博 士 学 位 论 文

特征耦合泛化及其在文本挖掘中的应用

Feature Coupling Generalization and Its Application in Text Mining

作 者 姓 名：李彦鹏

学科、 专业：计算机应用技术

学　　号：10709039

指 导 教 师：林鸿飞

完 成 日 期：2011-5-30

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：特征耦合泛化及其在文本挖掘中的应用

作 者 签 名 ：_____ 日期：_____年____月____日

摘　　要

文本挖掘（Text Mining）技术是利用计算机程序自动读取和理解自然语言文本，并从中发现有价值的信息，从而提高人们的工作效率。随着信息技术的飞速发展和互联网时代的来临，该技术拥有了很大的实际应用价值和广阔的应用前景。在处理文本挖掘问题的方法中，基于机器学习（Machine Learning）的方法得到了广泛的应用，在很多实验中取得了较好的效果。特征表示（Feature Representation）是机器学习方法中至关重要的一步，很大程度决定了系统效果的高低，然而在传统的基于局部特征的监督学习（Supervised Learning）策略中，由于已标注训练集中的样本数量有限，存在着数据稀疏问题，即产生了很多低频特征，由于缺少信息量这些特征在机器学习过程中往往得不到好的利用，这种影响在文本挖掘和自然语言处理任务中更为严重。针对此问题，本文研究如何利用未标注数据将这些被忽略的特征转化成更富有信息量的新特征，从而可以激发出这些特征潜在的作用，达到提高系统的性能的目的。本文提出了一种新的特征构建方法—特征耦合泛化（Feature Coupling Generalization, FCG），该方法利用原始特征在海量未标注数据中的共现信息以及特征间的概念层次关系生成新的特征。相比于原始特征，新特征具有更丰富的信息量和更泛化的表示。本文讨论了该方法中各种因素对系统性能的影响，并通过实验验证该方法在文本挖掘任务中的效果。

本文将 FCG 方法应用于三个经典的文本挖掘任务：命名实体识别（Named Entity Recognition）、关系抽取（Relation Extraction）、文本分类（Text Classification），对每个任务进行了详细的研究，从不同角度比较了经典特征与 FCG 方法的效果、检验了 FCG 方法所带来的贡献，并通过观测低频特征在不同方法中的效果分析了 FCG 方法有效的原因以及对数据稀疏问题的解决情况。实验结果显示，FCG 可以将传统方法中被忽略的低频特征转化为有效的特征，在传统方法的基础上有显著的提高，而且 FCG 方法可以很容易的应用于海量的未标注数据，这是相比于其他半监督学习方法（Semi-supervised Learning）的优势。更有趣的现象是，仅仅使用 FCG 方法所生成的新特征的效果普遍好于经典的特征，这说明该方法有可能在普遍的机器学习问题中取代经典特征表示方法，为特征生成的研究开辟了新的思路。此外，在公开评测数据上与其他研究者的对比结果显示，基于 FCG 方法的系统取得了很好的效果。

关键词：文本挖掘；机器学习；特征；命名实体识别；关系抽取；文本分类

Feature Coupling Generalization and Its Applications in Text Mining

Abstract

Text mining aims to automatically extract knowledge form plain text in natural language, which can help people to find useful information from large text corpora accurately and efficiently. With the rapid development of information science and World Wide Web, text mining becomes more and more useful in practice. In this area, the techniques based on supervised machine learning have been used with great success and achieve good results in a lot of experiments. Feature representation is one of the most important issues in machine learning, which has big impact on the performance of learning systems. However, in traditional supervised learning method for text mining, the limited amount of training data can lead to serious data sparseness problem in feature space, where a lot of low-frequency features cannot be utilized well due to insufficient information available. Addressing this problem, I develop a method that aims to convert these ignored features to effective ones so as to improve the performance of classification. I propose Feature Coupling Generalization (FCG) framework for creating new features from raw features based on feature co-occurrences in a large amount of unlabeled data and the concept hierarchy of raw features. The new features lead to a more informative and general representation than the raw features. In this thesis, I discuss various factors that influence the performance of FCG and examine its performance in text mining tasks.

