数据挖掘与组合学习*)

Data Mining and Ensemble of Learning Methods

刁力力　胡可云　陆玉昌　石纯一

(清华大学计算机科学与技术系智能技术与系统国家重点实验室　北京100084)

Abstract　Data-mining is a kind of solution for solving the problem of information exploding. Classification and prediction belong to the most fundamental tasks in data-mining field. Many experiments have showed that the results of ensemble of learning methods are generally better than those of single learning methods under most of the time. In the sense, it is of great value to introduce ensemble of learning methods to data mining. This paper introduces data mining and ensemble of learning methods respectively, along with the analysis and formulation about the role ensemble of learning methods can act in some important practicing aspects of data mining: Text mining, multi-media information mining and web mining.

Keywords　Data-mining, Classification/prediction, Ensemble of learning methods

一、概述

数据挖掘是信息爆炸问题的一种解决方案。在数据挖掘中，分类预测是最基本的任务之一。组合学习器是分类预测器的集合，这些分类器的单独决策被以某种方式组合起来(典型的方法是通过加权或无权重投票)以给新样本分类。实验表明，组合方法在多数情况下比单个分类预测方法要精确。因此，在数据挖掘中引入组合学习方法具有相当的现实意义。组合学习方法应用在数据挖掘的重要应用方面——文本、多媒体和网络信息挖掘中的初步实践验证了这一点。

二、数据挖掘

2.1　数据挖掘介绍

当前，在各种企业，商业领域中的交易记录与财务报表，科研领域所收集的数据(例如，气象卫星传回的气象图像)，其数据规模经常在数十兆、甚至上千兆字节。面对"堆积如山"的数据集合，传统的数据分析手段难以应付，因此，人们无法理解并有效使用这些数据。另外，传统的数据分析方法(比如统计)只能获得数据的表层信息，无法获得数据属性的内在关系和隐含的信息。这样，快速的数据产生与搜集技术和拙劣的数据分析方法之间形成了鲜明的对照，这就需要新的技术来智能地和自动地分析这些原始数据，以使消耗大量财力与物力所收集与整理的宝贵资源——数据得以利用。这就是数据挖掘和KDD技术产生的背景。KDD即数据库中的知识发现，指识别出存在于数据库中有效的、新颖的、具有潜在效用的、最终可理解的模式的非平凡过程。KDD是应用特定的数据挖掘算法和评价解释模式的一个循环反复过程，并要对发现的知识不断

*)本文得到中国国家自然科学基金重大项目(编号：79990580)、中国国家重点基础研究发展项目(编号：G1998030414)和清华大学信息学院基础研究项目资助。

Planning. In IJCAI-99, 1999

14 Murphy K, Ralston E, Friedlander D, Swab R, Steege P. The Scheduling of Rail at Union Pacific Railroad. In, Proc of AAAI Conf. 1997. 903～912

15 Nebel B, Dimopoulos Y, Koehler J. Ignoring irrelevant facts and operators in plan generation. In: Proc. of the 4th European Conf. on Planning, 1997

16 Nguyen XuanLong, Kambhampati S. Extracting effective and admissible State Space heuristics from the Planning Graph. To Appear in AAAI-00, 2000

17 Pell B, et al. An Autonomous Spacecraft Agent Prototype. In: Proc. 1th Intl. Conf. Autonomous Agents, 1997. 253～261

18 Smith D E, Weld D S. Conformant Graphplan. In: Proc. 15th National Conf. On AI, 1998

19 Smith D E, Weld D S. Temporal Planning with Mutual Exclusion Reasoning. In IJCAI-99, 1999

20 Weld D S, Anderson C R, Smith D E. Extending Graphplan to handle uncertainty and sensing actions. In AAAI-98, 1998

21 www.cs.toronto.edu/aips-2000/selfContainedAips-2000.ps

精深化，使其易于理解。数据挖掘是KDD的一个关键步骤，KDD利用数据挖掘算法，按指定方式和阈值抽取有价值的知识，包括数据挖掘前对数据的预处理、抽样及转换和数据挖掘后对知识的评价解释过程等。数据挖掘是人工智能、机器学习与数据库技术相结合的产物。它大体上有两种功能：预测/验证功能和描述功能，即用数据库的若干已知字段预测或验证其它未知字段值，及找到描述数据的可理解模式。具体地说，主要包括以下几个方面：数据分类、回归分析、聚类、概括、构造依赖模式、变化和偏差分析、模式发现、路径发现等。

