華中科技大學

本科生毕业设计（论文）参考文献译文本

译文出处：

Published in: IEEE Transactions on Systems, Man, and Cybernetics

Page(s): 397-415 **INSPEC Accession Number:** 9921439

Date of Publication: 18 April 2008 **DOI:** 10.1109/TSMCC.2008.919172

Authors: Jin Y, Sendhoff B

院　　系 自动化学院

专业班级 自动化 1401 班

姓　　名 王 壮

学　　号 U201414260

指导教师 潘林强

2018 年 03 月

译文要求

一、译文内容须与课题（或专业内容）联系，并需在封面注明详细出处。

二、出处格式为

图书：作者.书名.版本（第×版）.译者.出版地：出版者，出版年.起页～止页

期刊：作者.文章名称.期刊名称，年号，卷号（期号）：起页～止页

三、译文不少于5000汉字（或2万印刷符）。

四、翻译内容用五号宋体字编辑，采用A4号纸双面打印，封面与封底采用浅蓝色封面纸（卡纸）打印。要求内容明确，语句通顺。

五、译文及其相应参考文献一起装订，顺序依次为封面、译文、文献。

六、翻译应在第七学期完成。

译文评阅

导师评语

评分：_____（百分制）　　指导教师（签名）：_____

年　　月　　日

基于 Pareto 的多目标机器学习：概述与案例研究

Yaochu Jin, Senior Member, IEEE, and Bernhard Sendhoff, Senior Member, IEEE

概要：机器学习的本质是一个多目标任务。然而，传统做法不是只使用一个目标作为代价函数就是将多个目标汇总为一个标量代价函数。这主要是因为多数机器学习算法只能处理单个标量代价函数。在过去的十年中，在致力于使用基于 Pareto 多目标优化算法解决机器学习的问题上取得了巨大的进步，主要是因为使用进化算法和其他基于群体的随机搜索算法的成功。基于 Pareto 的多目标学习算法与其他算法相比具有更强的解决机器学习中各种类型的标量代价函数的能力，如聚类、特征提取、泛化能力的提高、知识提取和集合生成。不同多目标学习的一个共同的有益方法是可以通过分析有多个 Pareto 最优解组成 Pareto 解的集合的前沿来深入了解学习问题。本文概述了多目标机器学习的概况，重点介绍有监督学习。另外，还有一些研究案例被用来说明基于 Pareto 的机器学习方法的主要优点，例如，如何由 Pareto 最优解识别可解释的模型和可以概括所获得不可见数据模型。三种基于 Pareto 方法的多目标集合生成会进行详细比较和讨论。最后，提出了多目标潜在的有趣话题。

索引词——集合、基于进化的多目标优化、泛化、机器学习、多目标学习、多目标优化、神经网络、Pareto 优化。

I. 引言

机器学习与计算机自主学习的算法和技术密切相关，即通过经验自动改进[1]，[2]。任何机器学习方法都有两个步骤组成，首先选择候选模型，然后使用学习算法和可用数据估计模型的参数。通常，模型选择和参数估计结合在迭代过程中，大部分情况下，模型选择只凭直觉和经验一次完成。换句话说，用户凭经验选择模型，然后使用学习算法对模型进行参数估计。

机器学习算法可以大致的分为三类。一类是有监督学习，该模型应该近似的接近给定数据的输入和输出之间的映射，通常称为回归或者分类。第二类是无监督学习。数据的聚类就是一种典型的无监督学习算法，其中给定的一组数据将被分配给不同的子集（聚类），使得每个子集中的数据具有由距离测量定义的一些共同特征（相似性）。第三类是强化学习，其目的在于寻找代理人采取措施的方法，以最大限度地在给定环境中累积奖励。

所有的机器学习算法都会进行模型选择并根据一个或多个标准进行参数估计。在有监督

学习中通常的标准是使用误差函数反映近似质量。而在聚类中，最大化类内相似性，最小化类间相似性。在强化学习中，标准是价值函数，用来预测给定状态下给定动作的输出结果。因此，所有的学习问题都可以视为一种优化问题。此后，我们主要讨论有监督学习和聚类，因为在多标准强化学习中很少有工作被报道。此外，我们认为任何学习标准都是客观的，因为我们将从优化的角度来讨论学习问题。

现有的从优化角度下有监督学习的分类会在第二部分进行阐述，根据在学习算法中有多少目标被考虑以及采用单个标准还是基于 Pareto 的多目标优化算法。在第三部分和第四部分分别会对现有的基于 Pareto 的多目标有监督学习和无监督学习进行简要的概述。为了说明基于 Pareto 多目标的好处，在下一节会举例说明。第五部分会对实验案例研究包括神经网络模型、多目标进化算法（MOEA）、三种基准问题进行概述。如何从现有的 Pareto 前沿中确定可解释模型，如何从不可见数据中选择最有可能推广的模型，如何使用基于 Pareto 的优化算法生成集合会在第六部分说明。第七部分是总结与展望。

II. 单目标和多目标学习

我们将学习算法分为三类，即单目标学习、标量化多目标学习和基于 Pareto 的多目标学习。

A. 单目标学习

所谓单目标学习，是指只对一个目标函数进行优化的算法。以有监督学习为例，单目标学习算法往往是训练数据中的均方误差最小。

$$f = \frac{1}{N}\sum_{i=1}^{N}(y(i) - y^{d}(i))^{2} \qquad (1)$$

$\mathrm{y(i)}$和$\mathrm{y^{d}(i)}$分别是模型输出和期望输出，N 是训练数据集中数据对的数量。还可以使用其他几种误差度量作为目标函数。

最常使用的数据聚类算法是 k-均值聚类算法，最小化以下目标函数

$$f = \sum_{j=1}^{N}\sum_{x\in C_j}\left|\left|x - c_j\right|\right|^{2} \qquad (2)$$

$||.||$是数据节点x和类C_j的中心c_j之间的距离测量方法，K 是类的数量。

B. 标量化多目标学习

学习的本质是多目标的。在有监督学习中，记忆训练数据不是唯一的目标。还经常需要考虑其他几个目标。在回归和分类中，一个学习模型不仅应该对训练数据具有良好的近似性，还应该对同一问题的不可见数据具有良好的近似性。但是这个目标不能通过最小化(1)中的单个目标或任何相似误差测量方法实现。事实上，单纯最小化训练数据上的近似误差会产生过拟合现象，这意味着该模型在不可见数据可能表现不佳。换句话说，该模型不能归纳不可见数据。为了防止模型在训练数据上过拟合，必须控制模型复杂度。通常需要考虑的另一个共同目标是学习模型的可理解性或可解释性，当有监督学习用于从数据中提取知识时这点尤为重要。正如[4]中所提出的，机器学习的可解释性强烈依赖于模型的复杂性，一般来说，复杂度越低，模型越容易理解。在这两种情况下都必须考虑反应模型复杂度的第二个目标。为了控制复杂度，这两个目标可以汇总为一个标量目标函数

$$f = E + \lambda\Omega \qquad (3)$$

其中 E 是普通的误差函数，例如(1)中定义的误差函数，Ω是模型复杂度的量度，例如模型中的自由参数的数量，$\lambda > 0$是一个正超参数，由用户定义。通过这种方式，学习算法可以优化两个目标，尽管目标函数仍然是一个标量函数。

在神经网络规则化[5]，创建可解释的模糊规则[6]，[7]以及生成负相关集成成员[8]等机器学习中，广泛采用了标量化的多目标学习方法。不同于神经网络和模糊系统一些学习模型，如支持向量机[9]，稀疏编码[10]或学习任务，如受试者工作特征曲线(ROC）[11]，明确指出了复杂性控制不是必须的考虑不只一个目标，这在理论上属于标量化多目标学习的范畴。

与监督学习类似，在数据聚类中也可以考虑多个目标。一方面，很容易看出的是（2）中定义的目标函数强烈地偏向球形集群。对于具有不同类型的聚类结构的数据，其他目标函数可能更合适[12]。另一方面，也有人提出在开发聚类算法时应考虑反映扰动下聚类方案变化的稳定性[13]。

如果将一个标量化的目标函数用于多目标优化，则存在两个主要缺陷。首先，确定适当能够反映用户目的的适当超参数λ是很重要的。其次，只能得到单一的解决方案，从中不能获得对问题的深入了解。尤其对于多个目标是相互冲突的情况下，这点是尤为重要，因此不存在使所有目标同时优化的单一最佳解决方案。这对于多目标学习尤其如此，例如，减小近似误差常常导致模型的复杂性增加。除了上述两个缺点之外，从优化的角度来看，已经指出，

即使超参数被适当地指定，使用标量目标函数也不能实现期望的解决方案[14]。但是请注意，如果超参数在优化过程中动态改变，可以部分解决这个问题[15]。

基于 Pareto 的学习方法的另一个潜在优点是多目标化可以帮助学习算法摆脱局部最优，从而提高学习模型的准确性。一些经验证据已经在[16]和[17]中报道。然而，通过多目标化对学习曲线有利变化的严格证明仍有待证明。

C. 基于 Pareto 的多目标学习

使用 Pareto 方法来解决机器学习中的多个目标实际上是一个自然的想法。但是，这种方法直到十年前才被采用，并且仅在最近才流行。我们认为，原因是传统的学习算法和大多数传统的优化算法在使用基于 Pareto 的方法解决多目标问题时效率低下。在基于 Pareto 的多目标优化方法中，目标函数不再是标量值，而是矢量。因此，应该实现许多 Pareto 最优解，而不是单一的解决方案。

Pareto 最优是 Pareto 多目标优化中最重要的概念。考虑以下 m 个目标的最小化问题：

$$\min F(x),\ \ F = \{f_1(x), f_2(x), \dots, f_m(x)\}$$

如果$\forall j = 1,2,\dots,m, f_j(X) \le f_j(Y)$，并且对于$k \in \{1,2,\dots,m\}$存在$f_k(X) < f_k(Y)$则称解决方案 X 支配解决方案 Y。解决方案 X 如果不受任何其它可行的解决方案支配则被称为 Pareto 最优解。如前所述，如果目标相互矛盾，往往存在着多个 Pareto 最优解。由 Pareto 最优解组成的曲线或者曲面被称为 Pareto 前沿。实际中，我们通常不知真正的全局 Pareto 前沿在哪里，因此，由多目标进化算法(MOEA)得到的非支配解不一定是 Pareto 最优解。然而由多目标优化算法得到的非支配解基本可以认为是 Pareto 最优解。

基于 Pareto 的多目标学习算法遵循基于基于 Pareto 的多目标优化来处理学习问题。例如，(3)中的表量化双目标学习问题可以表示为基于 Pareto 的多目标优化，如下所示：

$$\min\{f_1, f_2\} \qquad (4)$$

$$f_1 = E \qquad (5)$$

$$f_1 = \Omega \qquad (6)$$

最流行的错误度量是(1)中所定义的均方误差。神经网络的复杂度可以是权重的平方之和：

$$\Omega = \sum_{i=1}^{M} {w_i}^2 \qquad (7)$$

或者是权重的绝对值之和：

$$\Omega = \sum_{i=1}^{M} |w_i| \qquad (8)$$

其中$w_i, i = 1,2,\dots,M$是神经模型中的权重，M 是所有权值的数量。前面提到的两种复杂性度量常常用于神经网络正则化，其中(7)被称为高斯正则化，(8)被称为拉普拉斯正则化。

