008: 1-4

[79] Sun Y, Todorovic S, Goodison S. A feature selection algorithm capable of handling extremely large data dimensionality. In: *Proc. 8th SIAM International Conference on Data Mining*. 2008: 530–540

[80] Ren J, Qiu Z, Fan W, et al. Forward Semi-supervised Feature Selection. In: *The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)* Osaka, Japan, 2008: 970-976

[81] Yaslan Y, Cataltepe Z. Co-training with relevant random subspaces. *Neurocomputing*, 2010, 73 (10-12): 1652-1661

[82] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: *Proceedings of the eleventh annual conference on Computational learning theory*. Madison, Wisconsin, USA: ACM, 1998: 92-100

[83] Zhao J, Lu K, He X. Locality sensitive semi-supervised feature selection. *Neurocomputing*, 2008, 71 (10-12): 1842-1849

[84] Zhang D, Chen S, Zhou Z H. Constraint Score: A new filter method for feature selection with pairwise constraints. *Pattern Recognition*, 2008, 41 (5): 1440-1451

[85] Sun D, Zhang D. Bagging Constraint Score for feature selection with pairwise

constraints. *Pattern Recognition*, 2009, 43 (6): 2106-2118

[86] Narendra P M, Fukunaga K. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, 1977, 100 (9): 917-922

[87] Chen X. An improved branch and bound algorithm for feature selection. *Pattern Recognition Letters*, 2003, 24 (12): 1925-1933

[88] Whitney A W. A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, 1971, 100 (20): 1100-1103

[89] Marill T, Green D. On the effectiveness of receptors in recognition systems. *IEEE transactions on Information Theory*, 1963, 9 (1): 11-17

[90] Stearns S D. On selecting features for pattern classifiers. In: *Proc. of the 3rd Int. Joint Conf. on Pattern Recognition*. 1976: 71–75

[91] Pudil P, Novoviová J, Kittler J. Floating search methods in feature selection. *Pattern recognition letters*, 1994, 15 (11): 1119-1125

[92] Zhang H, Sun G. Feature selection using tabu search method. *Pattern Recognition*, 2002, 35 (3): 701-712

[93] Meiri R, Zahavi J. Using simulated annealing to optimize the feature selection problem in marketing applications. *European Journal of Operational Research*, 2006, 171 (3): 842-858

[94] Liu H, Setiono R. A probabilistic approach to feature selection-a filter solution. In: *Proceedings of the 13th International Conference on Machine Learning*. Bari, Italy Morgan Kaufmann, 1996: 319-327

[95] Yang J, Honavar V. Feature subset selection using a genetic algorithm. *Intelligent Systems and Their Applications, IEEE*, 1998, 13 (2): 44-49

[96] Gheyas I A, Smith L S. Feature subset selection in large dimensionality domains. *Pattern Recognition*, 2010, 43 (1): 5-13

[97] Kira K, Rendell L A. The feature selection problem: Traditional methods and a new algorithm. In: *Proceedings 10th National Conference on Artificial Intelligence*.

Kanagawa, Japan: JOHN WILEY & SONS LTD, 1992: 129-134

[98] Liu H, Motoda H, Yu L. Feature selection with selective sampling. In: *Proceedings of the19th International Conference on Machine Learning*. Sydney, Australia, 2002: 395-402

[99] He X, Cai D, Niyogi P. Laplacian score for feature selection. In: *Advances in Neural Information Processing Systems*. 2006: 507-514

[100] Yubo C, Yunpeng C, Yijun S, et al. Semi-supervised feature selection under logistic I-RELIEF framework. In: *19th International Conference on Pattern Recognition* Tampa, Florida, USA: IEEE, 2008: 1-4

[101] Zhu Z, Ong Y S, Dash M. Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, 2007, 40 (11): 3236-3248

[102] Dash M, Choi K, Scheuermann P, et al. Feature selection for clustering-a filter solution. In: *Proceedings of the Second International Conference on Data Mining*. Maebashi City, Japan: IEEE, 2002: 115-122