大部分数据挖掘的方法都基于机器学习、模式识别、统计学等领域，它们分别从不同的角度进行数据挖掘：(1)关联规则开采。关联可分为简单关联、时序关联和因果关联等，关联分析的目的是找出数据库中隐含的关联。(2)决策树方法。利用信息论中的互信息(信息增益)寻找数据库中具有最大信息量的字段，建立决策树的一个节点，再根据字段的不同取值建立树的分支；在每个分支子集中重复建树的下层节点和分支过程，即可建立决策树。国际上最有影响和最早的决策树方法是由Quinlan研制的ID3方法，后人又发展了多种决策树方法。(3)神经网络方法。它模拟人脑神经元结构，以MP模型和Hebb学习规则为基础，建立了三大类多种神经网络模型：前馈式网络、反馈式网络和自组织网络。神经网络的知识体现在网络连接的权值上，是一个分布式矩阵结构；神经网络的学习体现在神经网络的权值的逐步计算上(包括反复叠代或累加计算)。(4)覆盖正例排斥反例方法。利用覆盖所有正例、排斥所有反例的思想来寻找规则，比较典型的有MICHALSKI的AQII方法、洪家荣改进的AQ15方法和AE5方法。(5)粗糙集方法。在数据库中，将行元素看成对象，列元素看成属性(分为条件属性和决策属性)。等价关系R定义为不同对象在某个属性(或几个)上取值相同，这些满足等价关系的对象组成的集合称为该等价关系的等价类。条件属性上的等价类E与决策属性上的等价类Y之间有三种情况：①下近似：Y包含E；②上近似：Y和E的交非空；③无关：Y和E的交为空。(6)概念树方法。对数据库中记录的属性字段按归类方式进行抽象，建立起来的层次结构称为概念树。利用概念树的提升方法可以大大浓缩数据库中的记录。对多个属性字段的概念树进行提升，将得到高度概括的知识基表，然后再将它转换成规则。(7)遗传算法。这是模拟生物进化过程的算法，由三个基本算子组成：①繁殖(选择)是从一个旧种群(父代)选出生命力强的个体，产生新种群(后代)的过程；②交叉(重组)选择两个不同个体(染色体)的部分(基因)进行交换，形成新个体。③变异(突变)对某些个体的某些基因进行变异。这种遗传算法可起到产生优良后代的作用。这些后代需满足适应值，经过若干代的遗传，将得到满足要求的后代(问题的解)。遗传算法已在优化计算和分类机器学习方面发挥了显著的作用。(8)公式发现。在工程和科学数据库(由实验数据组成)中，对若干数据项(变量)进行一定的数学运算，求得相应的数学公式。比较典型的BACON发现系统完成了对物理学中大量定律的重新发现。(9)统计分析方法。在数据库字段项之间存在两种关系：①函数关系(能用函数公式表示的确定性关系)；②相关关系(不能用函数公式表示、但仍是相关确定的关系)。对它们的分析采用如下方法：回归分析、相关分析、主成分分析。(10)模糊论方法。利用模糊集合理论对实际问题进行模糊评判、模糊决策、模糊模式识别和模糊聚类分析。模糊性是客观存在的，系统的复杂性越高、精确化能力就越低，即模糊性越强。这就是Zadeh总结出来的互克性原理。(11)可视化技术。可视化数据分析技术拓宽了传统的图表功能，使用户对数据的剖析性更清楚。(12)另外还有归纳逻辑编程、贝叶斯网络等方法。