In this work, FCG is applied to three text mining tasks: named entity recognition, relation extraction and text classification. In each task, I investigate the performance of classical features the features derived from FCG, examine the contribution of FCG and whether it overcomes the problem of data sparseness. The experimental results show that FCG can utilize well the features ignored by supervised learning and improve the performance of classical methods. In all tasks, FCG can utilize huge amount of unlabeled to generate new features, which is one of the advantages over other semi-supervised learning methods. Interestingly, I find that the individual performance of new features generated by FCG is at least as well as the classical features widely used in these tasks, which indicates FCG provides an alternative way for feature representation in machine learning. The results also show that the system based on FCG achieves state-of-the-art performance on public challenge datasets.

Key Words：Text Mining; Machine Learning; Feature; Named Entity Recognition; Relation Extraction; Text Classification

目　　录

1 绪论

1.1 文本挖掘

1.1.1 概述

文本挖掘（Text Mining）[1-4]是指利用计算机程序自动在文本中发现有价值的信息，是数据挖掘（Data Mining）[5] 的一个热门方向。文本作为数据的最重要的存储方式之一，其中蕴含着丰富的知识，如何通过计算机程序快速而准确的获取其中有价值的信息具有着重要的意义。而且文本作为人类语言的主要载体之一，是人类智能的重要表现，在文本挖掘的过程中必然涉及到对人类语言本质的探讨，也必将对人工智能的研究起着积极的推动作用。著名的图灵测试 [6] 就是将计算机对语言的理解能力作为判断计算机是否智能的标准。因此，文本挖掘的研究有着重要的理论和实践意义。

近年来随着互联网技术的飞速发展，存储在计算机中文本的数量也日益增多，例如网页的数量目前已超过了 3 百亿（参见 http://www.worldwidewebsize.com/），生物文献的数据库 PubMed（http://www.ncbi.nlm.nih.gov/pubmed）中的文献已经超过了 2 千万，如此海量的文本一方面为人们的学习提供了丰富的资源，另一方面却给人们的阅读带来了很大的困难。在这种背景下，文本挖掘技术有了很好的用武之地，有效地文本挖掘工具可以帮助人们快速准确的获取信息，并且可以通过已有的知识发现新的知识，起到创新的作用。近几年各种文本挖掘技术的研究取得了一定的进展。本文中讨论的广义的文本挖掘技术包含了自然语言处理技术（Natural Language Processing， NLP）、计算语言学（Computational Linguistics）和文本检索技术（Text Retrieval），因为这些技术本质上都是设计关于自然语言理解的算法从而从文本中获取有价值的信息，为了描述简洁，本文将这些概念都归属于广义的文本挖掘技术。

根据不同的应用背景和所关注文本的特点，文本挖掘的应用可以划分为若干个具体的任务[3-4]。有的任务所关注的是对词或词组的理解，如词性标注（Part-of-Speech Tagging）、命名实体识别（Named Entity Recognition）、词的语义类型标注（Semantic Role Labeling）、词的情感倾向性分类（Sentiment Classification）等。这些任务要求算法能够自动的识别出词或词组的词性、语义角色、专有名词的类别等人工指定的标签。很多情况下这些任务可能通过词或词组本身的信息就可以判断出相应的类别， 有时也需要考虑上下文的影响。比如同一个词会在不同的上下文中具有不同的词性词义等，必须综合该词自身的信息和上下文信息，才能做出正确的判断。有的任务侧重于对句子的理解，如句法分析（Syntactic Analysis）、关系抽取（Relation Extraction）、事件抽取（Event