比较(3)中描述的标量多目标学习和(4)中所描述的基于 Pareto 的多目标学习，我们发现我们不在需要基于 Pareto 的多目标学习中指定超参数。一方面，这样可以避免用户在学习之前确定超参数的负担，但另一方面，用户需要在学习之后根据偏好从所得到的 Pareto 最优解集中挑选一个或多个解决方案。这样可能会出现一个问题：标准化多目标学习和基于 Pareto 的多目标学习算法的区别在哪？正如我们在下一节中说明的那样，基于 Pareto 的多目标学习算法能够得到多个 Pareto 最优解，从中用户可以提取关于该问题的知识并在选择最终解决方案时做出更好的决策。

在下面的章节中，我们将简要回顾一下选择现有的基于 Paretobased 的有监督和无监督学习算法的研究。有关现有多目标学习研究的更新和更详细的介绍，请参阅[18]。

III. 多目标有监督学习

A. 早期想法

在 20 世纪 90 年代中期报道了将监督学习作为基于 Pareto 的多目标优化的第一个想法。其中神经学习问题最早被指定为多目标优化问题的工作之一在[19]中有报道，其中两个误差测量（L2 范数和 L∞范数）和一个复杂度测量（非零元素的个数）Volterra 多项式基函数网络和高斯径向基函数网络的最小最大化方法被最小化

$$f_1(W) = \left\| y(W) - y^d(W) \right\|_2 \qquad (9)$$

$$f_2(W) = \left\| y(W) - y^d(W) \right\|_\infty \qquad (10)$$

$$f_3(W) = C \qquad (11)$$

$$F(W) = min_w\{f_1'(W), f_2'(W), f_3'(W)\} \qquad (12)$$

其中 C 是非零权重的数量，$f_1'(W), f_2'(W), f_3'(W)$是$f_1(W), f_2(W), f_3(W)$的标准化值，W 是神经网络的权重矩阵。遗憾的是，单目标遗传算法已被用于实施学习过程，因此，只能得到一种解决方案。

文献[20]中讨论了处理竞争性学习目标的方法的薄弱性以及使用基于 Pareto 方法考虑权衡的必要性。在[21]中提出了一个重要的步骤，其中多层感知器网络的训练被表述为双目标优化问题。考虑到 MSE 和网络隐藏节点的数量。采用分支定界算法求解混合整数多目标问题。由于分支定界算法的能力有限，本文没有充分证明基于 Pareto 的机器学习方法的优点。

随着多目标进化算法(MOEAs)的日益普及，采用多目标进化算法(MOEAs)学习问题的想法变得越来越可行。基于 Pareto 的监督学习方法的现有研究大致可以根据其动机分为三类。

B. 泛化能力的改进

监督学习中的一个主要问题是生成不仅在训练数据上具有良好逼近性能的学习模型，而且还可以泛化到未知数据。为了达到这个目的，除了训练错误之外，还可以考虑几个目标。受到神经网络正则化的启发，使用基于 ε 约束的多目标优化方法将训练误差和绝对权重之和最小化[17]。Tikhonov 正则化术语被用作[23]中参数识别问题的第二个目标，而双目标问题则通过多目标实数编码进化算法解决。类似于[21]，前馈神经网络的训练误差和隐藏节点的数量使用基于 Pareto 的差分进化算法[24]被最小化。三种不同的正则化术语对复杂度最小化的影响已经在[25]中用多目标优化方法进行了讨论。与基于梯度的正则化算法得出的结论不同，它表明高斯正则化算法能够像拉普拉斯正则化算子一样在使用进化方法的同时有效地降低网络复杂度[26]。

提高神经网络泛化性能的另一个想法是尽量减少不同的，可能相互冲突的误差测量[27]，如欧几里得误差和鲁棒误差，它可以由

$$E_r = exp\left(\lambda\left|\vec{y} - \vec{y}^d\right|^p\right) \qquad (13)$$

其中 λ 和 p 是要定义的两个参数。在[28]中，研究了两种用于确定非支配性解决方案的不同方法，一种使用验证数据集而不是训练集，另一种使用增强方法。

[29]研究了基于多目标的神经网络的协作进化。算法中两个种群共同作用，模块（子网）种群和网络种群。模块群体再次由多个子群体组成，每个子群体都演变子网络（神经网络的子组件）的结构和权重。网络人口的染色体编码应挑选哪些子组件来构建整个神经网络。稳态遗传算法被用于网络种群。对于协同进化算法，确定模块群体中个体的适应值并不容易。在[29]中，讨论了评估模块适应性的几个标准。第一个标准涉及模块的性能，可以用不同的方式再次确定。例如，模块的性能可以是许多模型参与的最佳神经网络的平均适应值。或者，模块的性能可以通过更换或移除模块时最佳神经网络的平均适应度改变来确定。第二个标准

是模块存在的神经网络的数量，在优化期间将被最大化。第三个标准是模块的复杂性，包括连接数（NC），节点数量以及权重绝对值的总和。考虑网络人口的两个目标，即每个模块的性能和适应度。

除了前馈神经网络之外，使用基于 Pareto 的方法在准确性和复杂性之间的权衡也被考虑用于生成径向基神经网络[30]，[31]，支持向量机[32]-[34]，决策树[35]和分类器系统[36]。已经报道了基于 Pareto 的多目标学习在面部检测[37]，特征提取[38]，机器人学[39]和文本检索[40]中的有趣应用。

C. 规则提取中的可解释性增强

从数据或训练的神经网络提取逻辑或模糊规则是知识发现的重要途径。这里的一个关键问题是可解释性，也就是生成的规则的可理解性或透明度。有几个方面与规则的可解释性高度相关[41]，如紧凑性（规则数量，处所数量）和规则的一致性。对于模糊规则，模糊子集的划分应该很好区分，这样一个有意义的项可以附加到模糊子集上。可解释性的不同方面已经应用了标量化多目标优化[6]，[7]。

提高规则系统可理解性的第一个想法是从数据生成的大量规则中选择一个小的子集。基于 Pareto 的多目标遗传算法（MOGA）被用来生成模糊规则，通过将规则数量与分类错误进行权衡[42]。[43]和[44]也报道了类似的工作。更进一步的是包含第三个目标，该规则将规则长度(处所数量)[45]或选定输入变量的数量[46]最小化。为了提高模糊分割的可区分性，模糊子集之间的最大相似性除了精确性和紧凑性之外也被最小化[47]。为了进一步提高模糊分割的可区分性，考虑到分类和回归问题的精确性和紧凑性，在模糊规则的多目标优化中，将相似子集进行合并，删除单个子集，并将重叠子集进行分离[48][49]。

在从经过训练的神经网络中提取逻辑规则时，必须对几个目标进行优化，例如覆盖率，即通过规则集正确分类的模式数量，错误（即错误分类的模式数量）和紧凑性[50]。

基于 Pareto 的方法生成可解释的模糊规则的主要优点是用户能够从多个 Pareto 最优解中选择一个首选解决方案。

D. 不同集合的生成

如果集合的成员足够不同，那么学习模型的集合表现比单一学习模型好得多[51]。然而，在精确性和多样性之间存在权衡，因此合奏组成员高度多样化且足够准确是至关重要的[52]，

[53]。以前，通过使用不同的数据，不同的学习算法或不同的学习模型，促进了合奏成员的多样性[54]。另一种方法是开发一种学习算法，以减少训练误差并最小化合奏成员输出之间的相关性。传统上，集合成员之间的近似误差和输出相关性被归结为一个标量目标函数[8]，[55]。在[52]中，采用基于 Pareto 的方法来生成多样化和准确的种群，其中以下两个目标被最小化，

$$f_1 = \frac{1}{N}\sum_{i=1}^{N}(y(i) - y^d(i))^2 \qquad (14)$$

$$f_2 = \sum_{i=1}^{N}\left((y_k(i) - y(i)\right)\left[\sum_{j=1,j\neq k}^{N}\left((y_j(i) - y(i)\right)\right] \qquad (15)$$

其中$y_k(i)$是第 k 个集合成员的输出，$y(i)$是第 i 个训练样本集合的输出，N 是训练样本的数量，M 是集合中成员的数量。这项研究已经扩展到由三个层次的演变组成的演变综合框架[56]。第一个层由多个学习模型（如多层感知器，径向基函数网络和支持向量机）的混合演变而来。在第二层次上，使用不同的训练数据集来演化在第一层次上产生的混合集合。在第三层次上，在第二层次上产生的混合集合的所有同质学习模型的子集分别进化以最小化训练误差和相关性在集合成员之间。在每次迭代中，如果它基于训练误差和测试误差控制先前的最佳集合，则将由每种不同类型的模型组成的当前集合归档。档案中的集合作为最后的混合集合。

在[57]中已经提出了利用基于 Pareto 的学习来集成生成的不同想法，其中训练数据被分成两组，并且两个数据集上的差错被用作学习的两个目标

$$f_1 = \sum_{i=1}^{N_1}\left(y(i) - y_1^d(i)\right)^2 \qquad (16)$$

$$f_2 = \sum_{i=1}^{N_2}\left(y(i) - y_2^d(i)\right)^2 \qquad (17)$$

其中y_j^d是数据集$j, j = 1,2$中的训练数据，N_1, N_2是数据集的大小。应该注意的一点是，所使用的神经网络应尽可能地小，以避免在两个数据集上过拟合。

另一个想法是为了生成神经网络集合，将复杂性度量作为第二个目标[25,26]

$$f_1 = \sum_{i=1}^{N}\left(y(i) - y^d(i)\right)^2 \qquad (18)$$

$$f_2 = C \qquad (19)$$

其中 C 是神经网络中的 NC。这样，网络的多样性就可以通过不同的网络结构来实现，这是

通过种群成员总是具有不同的NC来保证的。对回归和分类问题的仿真结果表明，该方法对于生成神经网络集合是有效的。然而，应该注意的是，生成的非常简单的Pareto最优神经网络，其对训练数据的误差可能非常大。如果高精度模型是针对性的，这些网络不应该包含在集合中。一个在[25]和[26]中没有得到回答的问题是如何从非支配性解决方案中选择集合成员。我们将在案例研究中再次研究这个问题。

神经网络[29]的多目标合作协同进化方法也被用于生成神经网络集合[58]。在集合生成的情况下，一个种群演化单个神经网络，另一个演化神经网络集合。对于演变单一网络的人口，关于单一网络表现的目标，困难模式的表现（衡量，例如，通过对其进行错误分类的集合的数量来衡量）以及网络存在的整体的平均表现可以考虑用于评估单个网络的性能。此外，网络复杂性，合作能力和多样性是其他需要考虑的目标。除了[52]中使用的相关性度量，功能多样性，它测量两个神经网络输出之间的平均欧几里得距离，两个网络输出之间的互信息以及Yule's Q统计[59]也考虑了两个模型所产生的误差的相关性。对于种群集合而言，性能和模糊性是优化的两个目标。已经表明，使用多目标方法生成的集合的泛化性能明显优于经典方法生成的集合的泛化性能。

已经报道了径向基函数网络[60]和模糊规则系统[61]的基于Pareto的生成集合。

E. 其它

基于Pareto的多目标学习的早期工作受到特定应用的激励，即在不考虑泛化的情况下必须考虑多个目标。例如，在生成分类器的ROC曲线时，真实阳性率（TPR）和假阳性率（FPR）都要最小化。在[62]中，Niched Pareto GA [63]被用来生成神经网络分类器的ROC曲线[62]。已经表明，与通常通过改变训练后神经分类器的阈值来生成ROC曲线的传统方法相比，使用基于Pareto的方法可以获得更好的结果。请注意，传统上，ROC分析只是评估给定分类器的一种方法，但在基于Pareto的方法中，ROC曲线上的分类器是不同的。最近，大多数神经分类器使用基于Pareto的方法对ROC曲线生成已在[64]中进行了研究，并在[65]中研究了基于Pareto的多目标多类ROC分析。