2.2 数据挖掘中机器学习和知识发现的难点：大规模和高维

为适应数据挖掘的要求，学习算法需要扩大到有数以百万计例子、成千上万的特征、成百上千的类别的规模。非常大的数据挖掘问题已经开始在数据库挖掘应用中出现，每天可能会有上千万的事务。我们希望能有在较短的CPU时间内能分析如此大的数据集的机器学习算法。另一个大学习算法出现的领域是从全文数据库和WWW上的信息检索。在信息检索中，文档中的每个字都被当作输入特征，因此每个训练例子可以通过成千上万的特征来描述。最后，语音识别、对象识别和中日文识别也需要判别成千上万的类。

·在大规模数据中学习。三个不同的决策树算法扩展可以用来处理大规模数据。第一个方法是基于在决策树生长时智能地亚取样训练数据。为树选择测试的生长过程包括分析训练数据并选择能够最佳预测输出类的特征。在大的和冗余的数据集中，基于数据的取样来做选择是可能的。第二个方法基于开发聪明的数据结构以避免在内存中保存所有的训练数据。SPRINT方法可以把训练数据打散，让每个例子的属性(特征值)放入一个单独的磁盘文件中。SLIQ用的内存比SPRINT多，但速度稍快。大数据集中的第三种方法是利用决策树的组合。训练数据可以随机划分为N个不相交子集。于是每个子集都可以并行地生成独立的决策树。这些树可以联合投票以预测。尽管单个树精度不高，但组合通常性能较好，同时并行加快了建树的速

度。第四个方法是离散化特征值以解决实值输入特征划分问题，如 RIPPER。

·学习高维数据。流行的C4.5和BP算法并没有很好地处理大规模特征的能力。事实上，从统计角度看，有很多不相关、多噪特征的例子提供了很少的信息，因此需要选择特征。有3个主要方法：第一个方法是对训练数据执行初始分析并选择特征的一个子集提供给学习算法。第二个方法是在学习算法上试不同的特征子集，用这些特征估计算法性能，保留性能最好的子集。第三个方法是把对特征的选择和加权都直接集成到学习算法中，如 WINNOW 等。

虽然现存的一些方法可以在数百万计的训练例上以合理的计算时间进行学习，但我们还不知道在上百亿训练例中的情况。收集尽可能多的在极大数据集上的学习经验是很重要的。

2.3 数据挖掘的重要领域：文本挖掘、多媒体信息挖掘、网络挖掘

·文本挖掘：Internet 已经发展为当今世界上最大的信息库，但它是一个开放性的全球分布网络，资源分布很散，信息搜索困难。WWW 网上的信息，是以网页，也就是(超)文本的形式存放的，如何快速准确地从浩瀚的信息资源中寻找所需信息，这就涉及到我们要研究的文本知识挖掘。文本可被看做一种顺序数据。有许多适合用于顺序数据的 KDD 方法。WWW 网的发展，极大地促进了用于文本挖掘的发展。网络文本挖掘的主要任务包括：Web 文本的特征表示、文档总结、文本分类和文本聚类。

·多媒体信息挖掘：在现存的大量语言、图形图像、Video 等多媒体数据中找出合适的描述模式，理解其意义。主要用于提高多媒体信息的识别和检索效率。目前的应用主要包括语音理解和识别、图像识别和检索等。

·网络挖掘：从 WWW 中抽取、发现有用信息和知识的技术。它包括三个方面的内容：网络内容挖掘、网络结构挖掘、用户行为挖掘。在系统具有大量文档的情况下，查找用户需要的信息成为一件困难而麻烦的事情。智能导航是基于对用户行为的分析，引导用户到可能包含他所需要信息的链接，从而大大简化信息查找的过程。通过对大量用户共同特性的分析，还能够发现用户群共同的兴趣，可用来引导新用户对热点问题的访问。