Extraction）、词义消歧（Word Sense Disambiguation）、指代消解（Co-Reference Resolution）。这些任务需要处理的不仅仅是词或词组，而更多的涉及到分析当前句子中的语境和语义，因此属于更高级的自然语言理解问题。还有些任务侧重于对整篇文章的理解，如文本分类（Text Classification），文本检索（Text Retrieval），这些任务需要系统能够判断一篇文档的类别，找出用户所感兴趣的文档，将不相关的过滤掉。以上三类任务属于文本挖掘中的基本任务，从本质上都是分类或聚类的问题，很多情况下单独的任务并无法满足用户多种多样的需求，而是通常作为处理更高级的自然语言理解问题的基础。还有一类更为复杂的任务，如问答系统（Question-Answer System）、机器翻译（Machine Translation）、知识发现（Knowledge Discovery）、聊天机器人（Chat Robot）等，要求机器具有更高的智能，能够理解复杂的语境，几乎使机器具有了和人类交流的能力。然而这类任务相对于基本的分类和聚类问题具有更大的难度，目前的研究水平还无法满足实际应用中的需求，性能最好的系统和人类对语言的理解能力也有很大的差距[3-4]，不作为本文重点讨论的内容。

本文选择了三个有代表性的基础任务进行重点研究，分别是命名实体识别、关系抽取和文本分类。这三个任务在文本挖掘中起着重要的作用，是很多复杂的自然语言处理系统的基础，也是近年来研究的热门方向，目前既取得了一定的进展，也存在着很多问题限制了其性能的进一步提升。本文的研究沿着从特殊到一般再到特殊的思路，先从具体的任务入手，通过分析和实验找出该技术发展中所遇到的瓶颈问题，而后针对该问题提出一个解决方案，并将其抽象泛化到一个通用的方法，然后应用于解决其它问题。下面分别对这三个任务进行概述。

1.1.2 命名实体识别

命名实体识别[3-4]的目标是识别出文本中的专有名词，例如识别文章中出现的人名、地名、机构名等，再如在生物医学文献中识别出基因名、蛋白质、细胞、药物的名称等，由于生物医学文献的规模庞大，各种专有名词的数量也达到了数万之多，人工识别费时费力，因此命名实体识别的工作就变得很有意义。命名实体识别是文本掘系统中的一个重要的基础步骤，其精度的高低直接影响到其它步骤的所取得的效果。一个关键的问题是如何评价一个识别系统性能的高低，人们通常采取的方法是将系统识别的结果与人工标注的结果进行对比，利用规定的指标如准确率（Precision）、召回率（Recall）、F 值（F-score）进行量化。而且为了使不同系统的性能得到公平的比较，近年来人们开发出了很多公共的数据集和评测任务，每个系统都使用同样的数据集进行测试，按照同样的指标进行评估，这样不同系统性能之间的差异可以比较真实的反应不同方法的相对效

果，从而促进该项技术的发展。比较著名的命名实体识别评测有 MUC-7[7]、CoNLL-2003[8]，是关于新闻类文本中人名、地名、机构名等命名实体的识别。JNLPBA[9]、BioCreative I[10]、BioCreative II[11]评测是针对生物医学领域的，要求识别出文章中出现基因名、蛋白质名、细胞名等。这些评测都使用 F 值作为主要的评测指标，其中最好的结果在 73 至 87 之间（注：本文为了描述问题简洁，将评测指标中的百分号省略），这些差异主要取决于不同的命名实体类型。对于某些相对简单的命名实体如人名、地名识别精度已经到达了 90 以上，但是对于很大一部分复杂的命名实体如机构名、基因名等，识别的精度与人工的标注结果有较大的差距[7-11]，因此具有较大的提升空间和进一步研究的价值。