[103] Bobrowski L. Feature selection based on some homogeneity coefficient. In: *The 9th International Conference on Pattern Recognition*. Rome, Italy: IEEE, 1988: 544-546

[104] Slonim N, Bejerano G, Fine S, et al. Discriminative feature selection via multiclass variable memory Markov model. *EURASIP Journal on Applied Signal Processing*, 2003, 2: 93-102

[105] Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002: 301-312

[106] Liu H, Motoda H, Dash M. A monotonic measure for optimal feature selection. In: *Proceedings of the 10th European Conference on Machine Learning*. Chemnitz, Germany: Springer, 1998: 101-106

[107] Zhang T. On the consistency of feature selection using greedy least squares

regression. *The Journal of Machine Learning Research*, 2009, 10: 555-568

[108] Foroutan I, Sklansky J. Feature selection for automatic classification of non-gaussian data. *IEEE Transactions on Systems, Man and Cybernetics*, 1987, 17 (2): 187-198

[109] Ichino M, Sklansky J. Optimum feature selection by zero-one integer programming. *IEEE Transactions on Systems, Man, and Cybernetics*, 1984, 14 (5): 737-746

[110] Stracuzzi D J, Utgoff P E. Randomized variable elimination. *The Journal of Machine Learning Research*, 2004, 5: 1331-1362

[111] Caruana R, Freitag D. Greedy attribute selection. In: *Proceedings of the 17th International Conference on Machine Learning*. New Brunswick, NJ, USA: Morgan Kaufmann, 1994: 28-36

[112] Domingos F. Control-sensitive feature selection for lazy learners. *Artificial Intelligence Review*, 1997, 11 (1): 227-253

[113] Devaney M, Ram A. Efficient feature selection in conceptual clustering. In: *Proceedings of the 14th International Conference on Machine Learning*. Nashville, TN, USA: Morgan Kaufmann, 1997: 92-97

[114] Kim Y S, Street W N, Menczer F. Feature selection in unsupervised learning via evolutionary search. In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Boston, MA, USA: ACM, 2000: 365-369

[115] Scudder III H. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 1965, 11 (3): 363-371

[116] Vapnik V N, Chervonenkis A Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 1971, 16: 264-280

[117] Hosmer D W. A comparison of iterative maximum likelihood estimates of the

parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, 1973, 29 (4): 761-770

[118] O'Neill T J. Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, 1978, 73 (364): 821-826

[119] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977, 39 (1): 1-38

[120] Nigam K, McCallum A K, Thrun S, et al. Text classification from labeled and unlabeled documents using EM. *Machine learning*, 2000, 39 (2): 103-134

[121] Larsen J, Szymkowiak A, Hansen L K. Probabilistic hierarchical clustering with labeled and unlabeled data. *International Journal of Knowledge-Based Intelligent Engineering Systems*, 2002, 6 (1): 56-62

[122] Fujino A, Ueda N, Saito K. Semisupervised Learning for a Hybrid Generative/Discriminative Classifier based on the Maximum Entropy Principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2008, 30 (3): 424-437

[123] Rosenberg C, Hebert M, Schneiderman H. Semi-supervised self-training of object detection models. In: *Proceedings of the Seventh IEEE Workshops on Application of Computer Vision* Brekenridge, CO, USA: IEEE 2005: 29-36

[124] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training. In: *The 9th International Conference on Information and Knowledge Management*. McLean, Virginia, United States: ACM, 2000: 86-93

[125] Muslea I, Minton S, Knoblock C A. Selective sampling with redundant views. In: *Proceedings of the 17th National Conference on Artificial Intelligence*. Austin, Texas, USA: AAAI, 2000: 621-626

[126] Muslea I, Minton S, Knoblock C A. Active+ Semi-supervised Learning= Robust Multi-View Learning. In: *Proceedings of the 19th International Conference on Machine Learning*. Sydney, Australia: Morgan Kaufmann, 2002: 435-442