三、数据挖掘中引入组合学习方法的意义及可能性

在数据挖掘和机器学习过程中，分类预测是最基本的任务之一。传统的做法主要是借助于模式识别、数理统计等方法以寻找尽可能精确的预测规则。但在很多情况下，预测器的精度受领域知识和训练数据及其分布的影响很大。特别是对那些我们还未完全了解其本质特征的预测问题。比如说在文本挖掘领域，需要将各特定文档归入相应的主题中，此即所谓文本分类问题。本质上，要解决这类问题，预测器需要理解文本内容，然后才能通过学习训练文本发现规律并组合成预测规则(学习)，或进行归纳归类(预测)。由于技术上未能解决语言理解问题、替代的方法只能是对组成语句的基元——词或短语进行统计分析，更高级的办法是结合语法和语义模板寻找关键成分。对学习样本而言，这些关键成分被用来总结出预测规则；对待分类样本而言，这些关键成分作为输入由预测规则进行预测。上面描述的方法的主要问题是，它几乎可以肯定要忽略文章中的一大部分内容，而这些内容的重要与否预测器是无法确定的——它无法理解这些内容，因而也使预测少用了很多可能很有用甚至很关键的信息。从根本上限制了预测的正确度。所以，在预测器本身预测质量难于提高的情况下，寻找一般的提高已有预测器学习精度的方法是有相当价值的。组合学习的思想正是为实现这个目的提出来的，它通过综合(多为投票)多个学习器的预测结果给出最终预测结果。实践证明，这些学习器虽然单个学习性能都可以不太理想，但是综合后的结果一般来说都会比单个学习器好，很多情况下学习性能甚至会有质的飞跃。此外，众多实验也表明，组合比单个学习器更适合处理大规模数据集，这就表明了这种思想在数据挖掘中的潜力。

有不相关错误的单个分类器可通过投票使性能得到改善，但有一个更深层次的问题：为什么可以找到有不相关错误的分类器组合？还有另外一个问题：为什么我们不能找到和组合一样好的单个分类器？至少有三个原因可以用来解释为什么好的组合可以被建立及为什么找到和组合性能一样好的单个分类器是困难的或甚至是不可能的。让我们考虑机器学习的本质。机器学习算法通过搜索可能的假设空间 H 以找到最精确的假设(即，能最好地近似未知函数 f 的假设)。假设空间 H 两个最重要的方面是其大小和它是否包含了 f 的好的近似。如果假设空间很大，那么就需要大量的训练数据才能搜索到好的近似，每个训练样本排除 H 中所有错分它的假设，或降低其合理性。在两类问题中，理想的情况是每个训练例子都能消除 H 中假设的一半。这样，我们将仅需要 $O(\log|H|)$ 个例子来从 H 中选择出单个假设。需要组合的第一个原因是训练数据可能并没有提供足够的信息来从 H 中选择一单个最佳分类器。大多数学习算法考虑非常大的假设空间，所以甚至在去掉了错分训练例的假设后还会有很多假设得以保

留。偏好这里的部分假设(如,偏好简单的假设或先验概率高的假设)可能是合理的。然而,通常并没有太多的合理假设。利用保留下来的假设集,可以很容易地建立组合分类器。需要组合的第二个原因是:学习算法可能不能解决我们提出的困难的搜索问题。例如,寻找与一套训练例子相一致的最小决策树问题是NP-hard问题。因此,实际决策树算法用启发式搜索来引导到小决策树的贪婪搜索。相似地,寻找与训练例一致的最小可能的神经网络的权重也是NP-难的。神经网络因此使用局部搜索方法(如梯度下降)来找到网络的局部优化权重。这些有缺陷算法的后果是:甚至当我们的训练例和先验知识的综合(如偏好简单或贝叶斯先验概率)决定了单个最佳假设时,我们也不一定能够找到它,我们找到的可能是有些复杂的假设(或有较低后验概率的假设)。如果在轻微不同的训练集上运行学习算法或加入噪声,我们将得到不同的假设,因此组合将被看作是补偿有缺陷搜索算法的一种方法。需要组合的第三个原因是:假设空间H可能并不包含真实函数f。转而,H可能会包含几个对f一样好的近似。通过这些近似的加权组合,我们可以表示H之外的分类器。理解这一点的一个方法是可视化由学习算法建立起来的决策边界。决策树学习算法建立起来的决策边界是分段线性(更一般地,是超平面分段线性)。通过在不同近似区间上的投票,就可能对多边形的决策边界建立更好的逼近。因之,组合提供了克服假设空间表达能力不足的一种方法。