1.1.3 关系抽取

这里所讨论的关系抽取任务[3-4]是指在一段描述性文本中确定两个或多个指定实体之间的是否存在语义上的关系，例如人与人之间的关系、人与组织机构的关系、生物医学文献中蛋白质和蛋白质之间的关系等。相对于命名实体识别而言，关系抽取可以看作是文本挖掘更高级的应用，因为其中不但需要到命名实体识别技术作为铺垫还涉及到更多对上下文语义的理解。关系抽取系统有很大的应用前景，例如可以用来提取出搜索引擎返回结果中和查询串在语义上最相关的部分，从而提高信息检索的精度，也可以将抽取出的关系表示成图的形式，通过聚类分析的方式挖掘出新的有价值的信息，或通过已有的关系发现隐含的未知的关系。和命名实体识别任务类似，近年来人们开发出了很多关系抽取任务的公共评测数据集。比较著名的有 ACE 评测[12]，主要是针对新闻类文本中特定几类关系的抽取。生物医学领域的关系抽取也有很多，比如蛋白质-蛋白质关系抽取（Protein-Protein Interaction Extraction，PPIE）[4]是最热门的关系抽取任务之一，相关的公共数据集和评测有 AIMED[13-14]、LLL[15]、BioCreative II[16]、BioCreative II.5[17]等。然而由于这些数据集的评测方式和数据类型以及规模存在着较大的不同，每个评测中最优的结果之间也有较大的差异，不具有直接的可比性。例如 ACE、AIMED 和 LLL 数据集的相关评测中仅仅考虑了关系抽取的性能而不考虑命名实体识别效果的影响，而且任务限定在抽取在同一个句子中实体间的关系，不考虑同一篇文章多个句子之间体现的实体语义关系，这些评测结果的最高的 F 值在 60 至 80 之间浮动，取决于训练样本的数量和关系的类别。在 BioCreative 相关的 PPIE 评测任务中，评测的时候考虑了命名实体识别、命命名实体标准化，关系抽取的整个过程，其最终的结果是由所有步骤的效果共同决定的，而且该任务没有把关系抽取限定在一个句子内，因此该任务具有了更大的难度，最高的评测指标只有 30%左右的 F 值，远远低于其它的数据集。综上所述，在大

多数情况下，关系抽取的精度还无法达到实际应用的需要，与人工标注的结果也有较大的差异，其主要原因之一是在自然语言中描述两个实体关系表达方式的千差万别导致了很难在算法中枚举出所有可以抽取出关系的规则。

1.1.4 文本分类

文本分类[3-4]是文本挖掘中另一个经典的问题，该任务需要将一篇文章分类到指定的类别，例如判断一篇新闻类文章的体裁属于政治类、军事类、体育类还是娱乐类。与命名实体识别和关系抽取的最大不同是待分类的样本是整篇文章，而不是一个短语或一个句子。相比于命名实体识别和关系抽取而言，文本分类属于粗粒度的信息抽取和分类，尽管很多时候无法精确的提取出所需信息，但是可以在海量的文本中快速的获取可能对人们有价值的信息，从而很大程度的缩小搜索的范围，提高人们的工作效率。从广义上讲，文本搜索引擎也属于文本分类，可以当作将文本分类成用户可能感兴趣和不感兴趣两类，将用户可能感兴趣的文本推荐给用户，如果系统的分类和实际用户的需求越吻合那么系统的性能越高。在文本分类相关的公共数据和评测也比较多，如路透社语料[18]，20 新闻组语料[19]、TREC Genomics Track 2005 评测[20]、BioCreative II 分类任务评测[21]。这些任务中有些具有明确的实际应用背景，例如 Genomics Track 中的分类任务[20]是实现机器自动的将生物文本分类成是否与鼠类基因相关，相关的文本将被用于构建基因数据库。BioCreative 分类任务[21]是判断一篇生物医学文献中是否反映了蛋白质之间的交互关系。这些工作通常是由生物学家人工完成的，自动的文本分类程序可以很大的提高人们的工作效率。各个评测的最高纪录也不尽相同，20 新闻组和路透社语料的分类精度较高，达到了 90 以上。而专业文献的分类的精度较低，在 TREC 和 BioCreative 评测的任务中，很多任务的评测指标低于 80。因此，和命名实体识别与关系抽取任务类似，尽管在近年的研究中文本分类的技术有了较大的进展，但也存在着很多瓶颈问题未得到解决，导致了性能的提升受到了限制，但也说明了该领域有很大的提升空间和进一步研究的价值。

1.1.5 方法总结

一个有趣的现象是，在这些不同任务的评测中，取得精度较高的系统都使用了很相似的方法，即基于监督学习（Supervised Learning）的方法。这种策略的通常的步骤是：首先根据任务选择特征，将每个样本表示成特征向量的形式，然后使用分类器进行分类。在 CoNLL、JNLPBA，BioCreative 以及 BioCreative II 命名实体的评测中，取得精度较高的系统[8-11]都使用了词特征、n-gram 以及正则表达式特征（包括词缀、词形等）、词