系统控制是需要满足多个目标的另一个领域。在文献[66]中，基于Pareto的进化规划被用于最小化基于神经网络的控制器的下冲和整体跟踪误差。得到了许多Pareto最优解，并对某些典型Pareto解的控制性能进行了分析。

有监督的特征选择是机器学习任务之一，必须考虑所选特征的数量与使用特征的学习模

型的性能之间的权衡。因此，基于 Pareto 的多目标学习已经被研究[67]-[69]。

Ⅳ. 多目标无监督学习

在本节中，我们讨论已有的基于 Pareto 的多目标无监督学习的研究工作，主要是多目标数据聚类。在[70]中，基于 Pareto 的进化数据聚类考虑了四个目标。第一个目标是关注集群凝聚力，这有利于稠密集群，第二个目标是最大化集群之间的分离度，通过它们与整体质心的距离来衡量，第三个目标是为了减少集群的数量，第四个目标是最小化所选功能的数量。基于帕雷托的进化算法被用来实现多个 Pareto 最优解，而不是结合目标。通过分析单个 Pareto 最优解，可以确定显着特征和适当数量的聚类。

基于 Pareto 的数据聚类的优势已经在[71]中得到了令人信服的证明，其中可以通过分析 Pareto 前沿自动确定聚类的数量。在该论文中，两个目标被最小化以反映集群的紧凑性和数据点的连通性。簇的紧凑性由分区的总体偏差描述，并且连通性检查邻域中的数据点被分配给相同簇的程度

$$f_1 = \sum_{C_k \in C} \sum_{x_i \in C_k} \|x_i - c_k\|_2 \qquad (20)$$

$$f_2 = \sum_{i=1}^{N} \sum_{j=1}^{L} \gamma ij \qquad (21)$$

其中$\mathrm{C} = \{C_1, C_2, \ldots, C_k\}$是所有集群的总和，$C_k$是集群$C_k, k = 1,2, \ldots, K$的中心，$x_i$是集群$C_k$中的节点，K 是集群的数量，L 是提前定义的节点的数量，γij由下式定义

$$\gamma ij = \begin{cases} \frac{1}{j}, & \text{如果} x_i \text{和} \mathrm{NN}_j(x_i) \text{不在同一集群} \\ 0, & \text{其它} \end{cases}$$

其中$\mathrm{NN}_j(x_i)$是距离x_i最近的第 j 个邻接节点。

在偏差和连通性之间交换的 Pareto 最优解被绘制成使得 Pareto 最优解中包含的聚类数从左到右增加。有人认为，整体偏差随群集数量的增加而减少，并且当群集数量大于“真实”群集数量时，偏差最小化的增益将较小，而连接性的成本迅速增加。因此，如[72]所述，Pareto 最优的解决方案能够提供最大的性能增益，而不会增加聚类数量，从而提供正确的聚类数量。

V. 案例研究：实验设置

A. 神经网络模型

在案例研究中使用带有一个隐藏层的前馈神经网络。隐藏的神经元是非线性的，输出神经元是线性的。用于隐藏神经元的激活函数如下：

$$g(z) = \frac{x}{1 + |x|} \qquad (23)$$

在优化中，隐藏节点的最多为 10。权重初始化为-0.2~0.2 之间的值。

B. 基于 Pareto 学习的进化算法

*1). 神经网络的编码：*连接矩阵和权重矩阵被用来描述神经网络的结构和权重，见图 1. 连接矩阵指定了网络的结构，而权重矩阵决定了每个连接的强度。假设一个神经网络由总共 M 个神经元组成，包括输入和输出神经元，则连接矩阵的大小为 M*(M+1)，其中最后一列中的元素指示神经元是否连接到偏差值。在连接矩阵中，如果元素 $c_{ij}, i = 1,\ldots,M, j = 1,\ldots,M$ 等于 1，这意味着第 i 个和第 j 个神经元之间存在连接，并且信号从神经元 j 流向神经元 i。如果 j=M+1，则表明在第 i 个神经元中存在偏差。图 2 示出了连接矩阵和相应的网络结构。从图中可以看出，网络有两个输入神经元，两个隐藏的神经元和一个输出神经元。此外，两个隐藏的神经元都有偏差。

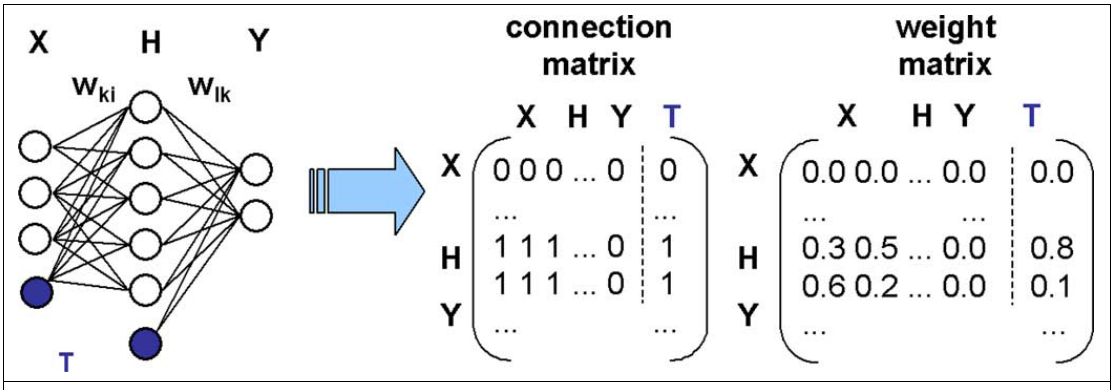

图 1　使用连接矩阵和权重矩阵对神经网络的结构和参数进行编码

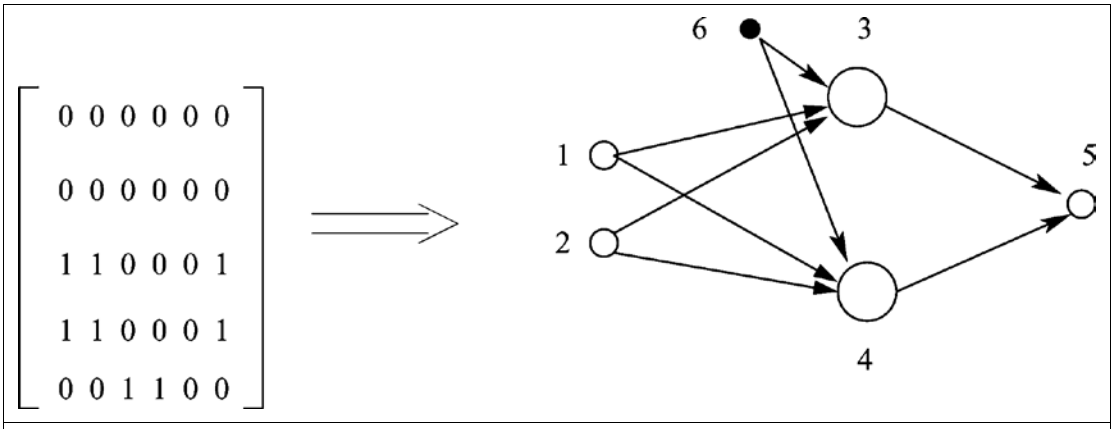

图 2　连接矩阵及其相应的神经网络结构的例子

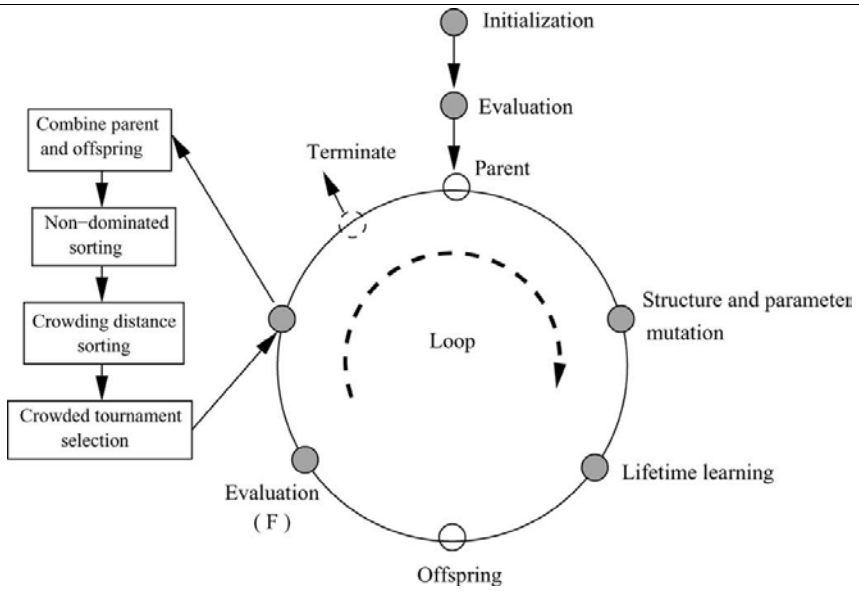

图 3　神经网络的进化多目标优化框架

*2). 结构和权重的变动：*演化算法已广泛用于优化神经网络的结构和参数，常常与基于梯度的局部搜索方法相结合[73]。在我们的案例研究中采用的神经网络的进化多目标优化框架如图 3 所示。与传统的进化优化相比，我们注意到在框架中只使用变异操作来改变神经网络的结构和参数，特定于神经网络，包括插入新的神经元或删除现有的神经元，添加或去除两个神经元之间的连接。高斯变异被应用于权重

$$\Delta w_{ij} = \mathrm{N}(0, \sigma_w) \qquad (24)$$

其中w_{ij}表示连接神经元 j 和神经元 i 的权重，σ_w是高斯分布的标准偏差。

*3). 终生学习：*变异后，使用 Rprop 算法[74]的改进版本进行学习，以精确调整权重。终生学习后，每个人对近似误差（f1）的适应度被更新。此外，在终身学习中修改的权重被编码回染色体，这就是所谓的拉马克式遗传。

Rprop 学习算法[75]被认为是一种快速和鲁棒的学习算法。在每次迭代中，权重按以下方式修改

$$\Delta w_{ij}{}^{(t)} = -\text{sign}\left(\frac{\partial E^{(t)}}{\partial w_{ij}}\right)\Delta_{ij}^{(t)} \qquad (25)$$

其中 sign(.)是符号函数，$\Delta_{ij}^{(t)} \geq 0$是步长，对于所有权重，初始化为Δ_0。每个重量的步长调整为

$$\Delta_{ij}^{(t)} = \begin{cases} \xi^{+}\Delta_{ij}^{(t-1)}, if\ \dfrac{\partial E^{(t-1)}}{\partial w_{ij}} \times \dfrac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ \xi^{-}\Delta_{ij}^{(t-1)}, if\ \dfrac{\partial E^{(t-1)}}{\partial w_{ij}} \times \dfrac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ \Delta_{ij}^{(t-1)}, otherwise \end{cases} \qquad (26)$$

其中$0 < \xi^{-} < 1 < \xi^{+}$。为了防止步长变得过大或过小，它们以$\Delta_{\min} \leq \Delta_{ij} \leq \Delta_{max}$为界。