[127] Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17 (11): 1529-1541

[128] Wang J, Luo S, Zeng X. A random subspace method for co-training. In: *In Proc.of the International Joint Conferenceon Neural Networks*. Hong Kong, China: IEEE, 2008: 195-200

[129] Blum A, Chawla S. Learning from Labeled and Unlabeled Data Using Graph Mincuts. In: *The 18th International Conference on Machine Learning*. Williamstown, MA, USA: Morgan Kaufmann, 2001: 19-26

[130] Zhu X, Lafferty J. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In: *Proceedings of the Twenty-Second International Conference on Machine Learning*. Bonn, Germany: ACM, 2005: 1052-1059

[131] Zhou D, Bousquet O, Lal T N, et al. Learning with Local and Global Consistency. In: *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT, 2004:

[132] Szummer M, Jaakkola T. Partially labeled classification with Markov random walks. In: *Advances in neural information processing systems*. British Columbia, Canada: MIT, 2002: 945-952

[133] Belkin M, Niyogi P. Using manifold structure for partially labeled classification. In: *Advances in Neural Information Processing Systems*. British Columbia, Canada: MIT, 2003: 929-936

[134] Wang F, Zhang C. Label propagation through linear neighborhoods. In: *Proceedings of the 23rd international conference on Machine learning*. Pittsburgh, Pennsylvania: ACM, 2006: 985-992

[135] Ando R K, Zhang T. Learning on Graph with Laplacian Regularization. In: *Advances in Neural Information Processing Systems*. Vancouver, B.C., Canada: MIT, 2007: 25-32

[136] Joachims T. Transductive learning via spectral graph partitioning. In: *Proceedings 20th International Conference on Machine Learning*. Washington, DC, USA: Morgan Kaufmann, 2003: 290-297

[137] Bennett K, Demiriz A. Semi-Supervised Support Vector Machines. In: *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT 1999: 368-374

[138] Xu L, Schuurmans D. Unsupervised and semi-supervised multi-class support vector machines. In: *The Twentieth National Conference on Artificial Intelligence*. Pittsburgh, Pennsylvania: AAAI, 2005: 904-910

[139] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 2006, 7: 2399-2434

[140] Zhou Z-H, Xu J-M. On the relation between multi-instance learning and semi-supervised learning. In: *Proceedings of the 24th International Conference on Machine learning*. Corvalis, Oregon: ACM, 2007: 1167-1174

[141] Zhu X, "*Semi-Supervised Learning Literature Survey*," Computer Science, University of Wisconsin-Madison 1530, 2008

[142] 周志华, 半监督学习中的协同训练风范, *机器学习及应用*, 周志华，王钰, 北京: 清华大学出版社, 2007, 259-275

[143] 肖宇,于剑. 基于近邻传播算法的半监督聚类. *软件学报*, 2008, (11): 2803-2813

[144] Zha Z J, Mei T, Wang J D, et al. Graph-based semi-supervised learning with multiple labels. *Journal of Visual Communication and Image Representation*, 2009, 20 (2): 97-103

[145] Subramanya A, Bilmes J. Large-Scale Graph-based Transductive Inference. In: *Workshop on Large-Scale Machine Learning: Parallelism and Massive datasets at Neural Information Processing Society*. Vancouver, Canada, 2009:

[146] Joachims T. Transductive inference for text classification using support vector

machines. In: *Proc. 16th International Conf. on Machine Learning*. Bled, Slovenia: Morgan Kaufmann, 1999: 200-209

[147] Chapelle O, Zien A. Semi-supervised classification by low density separation. In: *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*. Barbados, West Indies, 2005: 57-64

[148] Sindhwani V, Keerthi S S. Large scale semi-supervised linear SVMs. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM 2006: 477-484

[149] Weston J, Collobert R, Sinz F, et al. Inference with the universum. In: *Proceedings of the 23rd International Conference on Machine learning*. Pittsburgh, PA, USA: ACM, 2006: 1009-1016