性标注特征，并使用了隐马尔科夫模型（Hidden Markov Model，HMM）[22]、最大熵马尔科夫模型（Maximum Entropy Markov Model, MEMM）[23]、条件随机域（Conditional Random Field，CRF）[24]、支持向量机（Support Vector Machine，SVM）[25] 等线性分类器将这些特征进行统一。在关系抽取的各个评测中，应用最广泛的仍然是基于监督学习的方法，评测中效果较好的系统大都使用了各种基于词或词组、词性、句法等特征。此外，近年来很多系统采用了基于核函数方法（Kernel Method）[25-26]来表示样本，例如：Zhang 等人[27]使用了基于依存分析的核函数以及 SVM 分类器，在 ACE 评测数据集上举得了较高的精度。Miwa 等人[28]融合了词汇级别、浅层句法分、依存分析的特征表示，取得了 AIMED 数据集中较高的 F 值，该数据集为生物医学关系抽取中应用最广的公开数据之一。在文本分类中，机器学习方法的应用更为广泛，目前几乎所有评测中的系统都采用了基于机器学习的方法，其中基于词特征的文本表示和线性分类器的组合，如 SVM，Logistic 回归等，被使用的最为广泛，成为文本分类领域的经典方法。机器学习方法在各种文本挖掘任务中显示了较好的效果，相对于基于单纯的基于规则的方法有较大的优势。那么目前的机器学习的方法是否有进一步提升的空间呢？为了得到答案，人们也通过各种理论分析和实验进行着不断的探索，这也是本文研究的重点，在后续章节中将逐步对该问题进行详细的讨论。

1.2 机器学习

学习能力是人类智能的重要组成部分，人类从出生以来就具有了非凡的学习能力，能够不断地从过去的经验或其他人中学习得到新的知识，人类社会也因此而得到了不断的进步和发展。自从 20 世纪电子计算机问世以来，让机器能够像人类一样思考和学习也成为了人们所追求的梦想，一方面可以使计算机成为人类的助手，提高人类的工作效率，另一方面对探索人类思维的本质的研究有着重要的作用。1960 年 Rosenblatt 设计了第一个感知器（Perceptron）模型[29]，并指出该模型具有一定的从经验数据中学习的能力，从此越来越多的人致力于研究基于经验数据统计的机器学习方法，人们将多层感知器的方法称为神经网络（Neural Network）。1980 年 LeCun[30]提出了可以同时为神经网络中的多个感知器寻找参数的方法，即后向传播算法，从此基于神经网络的机器学习技术成为了研究的热门。然而神经网络在理论上存在的两个问题是：由于局部极小值的存在，优化的过程中只能找到近似的最优解；没有自动的控制过拟合（Overfitting）[25]的方法，从而导致在理论和实践的研究中受到了一定的限制。在 1995 年 Vapnik 等人[25]提出支持向量机技术，该方法通过基于最优超平面的正则化（Regularization）的方法在

一定程度上控制了过拟合现象，并且 SVM 中的训练函数是凸函数，可以求得最优解，在很多方面优于神经网络，在手写数字识别、文本分类等领域也取得了较好的效果[25]。该技术也导致了人们越来与关注基于“损失函数+正则项”的机器学习方法，将 SVM 的思想扩展到其它方法中，例如基于正则化的 Logistic 回归模型[31]、贝叶斯 Logistic 回归模型[32]、正则化的 Huber 线性分类器[33][34]等都在实际应用中取得和 SVM 相当甚至更好的效果。近年来随着各种信息的不断增多，很多从经典统计学直接派生出来的方法也得到了广泛的应用，如朴素贝叶斯（Naïve Bayes）模型、高斯混合模型（Gaussian Mixture Model）、Logistic 回归、隐马尔科夫模型都可以用来解决分类的问题。上述所有方法从本质上看仍然属于经典统计学方法，正则化技术也是数学中常用的解决不适定（ill-posed）[25]问题的技术之一。