在权重更新之后，有必要检查偏导数是否改变了符号，这表示前一步可能太大，并因此错过了最小值。在这种情况下，先前的重量变化应该缩回

$$\Delta w_{ij}{}^{(t)} = -\Delta_{ij}^{(t-1)}, \quad if\ \frac{\partial E^{(t-1)}}{\partial w_{ij}} \times \frac{\partial E^{(t)}}{\partial w_{ij}} < 0 \qquad (27)$$

回想一下，如果在第 t 次迭代中重量变化被收回，则应该将$\partial E^{(t)}/\partial w_{ij}$设置为 0。

在参考文献[74]中，有人认为(27)中体重回缩的情况并不总是合理的。只有偏导数变化符号和逼近误差增加时，才应收回重量变化。因此，(27)中的重量缩回条件被修改如下：

$$\Delta w^{(t)} = -\Delta_{ij}^{(t-1)}, \quad if\ \frac{\partial E^{(t-1)}}{\partial w_{ij}} \times \frac{\partial E^{(t)}}{\partial w_{ij}} < 0$$
$$and\ E^{(t)} > E^{(t-1)} \qquad (28)$$

在几个基准问题上已经表明，改进的 Rprop（称为 Rprop+）比 Rprop 算法具有更好的性能[74]。

*4). 选择：*多目标优化与标量优化最显着的区别在于选择方法。在我们的研究中，采用 NSGA-II[76]的选择方法，它由四个主要步骤组成。首先，父集群和子集群合并。这意味着 NSGA-II 是一种精英主义。其次，合并种群按非支配等级进行排序。在排序期间，合并人群中的非支配性解决方案被分配为 1 级，这属于第一个非支配性前沿。这些个体在暂时上被从种群中移除，而剩余种群中的非支配个体被识别，组成第二非支配解决方案，并且被指定为等级 2。这个过程重复直到合并种群中的所有个体被指定到等级 1 到 R，假定总共可以识别出 R 个非支配前沿。第三，计算反映特定解决方案附近的拥挤度的拥挤距离。在非支配前沿 j(j=1,...,R)中解决方案 i 的拥挤距离是目标空间中解决方案s_i^j的两个邻居之间的距离

$$d_i^j = \sum_{k=1}^{m} \left| f_k\left(s_{i-1}^j\right) - f_k\left(s_{i+1}^j\right) \right| \qquad (29)$$

其中 m 是目标空间多目标优化问题，解s_{i-1}^{j}和s_{i+1}^{j}是解s_i^j的两个相邻解。每个非支配前沿的边界解决方案都会分配一个很大的距离。在这里，拥挤距离越大，解决方案周围的拥挤程度就越小。第四，在非主导排名和拥挤之间进行对比。给定两个随机选择的个体，具有更好（较低）排名的解决方案将赢得比赛。如果两种解决方案具有相同的等级，那么具有较大拥挤距离的解决方案将获胜。如果这两个解决方案具有相同的排名和相同的拥挤距离，随机选择一个赢家。这个过程一直持续到产生所需数量的后代。

表 1　算法的参数设置

Neural Network Initialization	
maximum number of hidden neurons	10
initial weights	-0.2 ∼ 0.2
Evolutionary Algorithm	
population size	100
mutation rate	0.20
σ_w	0.1
Rprop$^+$ Algorithm	
ξ^+	1.2
ξ^-	0.5
Δ_0	0.01
$\Delta_{\max}$	50
$\Delta_{\min}$	10^{-6}

表 1 中总结了模拟中使用的参数设置。

C. 基准问题

*1). 威斯康星州乳腺癌数据：*加利福尼亚大学欧文分校（UCI）机器学习数据库存储的威斯康星州乳腺癌诊断问题由威斯康星大学麦迪逊分校医院 W.H.Wolberg 博士收集[77]。基准问题包含 699 个示例，每个示例都有 9 个输入和 2 个输出。输入为：团块厚度(x_1)，细胞大小均匀性(x_2)，细胞形状均一性(x_3)，边缘粘附性(x_4)，单个上皮细胞大小(x_5)，裸露核(x_6)，平淡染色质(x_7)，正常核仁(x_8)和有丝分裂(x_9)。所有输入都被标准化，更确切地说$x_1,\dots,x_9 \in \{0.1,0.2,\dots,0.8,0.9,1.0\}$。这两个输出是一个互补的二进制值，即如果第一个输出是 1，意味着“良性”，那么第二个输出是 0。否则，第一个输出是 0，这意味着“恶性”，第二个输出是 1。因此，只使用第一个输出。

*2). 糖尿病数据：*皮马印第安人糖尿病数据由 768 个数据对组成，其中 8 个属性在 0 和 1 之间归一化[77]。这 8 个属性是怀孕(x_1)，血浆葡萄糖浓度(x_2)，血压(x_3)，三头肌皮褶厚

度(x_4)，2 小时血清胰岛素(x_5)，体重指数(x_6)，糖尿病家系功能(x_7)和年龄(x_8)。在这个数据库中，268 个实例是正数(输出等于 1)，500 个实例是负数(输出等于 0)。

*3). 鸢尾花数据：*我们研究的第三个数据集是鸢尾花数据[77]。该数据集包含三类 40 个实例，其中每个类是一种鸢尾花。这三个类别是：山鸢尾(第 1 类，由-1 表示)，变色鸢尾(第 2 类，由 0 表示)和维吉尼亚鸢尾(第 3 类，由 1 表示)。四个属性被用于预测鸢尾花的类别，即萼片长度(x_1)、萼片宽度(x_2)、花瓣长度(x_3)和花瓣宽度(x_4)，所有的单位都是厘米。在这三类中，第一类与其他两类可线性分离，第二类和第三类不是线性可分的。为了简化知识提取，我们用三个输出来重新构造数据，其中类别 1 由{1,0,0}表示，类别 2 由{0,1,0}表示，类别 3 由{0,0,1}表示。

VI. 案例研究：结果

根据上一节介绍的多目标进化算法(MOEA)，我们在本节中说明使用基于 Pareto 的多目标学习的好处。我们生成了许多 Pareto 最优神经网络模型，将训练数据的准确性与网络复杂度进行交换。我们展示了三个基准问题，如何从提取到的可理解逻辑规则、最有可能归纳不可见数据的网络和得到的 Pareto 最优解集合中识别可解释神经网络之后，我们比较了 Abbass[57]，Chandra 和 Yao[52]以及 Jin 等人[26]提出的使用基于 Pareto 的多目标学习生成神经网络集合的三种方法。

A. 识别可解释模型

正如[4]所述，神经网络的可解释性主要取决于其复杂性。网络越简单，就越容易理解嵌入在神经网络中的知识。在模糊系统这点也是一样的[6]，[41]。

当我们以基于 Pareto 的方法将网络的准确性和复杂性最小化时，我们能够得到许多 Pareto 最优解，其复杂度范围从非常简单到非常复杂都有。我们认为，Pareto 前沿中的简单的 Pareto 最优神经网络是可以从可理解逻辑规则中提取到的可解释模型。在提供基准问题的例子之前，我们首先简要介绍一下我们在这个案例研究中采用的规则提取方法，它与[78]中使用的类似。考虑一个简单的具有单输入，一个隐藏神经元和一个输出神经元的神经网络，参见图 4。对于二元分类问题，如果输出小于 0.5，我们通常假设一个实例被标记为第 1 类。否则，它被标记为第 2 类。为了对决策更有说服性，我们还可以定义更强的标准，例如：

$$\begin{cases} y \geq 0.75, & class\ 1 \\ y \leq 0.25, & class\ 2 \\ 0.25 \leq y \leq 0.75, & undecided \end{cases} \tag{30}$$

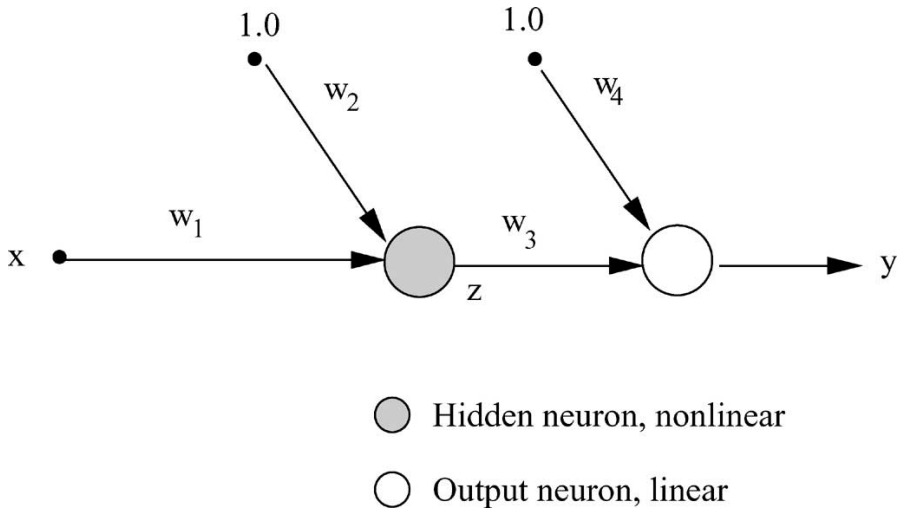

图 4　用于提取逻辑规则的典型简单网络

接下来，我们将展示如何使用定义的阈值从神经网络中派生出规则，设输出为隐藏神经元 z，则第一类的规则应设置为

$$\mathrm{w}_3\mathrm{z} + \mathrm{w}_4 \geq 0.75 \tag{31}$$

然后，我们可以得到 w

$$\mathrm{z} \geq \frac{(0.75 - \mathrm{w}_4)}{\mathrm{w}_3}, \quad if\ w_3 > 0$$

$$\mathrm{z} \leq \frac{(0.75 - \mathrm{w}_4)}{\mathrm{w}_3}, \quad if\ w_3 < 0$$

考虑第一个例子，定义$(0.75 - \mathrm{w}_4)\ /\mathrm{w}_3\ = \theta_1 > 0$，然后我们可以得到

$$\frac{\mathrm{w}_1 x + \mathrm{w}_2}{1 + |\mathrm{w}_1 x + \mathrm{w}_2|} \geq \theta_1 \tag{32}$$

因为$\theta_1 > 0$，所以$\mathrm{w}_1 x + \mathrm{w}_2$也必须大于 0 以满足第一类的条件，所以

$$\frac{\mathrm{w}_1 x + \mathrm{w}_2}{1 + \mathrm{w}_1 x + \mathrm{w}_2} \geq \theta_1 \tag{33}$$

继而得到

$$x \geq \frac{\theta_1 - \mathrm{w}_2(1 - \theta_1)}{\mathrm{w}_1(1 - \theta_1)}, \quad if\ \ \mathrm{w}_1(1 - \theta_1) > 0 \tag{34}$$

$$x \leq \frac{\theta_1 - \mathrm{w}_2(1 - \theta_1)}{\mathrm{w}_1(1 - \theta_1)}, \quad if\ \ \mathrm{w}_1(1 - \theta_1) < 0 \tag{35}$$

设置$[\theta_1 - \mathrm{w}_2(1 - \theta_1)]/[\mathrm{w}_1(1 - \theta_1)] = \theta_2$，就可以得到下面两条定义第一类的条件规则：

$$if\ \ x \geq \theta_2, \quad then\ \ class\ 1, if\ \ \mathrm{w}_1(1 - \theta_1) > 0$$

$$if\ \ x \leq \theta_2, \quad then\ \ class\ 1, if\ \ \mathrm{w}_1(1 - \theta_1) < 0$$