[150] Chapelle O, Sindhwani V, Keerthi S S. Optimization techniques for semi-supervised support vector machines. *The Journal of Machine Learning Research*, 2008, 9: 203-233

[151] Collobert R, Sinz F, Weston J, et al. Trading convexity for scalability. In: *The 23rd International Conference onMachine Learning*. Pittsburgh, USA: ACM, 2006: 201-208

[152] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. In: L D R a F Bergadano.*Machine Learning(ECML)*. Catania: Springer, 1994: 171-182

[153] Robnik-Šikonja M, Kononenko I. An adaptation of Relief for attribute estimation in regression. In: *Machine Learning: Proceedings of the Fourteenth International Conference*. Tennessee: Morgan Kaufmann, 1997: 296-304

[154] Robnik- ikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 2003, 53 (1): 23-69

[155] Kira K, Rendell L A. The feature selection problem: Traditional methods and a new algorithm. In: *Proceedings of the National Conference on Artificial Intelligence*. San Jose: JOHN WILEY & SONS LTD, 1992: 129-129

[156] 张丽新, 王家廞, 赵雁南等. 基于 Relief 的组合式特征选择. *复旦学报(自然科学版)*, 2004, (05): 893-898

[157] Hinneburg A, Aggarwal C C, Keim D A. What is the nearest neighbor in high dimensional spaces. In: *Proceedings of the 26th International Conference on Very Large Data Bases*. Cairo: Morgan Kaufmann, 2000: 506-515

[158] Beyer K, Goldstein J, Ramakrishnan R, et al. When Is Nearest Neighbor Meaningful? In: *the International Conference on Database Theory*. Jerusalem, Israel: Springer, 1999: 217-235

[159] Frank A, Asuncion A. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2010

[160] Mark H, Eibe F, Geoffrey H, et al. The WEKA data mining software: an update. *SIGKDD Explorations*, 2009, 11 (1): 10-18

[161] Quinlan J R. Induction of decision trees. *Machine Learning*, 1986, 1 (1): 81-106

[162] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, 1997, 29 (2): 103-130

[163] Landwehr N, Hall M, Frank E. Logistic model trees. *Machine Learning*, 2005, 59 (1): 161-205

[164] Dasgupta S, Littman M L, McAllester D. PAC generalization bounds for co-training. In: *Advances in Neural Information Processing Systems*. MIT Press, 2002: 375-382

[165] Balcan M F, Blum A. A PAC-style model for learning from labeled and unlabeled data. In: *The 18th Annual Conference on Learning Theory*. Springer, 2005: 111-126

[166] Brefeld U, Scheffer T. Co-EM support vector learning. In: *The Proceedings of The 21st International Conference on Machine Learning*. ACM, 2004: 121-128

[167] Abney S. Bootstrapping. In: *Proceedings of the 40th Annual Meeting of*

Association for Computational Linguistics. ACL, 2002: 360-367

[168] Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data. In: *Proceedings of the 17th International Conference on Machine Learning*. 2000: 327-334

[169] Soonthornphisaj N, Kijsirikul B. Iterative cross-training: An algorithm for learning from unlabeled Web pages. *International Journal of Intelligent Systems*, 2004, 19 (1-2): 131-147

[170] Quinlan J. Improved Use of Continuous Attributes in C4. 5. *Journal of Artificial Intelligence Research*, 1996, 4 (1): 77-90

[171] D eroski S, enko B. Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 2004, 54 (3): 255-273

[172] Opitz D, Maclin R. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 1999, 11 (1): 169-198

[173] Dietterich T G. Machine-learning research : Four Current Directions *AI magazine*, 1997, 18 (4): 97-136

[174] Bartels L M. Specification Uncertainty and Model Averaging. *American Journal of Political Science*, 1997, 41 (2): 641-674

[175] Montgomery J M, Nyhan B. Bayesian Model Averaging: Theoretical developments and practical applications. *Political Analysis*, 2010, 18 (2): 245-270

[176] Schapire R E. The strength of weak learnability. *Machine learning*, 1990, 5 (2): 197-227