四、组合分类方法综述

组合分类器是分类器的集合,这些分类器的单独决策被以某种方式组合起来(典型的方法是通过加权或无权重投票)以给新样本分类。在有监督的学习中最活跃的研究领域是构造好的分类器组合。其中的主要发现是当单个分类器间意见不一致时组合分类器通常比组成它们的单个分类器要精确得多。当然,如果单个分类器得出的假设错误超过0.5,则投票的结果其错误会随之增加。因此,组建成功的组合方法是用错误率小于0.5的单个分类器,而这些单个分类器的错误之间至少应在某种程度上无关。

4.1 组合技术分类

组合分类器中包括面向一般应用的和特殊应用的两类。一般性技术根据其特点分为:在训练例上改变分布、操纵输入特征、操纵输出目标、引入随机性等。面向特殊应用的组合包括用特定的方法使用于训练神经网络的BP算法或决策树等产生组合。特殊应用需要对具体情况进行具体分析,此处我们主要介绍一些一般性技术[1]。

·在训练例子上改变分布:Bagging和Boosting。操纵训练集的最直接的方法是Bagging。每一回运行Bagging都给学习算法提供有替代地随机从大小为m的原始训练集抽取出m个训练样本的集合。这种训练集被称作原始训练集合的Bootsrap复制。这种技术也叫做Bootstrap综合,即Bagging[2]。另一个训练集上取样的方法是通过保留不相交子集的方法构造训练集。例如,训练集可被随机地分成10个不相交子集。那么,10个互相重叠的训练集可通过在每一回构造训练集中分别去除不相交子集中的一个而得到。这就是10面交叉验证所用到的构建训练集的方法。于是这样建立起来的分类器组合就称作交叉验证社团[3]。第3个操纵训练集的方法是由AdaBoost描述的。AdaBoost在训练例子上维护一套概率分布,在每一回迭代中AdaBoost在每个例子上调整这种分布,成员分类器在训练例子上的错误率被计算出来并按照它在训练例子上调整概率分布,权重改变的作用是在被误分的例子上放置更多的权重,最终分类器通过单个分类器的加权投票建立起来,每个分类器按照其在训练集上的精度而加权[4]。

·操纵输入特征:在一个识别火山的项目中,Cherkauerd[5]训练了32个神经网络的组合。这32个神经网络的组合是基于119个可获得的输入特征和4个不同的网络大小的8个不同子集,输入特征子集按照不同的图像处理过程(例如基元分析和快速傅立叶变换)被分成不同的组,最后的组合分类器可以在识别火山上达到人类专家的水平。Tumer和Ghosh应用相似的技术到有25个输入特征的声纳数据集中。然而,他们发现甚至仅删除几个输入特征也会极大地损害单个分类器的性能,导致组合分类器性能不理想,很明显,这种技术仅在输入特征高度冗余的情况下才有效。