但是，请注意，可能会发生无法从神经网络中提取规则的情况。例如，如果$\forall z, w_3 x + w_4 < 0.75$。在这种情况下，神经网络无法区分这两个类。

1). 威斯康星州乳腺癌数据：对于规则提取，所有可用数据都用于训练神经网络。典型运行的 Pareto 最优解图如图 5 所示。正如我们稍后将要展示的那样，从不同运行中获得的最简单的 Pareto 最优神经网络几乎是相同的。

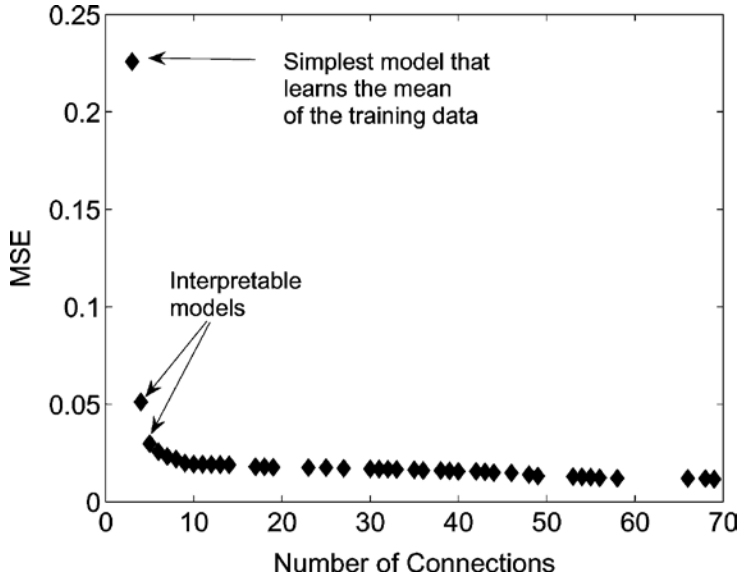

图 5　得到的由 41 个解组成的乳腺癌数据的典型 Pareto 前沿

现在让我们看看最简单的 Pareto 最优神经网络。最简单的神经网络总共有三个连接，其中没有选择输入。换句话说，神经网络的输入是恒定的，参见图 6.有趣的是，这个神经网络准确地学习了训练数据的平均输出。

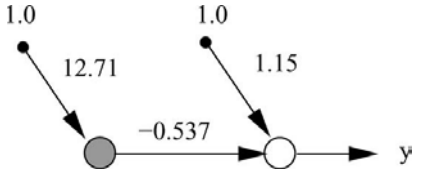

图 6　乳腺癌数据最简单的 Pareto 最优网络模型，它准确地利用训练数据的均值

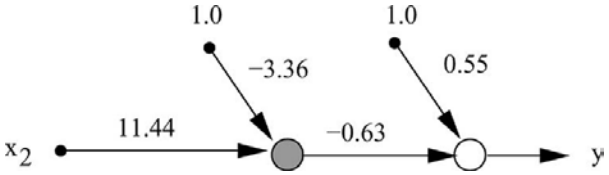

图 7　用于乳腺癌数据的具有四个连接的 Pareto 最优网络模型

第二个最简单的网络如图 7 所示，它有四个连接。在九个输入属性中，只有x_2（细胞大小的一致性）被选中，这意味着x_2可能是确定一个实例是良性还是恶性的最重要特征。网络的 MSE 是 0.051。从网络中，使用先前描述的规则提取方法可以提取以下两个规则：

$$if\ \ x_2 \leq 0.2, \quad \text{良性}$$

$$if\ \ x_2 \geq 0.4, \quad \text{恶性}$$

有了这两条简单的规则，602 个实例的正确分类率为 97.0%，其余 97 个实例未确定，回顾阈值设置为 0.75 和 0.25 以确保决策足够准确。但是，如果我们将分类阈值设置为 0.5，

则所有实例的正确率为92.4%，并得到以下规则：

$$if\ \ x_2 \le 0.3,\ \ 良性$$

其它情况为恶性

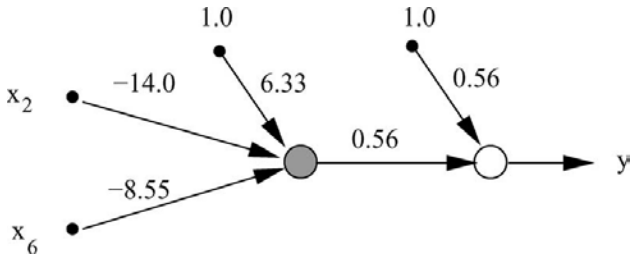

图8　用于乳腺癌数据的具有五个连接的Pareto最优网络模型

下一个简单的Pareto最优神经网络有五个连接，其中x_2和x_6都被选作输入特征（见图8）。该模型的MSE为0.029。从这个神经网络中，可以提取以下两个规则：

$$if\ \ 14x_2 + 8.55x_6 \le 5.81,\ \ 良性$$

$$if\ \ 14x_2 + 8.55x_6 \ge 7.55,\ \ 恶性$$

使用这两个规则，680个实例的正确分类率为97.2%，其余19个实例未确定。如果阈值设置为0.5，则所有实例的正确率为96.4%，并得到以下规则：

$$if\ \ 14x_2 + 8.55x_6 \le 6.45,\ \ 良性$$

其它情况为恶性

2). 糖尿病数据:对糖尿病数据进行相同的实证研究。得到的Pareto前沿如图9所示。

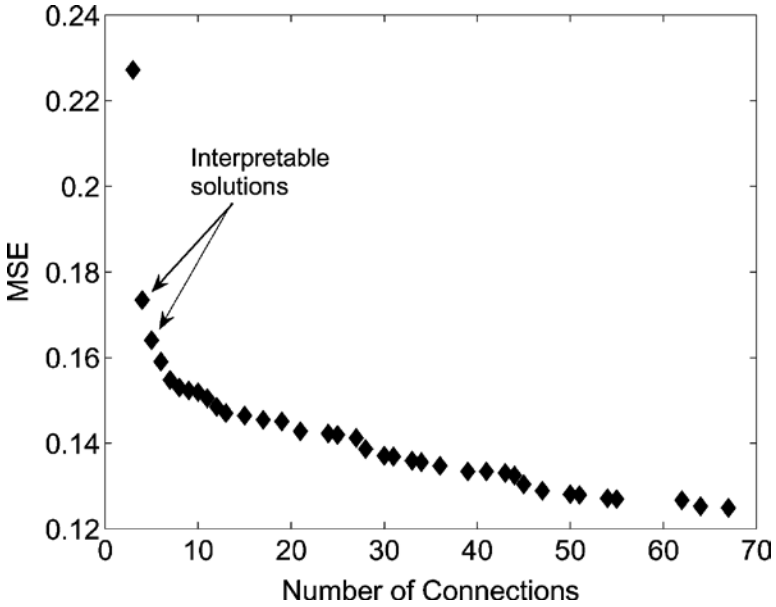

图9　得到的由37个解组成的糖尿病数据得典型Pareto前沿

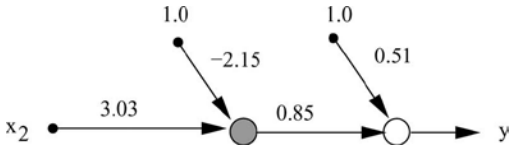

图 10　具有四个糖尿病数据连接的 Pareto 最优网络模型

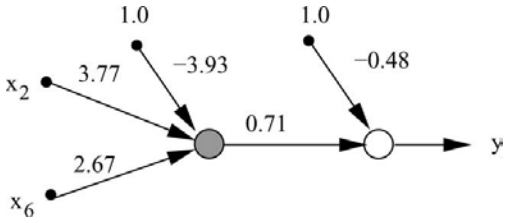

图 11　具有五个糖尿病数据连接的 Pareto 最优网络模型

与乳腺癌数据相同，最简单的 Pareto 最优解包含三个连接并学习输出值的平均值。具有至少一个选定属性的两个简单 Pareto 解决方案绘制在图 10 和 11。两种简单网络模型的 MSE 分别为 0.17 和 0.16。

从具有四个连接的神经网络（参见图 10）中，可以提取以下两个规则：

$$If\ x_2\ \leq\ 0.83, then\ positive$$

$$If\ x_2\ \geq\ 0.56, then\ negative$$

通过应用上述两个规则，我们能够对正确分类率为 85.4%的 413 个实例做出决定。剩下的 355 个实例不能用这两个规则来确定。

如果我们将阈值设置为 0.5，则获得以下规则：

$$\begin{cases} If\ x_2 \leq\ 0.72, then\ positive \\ otherwise\ negative \end{cases} \qquad (37)$$

使用上述规则的正确分类率在所有 768 个实例中为 75.0%。

当阈值设置为 0.75 和 0.25 时，可以针对图 11 中的神经网络获得以下规则：

$$If\ 3.77x_2 + 2.67x_6 \leq\ 4.54, then\ positive$$

$$If\ 3.77x_2 + 2.67x_6 \geq\ 3.46, then\ positive$$

有了这两条规则，正确的分类率为 85.4%，其余 308 个例子未定。如果阈值设置为 0.5，则我们有以下规则：

$$\begin{cases} If\ \ 3.77x_2 + 2.67x_6 \leq\ 3.97, then\ positive \\ otherwise\ negative \end{cases} \qquad (38)$$

根据上述规则，所有 768 个实例的正确分类率为 77.0%。

3). 鸢尾花数据：来自鸢尾花数据的 Pareto 前沿如图 12 所示，其中包括 20 个解决方案（两个 Pareto 最优解具有相同的 MSE 和复杂度）。再次，具有七个连接的最简单的网络近似于输出的平均值。

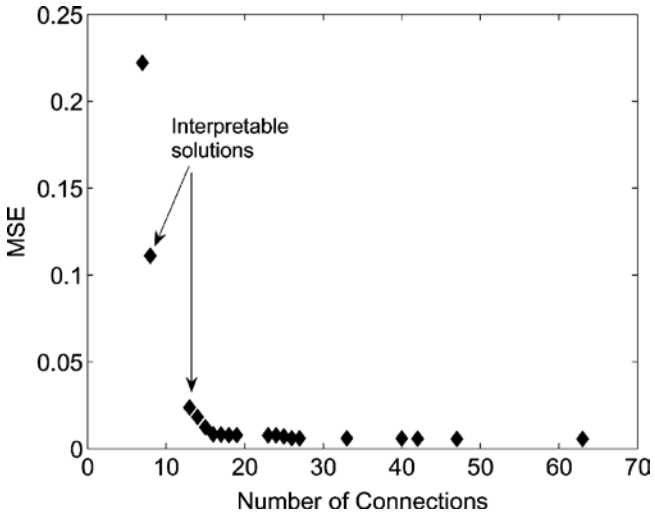

图 12　得到的由 20 个解组成的鸢尾花数据的典型 Pareto 前沿

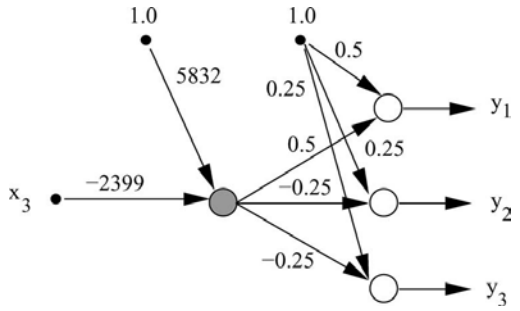

图 13　具有 8 个鸢尾花数据连接的 Pareto 最优网络模型。在这个模型中 x_3 被选为输入

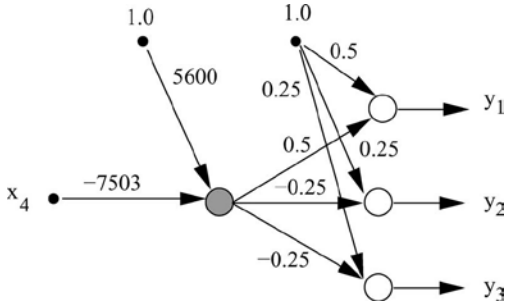

图 14. 具有 8 个鸢尾花数据连接的 Pareto 最优网络模型。在这个模型中 x_4 被选为输入

具有八个连接的两个 Pareto 最优网络绘制在图 13 和图 14 中。从图中我们注意到只有一个属性（x3 或 x4）被选中。从图 13 的网络中，可以提取以下规则：如果