[177] Breiman L. Bagging predictors. *Machine learning*, 1996, 24 (2): 123-140

[178] Kearns M, Valiant L. Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the ACM (JACM)*, 1994, 41 (1): 67-95

[179] Freund Y, Schapire R E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 1997, 55 (1): 119-139

[180] Buc F A, Grandvalet Y, Ambroise C. Semi-supervised marginboost. In: *Advances in Neural Information Processing Systems*. Vancouver, British Columbia, Canada: MIT, 2001: 553-560

[181] Bennett K P, Demiriz A, Maclin R. Exploiting unlabeled data in ensemble methods. In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, Canada: ACM 2002: 289-296

[182] Mallapragada P, Jin R, Jain A, et al. Semiboost: Boosting for semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, 31 (11): 2000-2014

[183] Zhu J, Rosset S, Zou H, et al. Multi-class adaboost. *Statistics and Its Interface*, 2009, 2: 349-360

[184] John G H, Langley P. Estimating continuous distributions in Bayesian classifiers. In: *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. Montreal, Quebec, Canada: Morgan Kaufmann, 1995: 338-345

[185] Singh A, Nowak R D, Zhu X. Unlabeled data: Now it helps, now it doesn't. In: *Neural Information Processing Systems*. Vancouver, British Columbia, Canada: Curran Associates, Inc., 2008: 1513-1520

[186] Tata S, Patel J M. Estimating the selectivity of tf-idf based cosine similarity predicates. *ACM SIGMOD Record*, 2007, 36 (4): 75-80

[187] Ackermann M R, Bl mer J, Sohler C. Clustering for metric and non-metric distance measures. In: *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2008: 799-808

[188] Jacobs D W, Weinshall D, Gdalyahu Y. Classification with nonmetric distances: Image retrieval and class representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, 22 (6): 583-600

[189] Ma Y, Lao S, Takikawa E, et al. Discriminant analysis in correlation similarity measure space. In: *Proceedings of the 24th international conference on Machine*

learning (ICML). New York, USA, 2007: 577-584

[190] Tan X, Chen S, Zhou Z H, et al. Face recognition under occlusions and variant expressions with partial similarity. *IEEE Trans. Information Forensics and Security*, 2009, 4 (2): 217-230

[191] Tenenbaum J B, Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000, 290 (5500): 2319-2323

[192] Varini C, Degenhard A, Nattkemper T W. ISOLLE: LLE with geodesic distance. *Neurocomputing*, 2006, 69 (13-15): 1768-1771

[193] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, 290 (5500): 2323-2326

[194] Cruz-Barbosa R, Vellido A. Semi-supervised geodesic Generative Topographic Mapping. *Pattern Recognition Letters*, 2010, 31 (3): 202-209

[195] Lebanon G. Learning Riemannian Metrics. In: *Proc. of the 19th Conference on Uncertainty in Artificial Intelligence (UAI)*. Acapulco, Mexico, 2003: 362-369

[196] Fischer B, Roth V, Buhmann J M. Clustering with the connectivity kernel. In: *Neural Information Processing Systems*. Cambridge, MA: MIT, 2004: 89-96

[197] Chang H, Yeung D Y. Robust path-based spectral clustering. *Pattern recognition*, 2008, 41 (1): 191-203

[198] Chapelle O, Sch¨olkopf B, Zien A. *Semi-Supervised Learning*. Cambridge, Massachusetts London, England: The MIT Press, 2006

[199] Ben-David S, Lu T, Pál D. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In: *Proceedings of the 21th Annual Conference on Learning Theory (COLT)*. Helsinki, Finland, 2008: 33–44

[200] Magdon-Ismail M. No free lunch for noise prediction. *Neural Computation*, 2000, 12 (3): 547-564

附录 1 攻读博士学位期间发表及录用的论文目录

[1] Enmin Song, **Dongshan Huang**, Guangzhi Ma, Chih-Cheng Hung. Semi-supervised multi-class Adaboost by exploiting unlabeled data. *Expert System with Application*, 2011, 38 (6): 6720-6726.(第一署名单位：华中科技大学计算机科学与技术学院，**SCI，收录 IF 2.908**)