同时人们也尝试着使用统计学的理论对这些方法进行分析，建立机器学习领域所特有的理论。1962 年 Novikoff[35]分析了感知器模型的可学习性，证明了经过足够多次的学习以后，神经网络可以将训练数据分开，对机器学习理论的研究起着重要的作用。自 1968 年到 1989 年期间，Vapnik 和 Chervonenkis[25]分析了当经验样本数量有限的时候，影响学习泛化能力的因素，对传统的大数定律进行了推广，提出了 VC 维（Vapnik-Chervonenkis Dimension）理论，并提出了一个可以通过训练集来估计分类器泛化性能的风险上届。1984 年 Valiant 等人提出了 PAC（Probably Approximately Correct）理论[36]，仍然是将统计学中的理论引入到机器学习领域，提出了一系列影响分类器学习能力的因素以及在具有噪音的情况下选择分类的基本原则，也得到了和 VC 理论很类似的结果。2002 年 Partlett 等人[37]提出了 Rademacher 复杂度的概念，得到了一个控制分类器泛化能的上届，而 Rademacher 复杂度的本质是对 VC 维的经验估计[26]，使得该上届更易于计算。2002 年等人[38]将 PAC 理论与 Bayes 方法结合，提出了 PAC-Bayes 上届，得到了风险上届的另一种表达形式。上述理论从本质上是对统计学中大数定律（Law of Large Number）、Chernoff 上届、Hoeffding 不等式等经典理论的推广，描述了在样本充足或不足的情况下影响概率收敛性的因素。但是这些理论中最终的结果却无法作为分类器性能精确的度量，得到上届很松，即理论分析的结果和实际有较大的偏差，在相关文献中都有讨论[39-40]。以上所介绍的方法都属于监督学习（Supervised Learning），是机器学习研究的最核心部分之一，而且有一套相应的理论体系，在实践中也应用的最为广泛。当然机器学习中还包括很多其它的研究领域，如特征选择（Feature Selection）、半监督学习（Semi-supervised Learning）、聚类（Clustering）、强化学习（Reinforcement

Learning）等。由于其涉及面非常广泛，本文只选择与本文核心内容密切相关的部分进行调研，关于特征选择和半监督学习的调研将在后续相关章节中给予详细介绍。

监督学习方法的通常步骤是：首先通过人工方式将样本表示成特征向量的形式，然后通过机器学习算法在训练集中统计得到每个向量的权重，而后根据这些权重将特征进行组合，从而达到预测样本类别的目的。然而需要指出的是，这些方法距离让机器自动学习的目标还有很大的差距，远远没有达到“机器学习”字面上的含义。因为这些方法都是建立在已经给定特征表示的基础上，而这些特征还需要人工进行赋予，特征的质量很大程度上决定了机器学习算法最终的性能。因此仅仅关注分类器的研究是有很大局限性的。在机器学习应用领域中很多实验表明，特征的设计比分类器的选择更大程度的影响了机器学习方法最终的效果。例如：在文本挖掘相关的评测中，名列前茅的系统都使用了很相似的特征而分类器却不完全一样，而没有使用这些特征的系统尽管使用了同样的分类器，效果却相差很大[7-21]。例如在命名实体识别中，词特征、词缀特征、词形特征等起到了很关键的作用，在关系抽取中，词特征、句法分析特征被大多数系统所使用，在文本分类中不考虑位置和顺序的词特征，即著名的 Bag-of-Words（BOW）特征几乎成了每个系统中必备的方法。由此可见，如何构建有效的特征是机器学习系统设计中至关重要的一步，因为它是任何机器学习算法的源头，比分类器的概念更接近机器学习的本质。试想一个简单的情况，如果把分类任务中最终得到的类别判定规则看作特征，那么机器学习的过程就可以看做是生成该特征的过程，那么就得到了监督学习的另一种定义，从原始的特征转化到最终类别判定特征的过程，预期的目标就是最终生成的特征和人工标注的该特征的吻合度最大。很有趣的是这种关于机器学习的定义与 Hinsky 和 Michalski 所定义的学习的本质有很大的相似，Minsky 认为“Learning is making useful changes in our minds”[41]，Michalski 认为"Learning is constructing or modifying representations of what is being experienced."[42]。尽管这两种定义比较笼统，但比大多数基于统计学的定义更接近与学习问题的本质，因为其中包括了特征生成的部分，即事物的表示从一种状态变为另一种状态。