$$\text{如果 } x_3 \leq 2.4, \text{则为山鸢尾} \qquad (39)$$

类似地，下面的规则可以从图 13 中的网络提取：

$$\text{如果 } x_4 \leq 0.80, \text{则为山鸢尾} \qquad (40)$$

可以很容易地验证这两个规则能够正确地将山鸢尾与其他两个类别分开。

具有 13 个连接的神经网络模型如图 15 所示。有趣的是，只有 x4 用于分类。从这个神经网络，我们可以提取以下三个规则：

$$\begin{cases} x_4 \leq 0.6, & \text{山鸢尾} \\ 1.1 \leq x_4 \leq 1.6, & \text{变色鸢尾} \\ x_4 \geq 1.7, & \text{维吉尼亚鸢尾} \end{cases}$$

所有实例的正确分类率为 91.3%。请注意，当鸢尾花数据的阈值设置为 0.5 时，分类

率几乎相同。

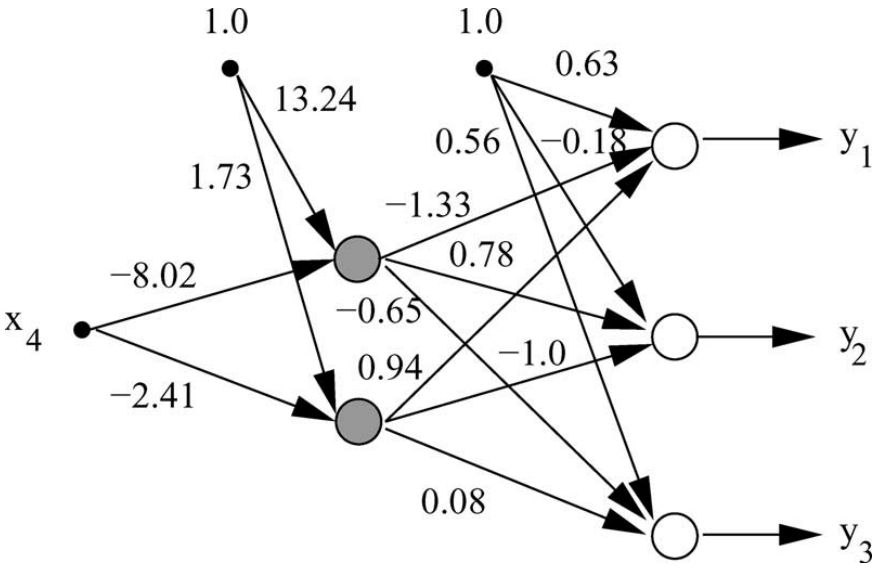

图 15　具有 13 个鸢尾花数据连接的 Pareto 最优网络模型被选为输入

从三个基准问题中，我们可以得出结论：通过将精度与复杂度进行权衡，基于 Pareto 的多目标优化算法能够找到解决问题最简单的最简单结构。此外，简单的 Pareto 最优网络能够捕获数据中嵌入的主要知识，从而可以提取可解释的逻辑规则。与从经过训练的神经网络中提取规则的其他方法[79]，[80]相比，基于 Pareto 的方法非常简单和高效。此外，正如我们在三个基准问题上所展示的，多个可解释 Pareto 最优解决方案还提供了额外的知识，可以帮助用户理解问题。

B.　基于 Pareto 前沿的模型选择

模型选择在机器学习中是一个深入研究的课题[81]，[82]。如果有足够的数据可用，模型选择的最佳方法是将数据分成三个子集，其中第一个子集(训练数据)用于构建模型，第二个(验证数据)用于估计选择模型，第三个(测试数据)用于访问所选模型的泛化误差。在实际应用中数据通常不足，在这种情况下可以使用信息理论标准[81]，[82]，如 Akaike 信息标准(AIC)和贝叶斯信息标准(BIC)，或重采样技术如 k 折交叉验证[82]。

在本节中，我们将证明，Pareto 方法在处理精度-复杂度权衡的问题上，提供了一个经验性、有趣的选择，以选择出对不可见数据上同样具有良好概括性的模型。基本的论点是模型的复杂性应该与要学习的数据和学习算法的能力相匹配。当模型的复杂性过大时，学习对随机影响变得敏感，对不可见数据的结果将是不可预测的，即可能发生过度拟合。受到多目

标数据聚类[71]中确定正确聚类数量的工作的启发，数据的适当复杂性可以通过*归一化性能增益(NPG)*来确定，

$$\mathrm{NPG} = \frac{MSE_j - MSE_i}{C_i - C_j} \qquad (41)$$

其中 MSEi，MSEj 和 Ci，Cj 是训练数据上的 MSE，以及第 i 和第 j 个 Pareto 最优解的 NC。当解决方案按照日益复杂的顺序排列时，以下关系成立

$$C_{i+1} > C_i$$

$$MSE_{i+1} \leq MSE_i$$

我们假设如果模型的复杂性低于数据的复杂性，复杂性的增加将导致性能(NPG)的显着增加。随着复杂性的不断增加，NPG 逐渐降至零。此时，模型的复杂性与数据的复杂性相匹配。复杂性的进一步增加可能会进一步提高训练数据的性能，但随着训练数据过度拟合的风险增加。

我们现在要验证三种基准问题的建议方法形式选择。在这部分模拟中，可用数据被分成一个训练数据集和一个测试数据集。对于乳腺癌数据，525 例用于训练，174 例用于测试。糖尿病数据的训练集包含 576 个样本和测试集包含 192 个样本。最后，120 个实例用于训练，其余 30 个实例用于测试的鸢尾花数据。

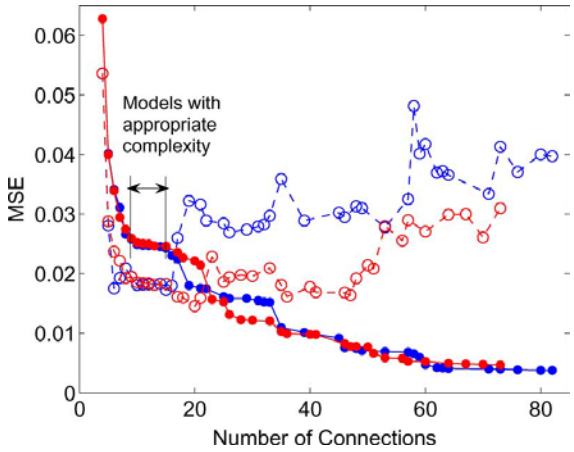

图 16　来自两次独立运行的 Pareto 最优解的准确性与复杂性：乳腺癌数据。点表示训练数据，圈表示测试数据

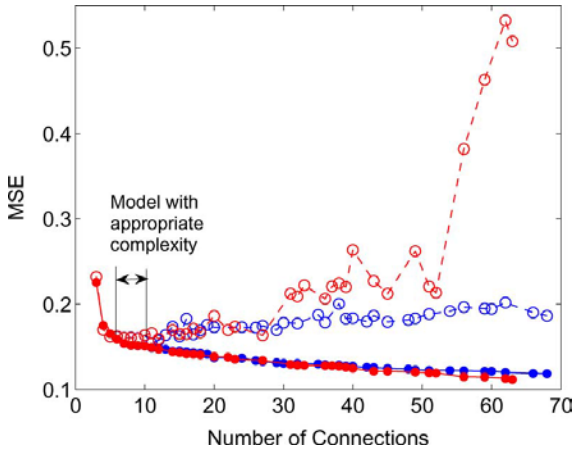

图 17　来自两个独立运行的 Pareto 最优解的准确性与复杂性：糖尿病数据。点表示训练数据，圈表示测试数据

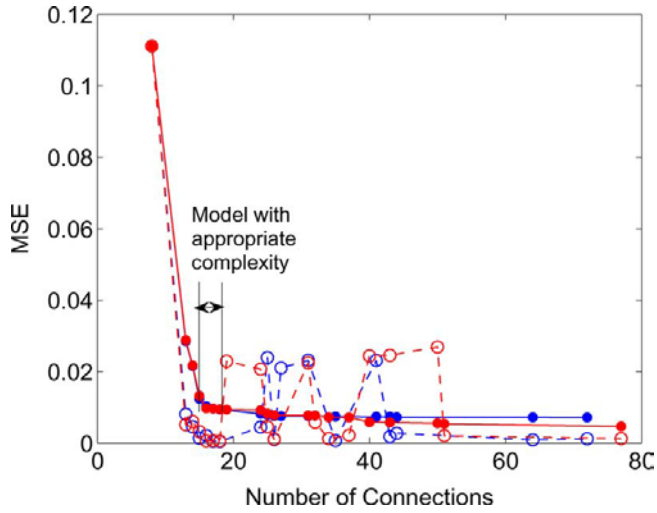

图 18　两次独立运行的 Pareto 最优解的精度与复杂性：鸢尾花数据。点表示训练数据，圈表示测试数据

由三个基准问题的两次独立运行产生的 Pareto 前沿在图 16、17 和 18 中给出。圆点表示训练数据集上的结果，圆圈表示测试数据的结果。这三个问题的两个独立运行的 NPG 绘制在图 19、20 和 21 中。