[2] **Dongshan Huang**, Enmin Song, Guangzhi Ma, Dapan Dai, Chih-Cheng Hung, Feature Selection: A feature subset category resolve based filter approach, *Journal of Information*. (第一署名单位：华中科技大学计算机科学与技术学院，**SCIE，In press**)

[3] **Dongshan Huang**, Enmin Song, Guangzhi Ma, Non-metric based semi-supervised nearest neighbor classification, *Journal of Experimental & Theoretical Artificial Intelligence*.(第一署名单位：华中科技大学计算机科学与技术学院，**SCI**，**Submitted)**

[4] Guangzhi Ma, Chih-Cheng Hung, Li Su, **Dongshan Huang**, Multiple costs based decision making with back-propagation neural network, *Decision Support Systems*.（第一署名单位：华中科技大学计算机科学与技术学院，**SCI**，**有条件接收**）

[5] **Dongshan Huang**, Enmin Song, Guangzhi Ma, Huirong Zhan, Chih-Cheng Hung. A new cross-training approach by using labeled data. In: *Proceedings of the 24th Annual ACM symposium on Applied Computing*. Honolulu, Hawaii,USA: ACM, 941-942, 2009.(第一署名单位：华中科技大学计算机科学与技术学院，**EI，收录)**

[6] 宋恩民, **黄东山**, 马光志, 肖强. 评估子集类区分能力的特征选择方法. 华中科技大学学报(自然科学版), 2011, 39 (02): 1-5.(第一署名单位：华中科技大学计算机科学与技术学院，**EI，收录)**

[7] Xinqiao Lv, **Dongshan Huang**, Enming Song, Ping Li, Chunshan Wu, One radical-based on-Line chinese character recognition (OLCCR) system using support vector machine for recognition of radicals. In: *The International Conference on Bioinformatics and Biomedical Engineering*. IEEE, 2007.558-561.(第一署名单位：华中科技大学计算机科学与技术学院，**EI，收录)**

[1] **国家软件著作版权:** 主观导向多维数据分析系统，登记号 2008SR33782，申请日：2008 年 10 月 6 日，版本号 V1.0，申请单位：华中科技大学。

附录 2 攻读博士学位期间参加的科研项目

课题名称 1：基于网格的数字化医疗决策支持系统

课题来源：国家科技部 **863** 项目（2006AA02Z347）

起始时间：2007.06-2009.07

主要职责：项目主要研发者，负责决策支持、数据可视化分析等模块的方案设计、算法分析及相关材料的整理

项目结果：2009 年顺利通过国家科技部验收，并获 2010 年湖北省科技进步二等奖

课题名称 2：面向医学应用的多模型数据分析系统

课题来源：校内合作预研课题

起始时间：2009.07-2010.11

主要内容：与华中科技大学协和医院合作建立了颈椎骨龄定量分期模型(153 个检测指标)；与华中科技大学校医院（53 个检测指标）建立了老年痴呆综合评价量化模型，自动筛选少量有效指标，建立较高精度的诊断模型。

主要职责：项目主要研发者，项目的总体框架设计、特征选择、分类识别算法研究及编码实现等。

项目结果：目前已经开发出系统的第 1 个版本，并成功的用于老年痴呆数据建模分析等实际应用。

课题名称 3：光谱舌像分析与疾病诊断系统

课题来源：与天津大学合作预研课题

起始时间：2010.11-2011.05

主要内容：包括高光谱舌像信息系统建立、预处理、特征提取和辅助诊断几部分

主要职责：项目主要研发者，项目的总体框架设计、特征选择、分类诊断算法研究

项目结果：目前系统已经在高光谱光学仪器上开发出软件原型系统，并已能自动采集数据和进行初步辅助诊断。

特征选择及半监督分类方法研究

作者： 黄东山
学位授予单位： 华中科技大学