·操纵输出目标,ECOC(Error-correcting output codes)是这种方法的一个代表,是由Dieterich和Bakiri于1995年提出的一个鲁棒的通过把多类学习问题简化为一系列的两类问题以解决大类问题的方法。假定类别数K很大,新的学习问题可以通过随机地把这K个类划分为两个子集A_i和B_i建立起来。输入数据可以被重新赋予标签,使得在A_i中的每个原始类得到标签"0",B_i中的每个原始类得到标签"1"。这些重新被标记的数据被传递给学习算法,该学习算法据此建立分类器h_i。重复该过程L次(产生不同的子集A_i和B_i),可以得到L个分类器的组合:$h_1,\cdots,h_L$。现在给定一新的数据点,ECOC让每个h_i来分类它。如果$h_i(x)=0$,A_i中的每个类得到一次投票;如果$h_i(x)=1$,B_i中的每个类得到一次投票。在L个分类器都投票后,有最高票数的类被选作综合预测的结果。这种思想

的一个等价方法是：每个类 j 被编码成 L 位的码字 C_j，其第 l 位为1当且仅当 $j\in B_l$。第 l 个学习过的分类器试图预测这些码字的第 l 位。在 L 个分类器用于分类某新点 x 时，它们的预测被组合进一个长为 L 的字码中，ECOC 选择的预测是其码字 C_j 与这 L 位的输出字码距离(汉明距离)最近的类 j。设计好的错误矫正码的方法可被用于选择好的码字 C_j(或等价地，选择子集 A_l 和 B_l)。Dieterich 和 Bakiri 宣称该技术在很多困难的分类问题上可以提高 C4.5和 BP 算法的性能。AdaBoost.OC 是 AdaBoost 和 ECOC 的结合。其性能优于 ECOC 和 Bagging。它的主要好处是实现的简单性：它可以使用任何可解决两类问题的学习算法[5]。

·注入随机性。在训练神经网络的 BP 算法中，神经网络的初始权重是随机设置的。如果使用不同的初始权重，则结果分类器也会大不相同。组合这些不同的分类器可以得到组合分类器，然而，操纵训练集可能会更有效。对于决策树算法 C4.5来说，很容易注入随机性。C4.5的关键决策是在决策树中每个内部节点选择某特征以测试。在每个内部点，C4.5应用象信息增益这样的准则来对不同的可能的特征测试进行排序。它于是选择排在首位的特征值测试。对有 V 个可能取值的离散特征，决策树把数据分裂成 V 个子集，依赖于被选择特征的值。对实值特征，决策树把数据分裂成两个子集，依赖于选择特征的值是大于还是小于某选定阈值。Dietteich&Kong 使 C4.5在排在前面的最佳测试中随机地选择。结果表明随机选择有很好的作用，尤其在字母识别任务中。Raviv&Intrator 组合在训练数据上伴随着输入特征上的噪声的 Bootstrap 取样。为了训练组合神经网络的每一个成员它们从原始训练数据中有替代地抽取训练例，每个训练例的输入特征都通过加入高斯噪声来扰动。这种方法在合成测试任务和药物分析任务中有大的改善。和注入随机性密切相关的方法是 Markov Chain Monte Carlo 方法。它已被用于神经网络和决策树。MCMC 方法的基本思想是建立一个 Markov 过程，它产生假设的一个无限序列。在贝叶斯环境下，目的是以概率 $P(h_l|S)$ 产生假设 h_l，其中 S 是训练样本。$P(h_l|S)$ 以通常的方式作为似然 $P(S_l|h_l)$ 和 h_l 的先验概率的积来计算。为应用 MCMC，定义一套在假设间互相转换的算子。对于神经网络，这样的算子可能会调整网络的权重之一。对决策树，算子可能会在树中交换父子节点，或以一个节点代替另一个节点。MCMC 通过维护当前假设 h_l 运作。每一步，它选择一个算子，应用它(以获得 h_{l+1})，并在训练数据上训练结果分类器的似然。它据此决定是接受 h_{l+1} 还是抛弃它返回 h_l。在不同的技术条件下，可以证明，这种过程最终可以收敛到 h_l 与其后验概率成正比的静态概率分布上。一个标准的方法是运行 Markov 过程很长一段时间并收集一套 Markov 过程中得到的 L 个分类器。这些分类器于是按照其后验概率通过加权投票组合起来。