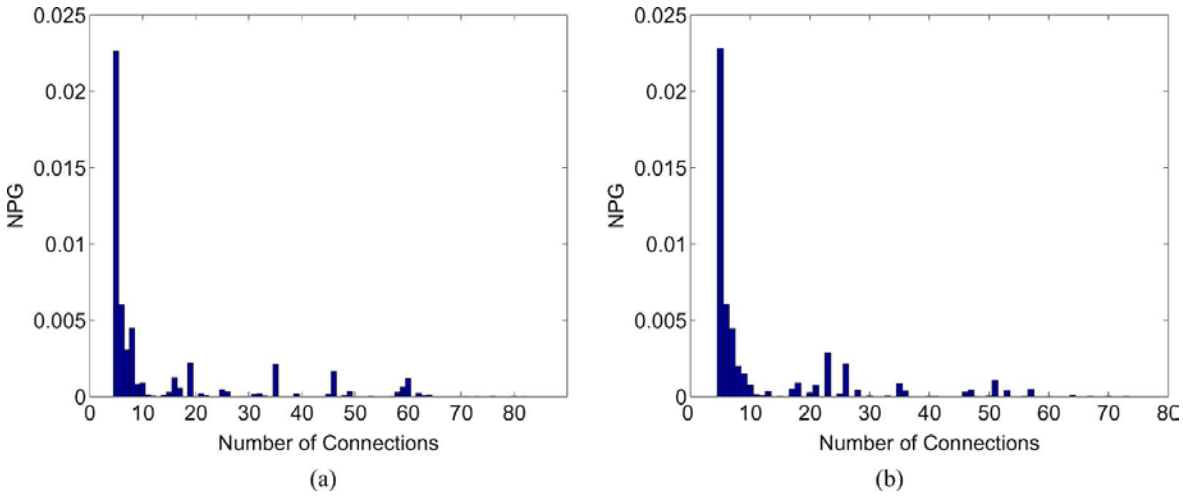

图 19　乳腺癌数据的两次独立运行的 NPG

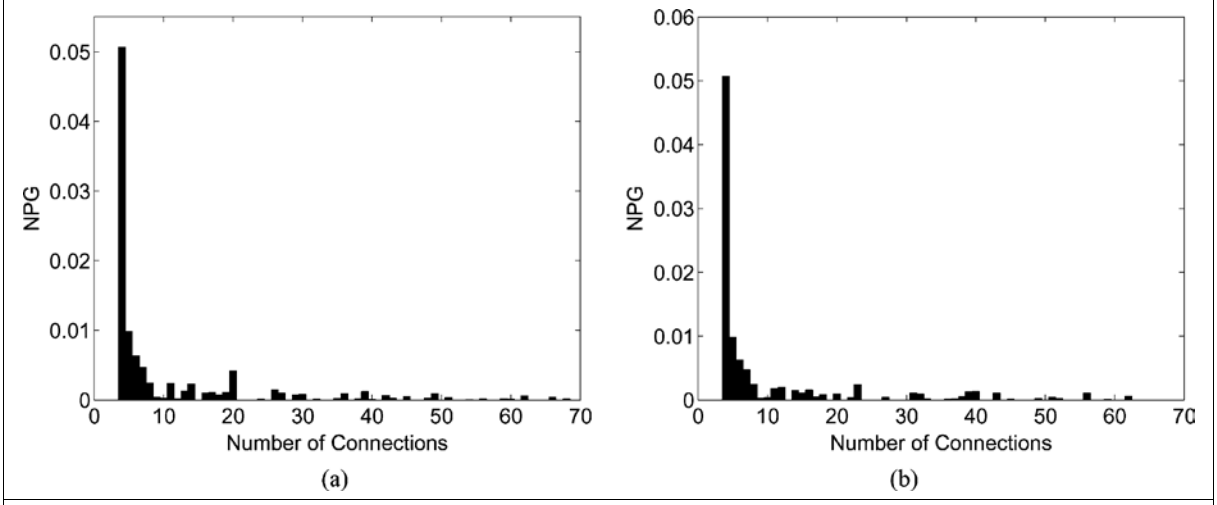

图 20　用于糖尿病数据的两次独立运行的 MPG

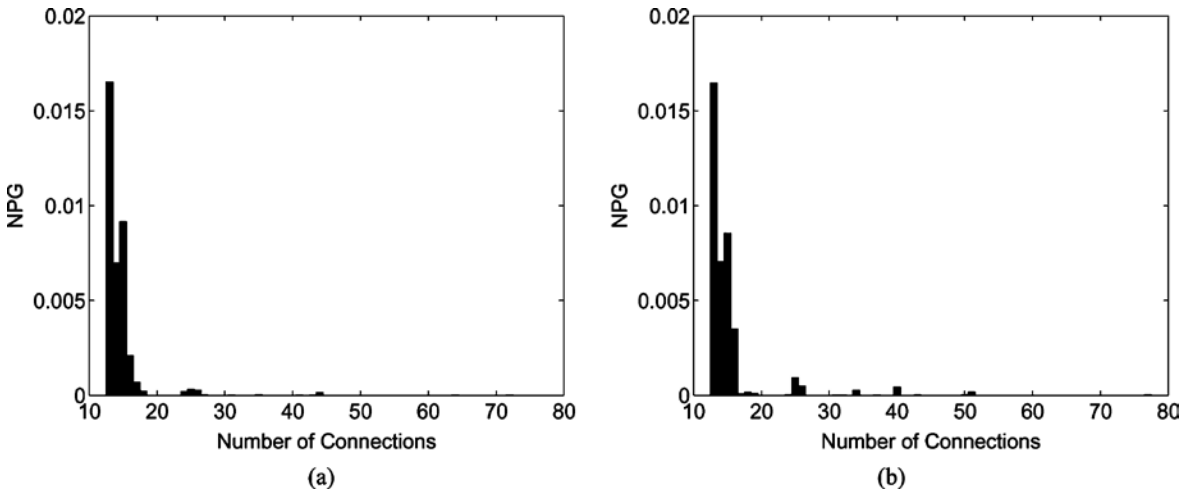

图 21　来自鸢尾花数据的两次独立运行的 NPG

我们首先分析乳腺癌数据的结果。从图 19 中我们注意到，当 NC 在 12 和 14 之间时，NPG 在性能增益的第一个峰值之后降到 0。同时，从图 16 可以看出，来自不同运行的训练数据的学习性能在 NC 大于 17 时开始波动。这两个事实表明该问题的神经网络的适当复杂度在 12 和 17 之间。从图 16 我们可以看出，当测试数据的误差在复杂度在建议范围内。

可以对糖尿病数据和虹膜数据进行类似的观察。对于糖尿病数据，当神经网络的 NC 大约为 10 时，NPG 首先下降到 0。另外，在 NC 达到 13 之后，两次运行之间的差异变大。从这两个观察，我们得出结论：神经网络对糖尿病数据应该在 8-10 左右。出于同样的原因，对于鸢尾花数据，神经网络的 NC 应该在 16 到 18 之间。

所提出的模型选择方法是经验性的，需要在更多问题上进行验证。为了清楚起见，我们只在上述分析中绘制两次独立运行的结果。图 28-30 中绘制了 10 次独立运行的结果。从这些结果中，我们可以证实，当训练数据的学习性能在不同运行中稳定时，神经网络的泛化性

能是好的。

使用我们的经验方法来选择单个模型是很困难的。相反，如果选择多个可能具有良好泛化性能的模型来构建一个集合，它将会更加可靠。这个主题将在下一节讨论。

C. 生成多样、准确的集合成员

在本节中，我们比较了三种基于 Pareto 的多目标方法对集合生成的影响。第一种方法在 Abbass 中提出[57]，其中两个数据集的准确性作为两个目标。Chandra 和 Yao[52]描述了第二种方法，其中考虑了准确性和多样性之间的折衷以生成合奏。Jin 等人提出了本节研究的最后一种方法。[25]，[26]，其中神经网络的准确性和 NC 被认为是两个相互冲突的目标。除了在 Abbass 的方法中，三个基准问题的训练数据被平均分成两个数据集，因此两个数据集上的近似误差可以作为两个目标来计算，实验装置与以前的研究相同。

另一个在 Abbass[57]中没有明确提到的问题是在多目标学习环境下的终生学习。请注意，RProp 被采用为终身学习算法，仅适用于单目标学习。在钱德拉和金的方法中，这不是问题，因为终身学习只适用于其中一个目标。然而，当两个目标都是近似误差时，应该应用多目标生命期学习，这对于基于梯度的学习算法来说并不简单。Jin 等人[83]，建议终生学习应该在两个目标之间随机切换，以实现不同的 Pareto 最优解。在这项研究中，终生学习以相等的概率在两个目标之间切换。为了进行比较，还进行了模拟，其中终身学习是单目标性质的，即 RProp 应用于两个数据集的组合。

图 22(a)绘制了 10 次乳腺癌数据运行的 Pareto 最优解，其中寿命学习在两个数据集之间切换，而图 22(b)中将寿命学习应用于两个数据集的组合。在图中，圆点表示训练数据的结果，圆圈表示测试数据的结果。从这些结果中，我们可以做出以下观察。首先，通过切换两个数据集之间的终身学习，可以实现更多不同的解决方案。其次，训练数据的良好性能不能确保测试数据的良好性能。正如[53]中所建议的那样，合奏成员应该既准确又多样。换句话说，成员精度不高的乐团表现不佳。这表明，如果 Pareto 最优解被用作集合成员，则集合的质量将会很差。第三，对数据组合的终身学习导致严重的过度配合。