1.3 特征表示

但是遗憾的是在机器学习领域人们对于特征的研究还仅仅处于起步阶段[43]，很多关键性的问题没有解决甚至没有定义，比如什么是特征？什么是好的特征？都没有严格的定义。目前的特征构造方法大都是由人工来完成的，本文将特征的生成分为三个步骤：1）人们“创造”出某种原始特征；2）通过自动的算法对原始特征进行重组，产生新的

特征；3）过滤掉不相关的特征。步骤 1 涉及到人类学习和机器学习最本质的问题，即人类是如何表示所观察到得事物的，该领域的进展目前几乎处于空白阶段，很多关键性的问题没有解决，还不存在可以自动的将所观测的事物（即 Michalski 所描述的“what is being experienced”）转化成计算机可以识别的特征的通用算法。目前人们的研究侧重在步骤 2 和步骤 3 中，这两个步骤算法的成功会减轻步骤 1 中人工工作的负担，同时可以发现步骤 1 里没有发现的更好的特征，从而提高机器学习的效果。

特征选择（步骤 3）[44-45]是目前机器学习研究中最热门的领域之一，近年来提出了很多特征选择算法，如信息增益（Information Gain）[45]，Chi 方检验（Chi-Square）[45]，Relief 方法[46]。从广义上说，很多分类器可以看作特征选择的方法，比如支持向量机，Logistic 回归，神经网络，这些都是根据对训练样本的统计，为每个特征找出适当的权重，而分类器的效果取决于那些权重高的特征，而忽略掉那些权重低的特征。同样很多人研究如何用分类器做特征选择，提出了基于 Wrapper 的特征选择策略[47]。而这些方法的作用更多的在于去除冗余的信息，而不会增加新的信息，在很多任务中人们发现如果选择了合适的分类器，特征选择对提高结果的作用不明显，甚至会因为去掉了相关的特征而降低效果，其作用更多的在于减少计算量、提高效率。

特征抽取（Feature Extraction）[48]的概念最早源自在图像识别领域，其目的是将原始的像素特征进行组合，转化为更有区分度的新特征，从而能结合后续的分类模型达到更好的识别效果，该策略可以归属于步骤 2，经典的算法包括主成分分析（Principle Component Analysis，PCA）[49]，独立成分分析（Independent Component Analysis，ICA）[50]，ISOMAP[51]，LLE[52]，Sparse Coding[53]等，在文本处理领域应用较多的是潜在语义分析（Latent Semantic Analysis，LSA）。这些方法的共同特点是，通过原始特征的线性或非线性组合，产生新的特征，从直觉上组合后的特征可以利用原始特征之间的关联，在某些方面体现了优势，在图像处理领域的一些实验中也取得了较好的效果[53-54]。例如等人[54]发现使用 PCA 的人脸识别系统比基于像素的方法取得了更好的识别效果，等人[55]发现使用了 Sparse Coding 的方法，在若干图像识别和文本分类的数据集上，效果至少和基于像素或词特征的方法相当。然而，这些方法的最大问题是，无法从理论上说明这样的变换为什么会带来预测效果的提高，每一种特征的变换都是基于某一种假设，但是无法证实满足这些假设会导致分类效果的提升。并且人们发现在很多情况下，这些方法的引入并不能带来明显的提高，和特征选择类似，仅仅起到降低特征维度的作用，尤其是在自然语言处理和文本挖掘的任务中并不是主流的方法。另一个问题是此类方法大多依靠矩阵变换来实现，可以处理数据的规模有限，很难应用于高维度的数据，而自然语