4.2 待解决的问题

组合是一种成熟的通过综合不那么精确的分类器获得高精度分类器的方法。然而，关于如何用最好的方法建立组合以及如何最好地理解组合的决策仍然有很多问题。面对一个新学习算法，建立和应用分类器组合的最好方法是什么?原则上，不会只有一个最好的组合方法，就如不会有单个最好的学习算法一样。然而，有些方法可能一致地比另一些方法好。某些情况下一些方法也会比另一些方法好。实验研究表明 AdaBoost 是构造决策树组合的最好方法之一。Schapire 比较了 AdaBoost.M2和 AdaBoost.OC 与 Bagging 和 ECOC，结论是 AdaBoost 一般性能更好，Quinlan 发现，在噪声很大的训练数据上，AdaBoost.M1的性能会很差。Dietterich&Kong 发现组合 Bagging 和 ECOC 改变了两个方法的性能，这就表明结合其他的组合方法也应得到研究。他们也发现 ECOC 在高度局部化的算法上性能并不好。在构造神经网络、规则学习系统及其它学习系统的组合方面几乎没有系统的研究，这个领域还有待探索。虽然组合提供了很精确的分类器，也有一些问题可能会限制它们的实际应用。一个问题是组合可能会需要大量内存和计算。例如，200个决策树的组合(它在识别字母中性能最佳)需要59M 的空间。所以，重要的研究方向是找到转化这些组合到不那么冗余的表达方式，途径可能是删除组合中高度相关的成员，或通过表示变换。另一个问题是组合没有提供对其决策的直观解释。一个决策树通常可以被人类用户理解，但200个投票的决策树组合就很难令人理解了。是否可以找到从组合中获得至少是局部的解释的方法?

五、组合学习方法在文本、多媒体和网络信息挖掘中的应用研究

有些组合分类方法已经确定可以用于上述的部分领域，但还存在缺陷，如效率等，而且组合方法和特定领域的专门知识间的结合还不够深入，还远未达到实用的地步，尚有待进一步探索。

5.1 文本分类和检索

大多数文本分类研究集中于二值问题，其中文档被分类成与某预定义的主题相关或不相关。然而，有一些象 Internet 新闻、电子邮件和数字图书馆这样的文本性数据是由不同主题组成的，也因此提出了多类分类问题。进一步，很多文档还可能兼类，即学习问题中的多标签情况。这方面的研究还不足。分类多标签的文

本分类的一般的解决方案是把任务打散成不相交的二值分类问题，每个类对应这这样的一个二值问题。为分类新的文档，所有两类分类器都需预测并组合成单个决策。结果是文档可能属于的类的列表，或是对可能主题的排序。这种方法的缺点是它忽略了这些类别间的任何相关性。已有实验表明，Boosting 组合学习方法中 AdaBoost.MH 和 AdaBoost.MR 两算法的扩展可以用于高效地处理标签集合[7]。这种方法可以预测所有而且仅仅是所有的正确标签，因此学习算法可以按照其预测与给定文档相联系的标签集的好的近似的能力进行评估。它还可以被设计成对标签进行排序，其中正确标签具有最高的级别。这种算法的时间和空间复杂度可以为 $O(mk)$，其中 m 是训练文档数，k 是类别数，实验结果表明，这个系统的一般性能比其它通用算法好，但速度慢。有待继续进行的工作就是应用理论和实验研究的成果，促进组合学习方法用于超大规模的文档分类和学习中，并研究并行文档分类算法和其它提高学习效率的方法。另外，弱学习器设计方面可以考虑换用不那么弱的分类规则，也可考虑与已有分类方法结合起来，即：与传统方法（如 RIPPER、Rocchio、Sleeping-Experts、朴素贝叶斯和概率 TF-IDF 等）结合或与新兴文本预测方法（应用小波、分形等特征空间转换方法进行分类）及其组合相结合。

5.2 图像识别和检索

现有识别算法准确度都不理想，且易受数据扰动干扰。希望组合能提高性能。

5.3 语音识别和理解

语音识别问题可以用与本文分类问题类似的方法加以解决。

5.4网络导航：用户行为分析和偏好排序

执行电影推荐任务的系统首先会询问用户对所看电影的偏好排序。系统然后检查已知的其他用户对电影的偏好排序信息并组合这些消息来向用户提供他可能感兴趣的电影列表。为做到这一点，系统需要找寻与该用户偏好相似的人群的选择。这类问题的特点是组合表示的是相对偏好而不是绝对选择。也就是说，即使排序中的每个级别会有一个数值，但我们通常可以忽略掉它，仅仅注意其相对位置，当不同用户用完全不同的打分范围来进行排序时就更是如此。在宏搜索和信息检索中也有类似的问题。宏搜索的目的是组合对 WWW 搜索策略的排序。每个搜索策略是象输入查询、执行一些查询的简单变换并发送之到特定的搜索引擎上。每个策略的输出是查询答案的 URL 列表，其目的是组合对给定查询集合性能最好的策略。对电影推荐问题，存在很多不同人对电影排序的大的数据集合。近邻算法和回归算法曾用于此问题。在机器学习中尽管排序应用很广泛，本问题的研究还很不够。组合排序的方法基本上是基于近邻方法或数值优化技术（排序被看作实值打分，组合不同排序的问题被简化为找到一套最小化组合打分和用户反馈间差距的数值搜索问题）。这些方法实践中性能可能不错，但它们并不能保证当我们把分数看作表达偏好的方法时组合的系统将和用户的偏好完全一致。Cohen，Schapire 和 Singer 提出了一个操纵和组合排序的框架以建立偏好图，其问题被简化为 NP-Complete 的组合优化问题。因此，组合不同的排序是一种近似的方法。这可以被描述成在线算法。AdaBoost 组合学习方法的扩展可用于解决此类问题[8]。它基于与上面类似的框架，对有充足的时间找到最佳组合的情况尤为适合。所以，这两种方法可以互补，它们的结合，可以为解决组合多排序问题提供一个可行的方法。

结束语 组合学习方法的分类预测精度较高，在数据挖掘中的应用前景非常广泛。它的主要问题是其速度较单个学习方法慢。因为它需要生成多个分类器以进行投票。如何在学习质量和效率两者间找到满意的折衷点是值得着重研究的问题。

参考文献

1 Dietterich T. Machine learning research: four current directions. Oregon State University, 1999

2 Breiman L. Bagging predictors. Machine learning, 1996, 26(2): 123～140

3 Parmanto B, Munro O W, Doyle H R. Improving committee diagnosis with resampling techniques. In: Touretzky, D. S., Mozer, M. C., & Hesselmo, M. E. Eds. Advances in Neural Information Processing Systems, Cambridge, MA: MIT Press, 1996, 8: 882～888

4 Freund Y, Schapire R E. A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Science, 1997, 55(1): 119～139

5 Cherkauer K J. Human expert-level performance on a scientific image analysis task by a system using combined artificial neural networks. In: Chan, P. Ed. Working Notes of the AAAI Workshop on Integrating Multiple Learned Models, 1996. 15～21

6 Schapire R E, Singer Y. Using output codes to boost multiclass learning problems. In: Machine Learning: Proc. of the Fourteenth Int Conf. 1997. 313～321

7 Schapire R E, Singer Y. BoostTexer. A system for multiclass multi-label text categorization. Machine Learning, 1998

8 Freund Y, Iyer R, Schapire R E, Singer Y. An efficient boosting algorithm for combining preferences. In: Machine Learning: Proc. of the Fifteenth Int Conf. 1998