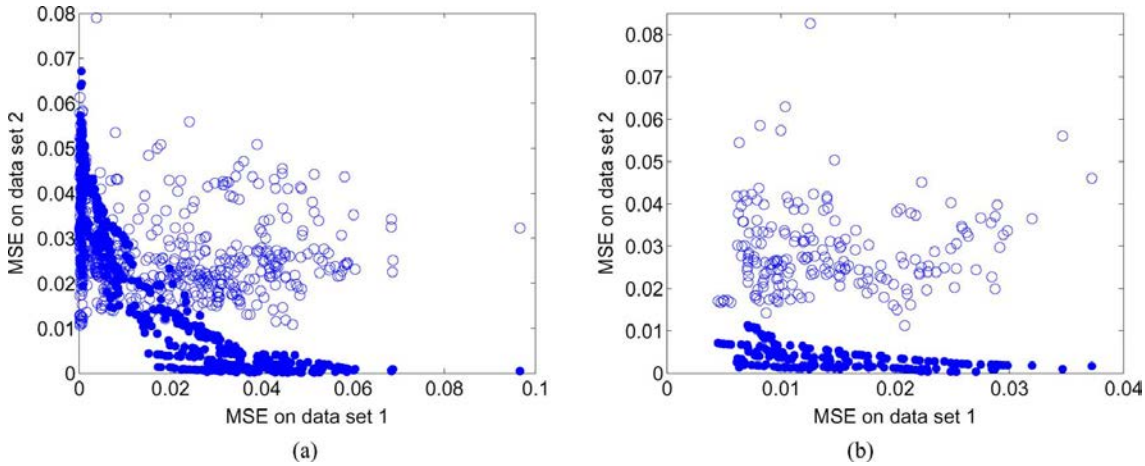

图 22　使用 Abbass 方法实现的非支配性解决方案：乳腺癌数据。(a)终身学习在两个数据集之间切换。(b)适用于数据组合的终身学习

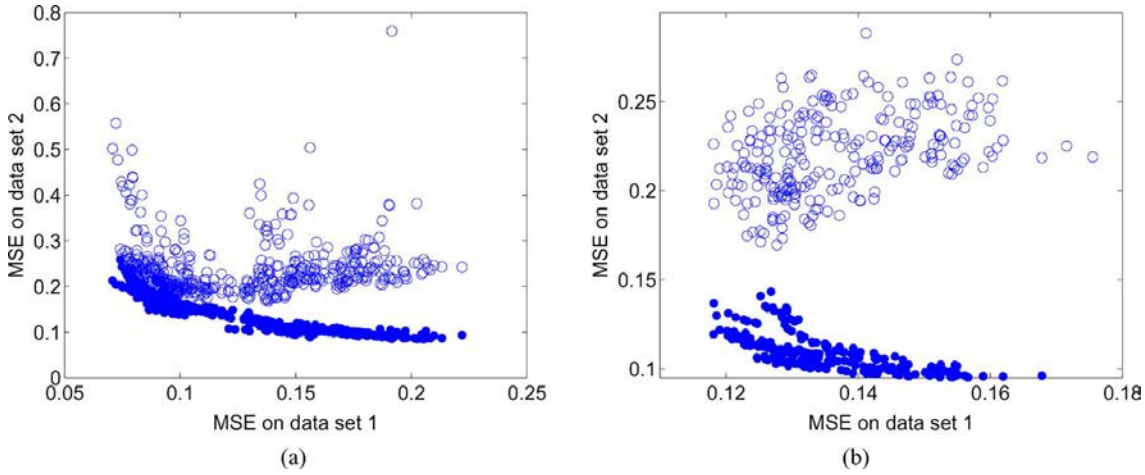

图 23　使用 Abbass 方法实现的非支配性解决方案：糖尿病数据。(a)终身学习在两个数据集之间切换。(b)适用于数据组合的终身学习

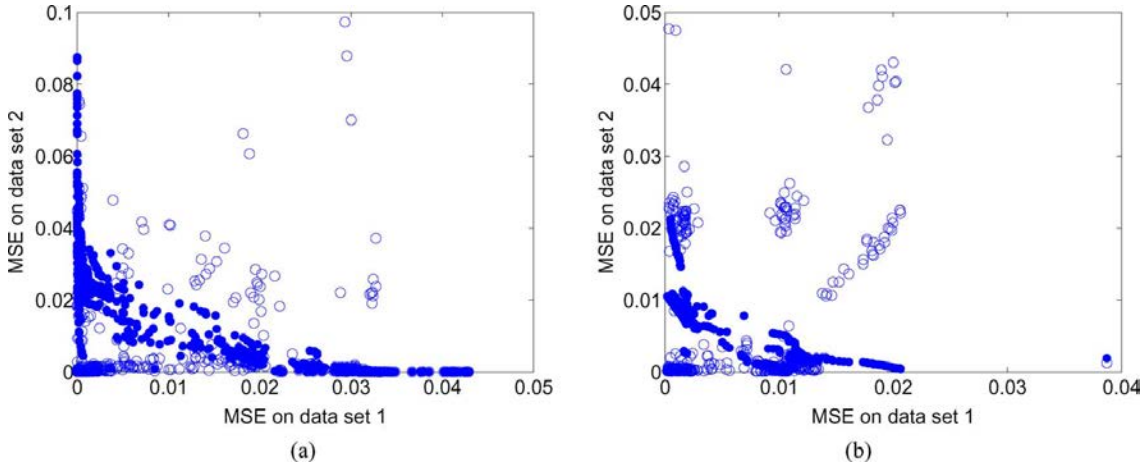

图 24　使用 Abbass 方法实现的非支配性解决方案：鸢尾花数据。(a)终身学习在两个数据集之间切换。(b)适用于数据组合的终身学习

糖尿病数据和虹膜数据得到了非常相似的结果，这些数据绘制在图 23 和 24 中。同样，从 Pareto 最优解中选择合适的集合成员也很困难。

使用 Chandra 和 Yao 的方法对三个基准问题进行的仿真结果如图 25、26 和 27 所示。再次，训练和测试数据的结果由圆点和圆圈表示。从图中我们发现，无论多样性如何，所实现的 Pareto 最优网络模型倾向于过度拟合数据，特别是糖尿病数据和虹膜数据。

最后，我们来看一下 Jin 等人的方法，它通过生成具有不同复杂性的神经网络来确保集合成员的多样性。结果如图 28、29 和 30 所示。从图中我们可以看到，在所有三个例子中，当神经网络的复杂度适当低时，测试数据上的 MSE 受到很好的约束。正如前一节所述，使用 NPG 可以估计与数据匹配的所需复杂性。通过选择这些网络作为集合成员，我们可以拥有一个神经网络集合，其成员既精确又多样。网络的多样性由神经网络复杂性的差异所保证。

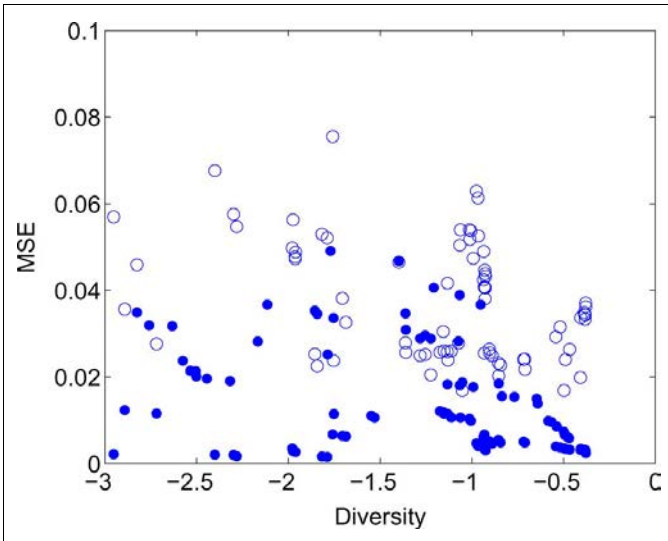

图 25　使用 Chandra 和 Yao 的方法获得的非支配性解决方案：乳腺癌数据

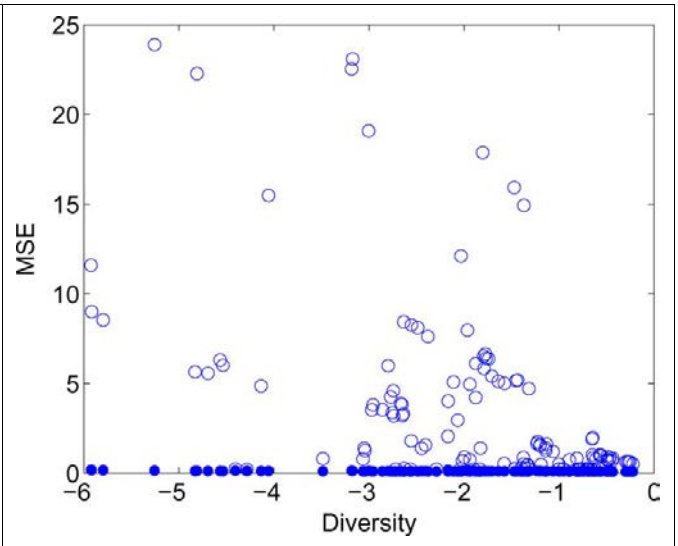

图 26　使用 Chandra 和 Yao 的方法获得的非支配性解决方案：糖尿病数据

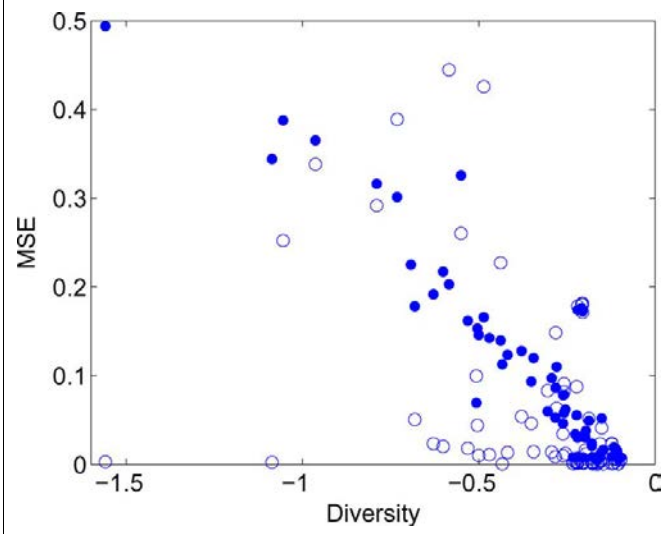

图 27　使用 Chandra 和 Yao 的方法获得的非支配性解决方案：鸢尾花数据

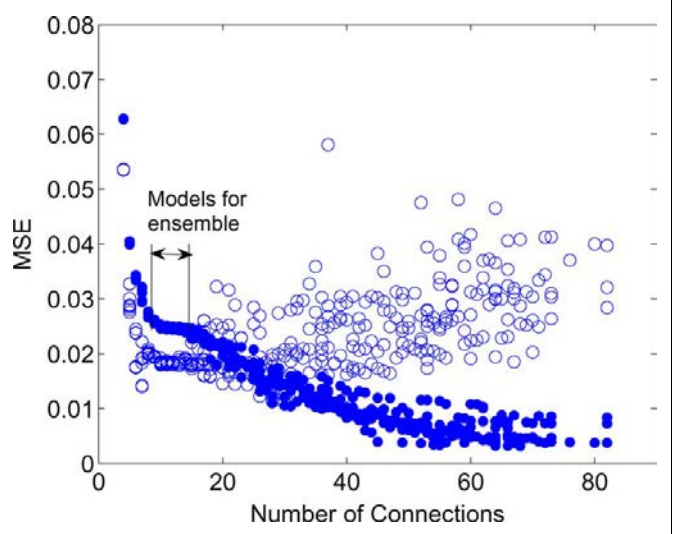

图 28　使用 Jin 等人的方法获得的非支配性解决方案：乳腺癌数据

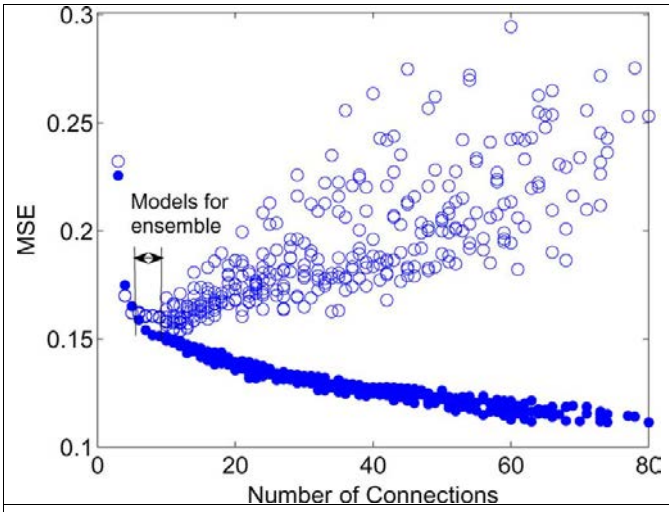

图 29　使用 Jin 等人的方法获得的非支配性解决方案：糖尿病数据

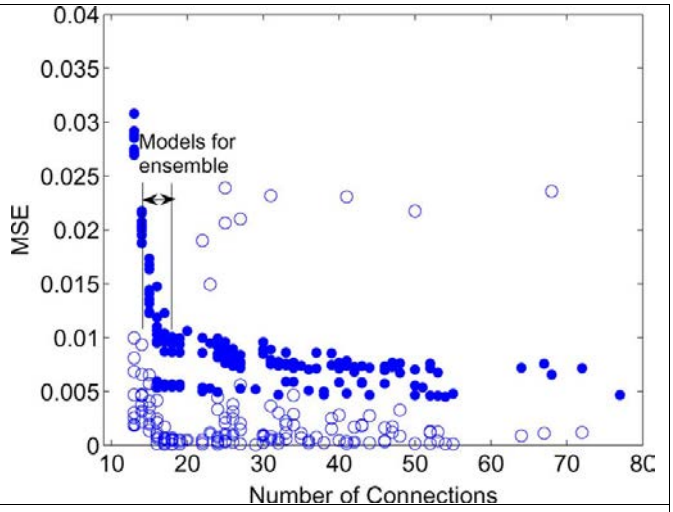

图 30　使用 Jin 等人的方法获得的非支配性解决方案：鸢尾花数据

比较三种基于 Pareto 的方法来进行集合生成，我们得出结论，从实现的 Pareto 最优解中选择集合成员来构建 Abbass 以及 Chandra 和 Yao 的方法中的集合是不直接的。为了确保集体的未见数据的良好性能，必须采用其他方法，例如交叉验证。相反，当使用 Jin 等人的方法产生 Paretooptimal 解时，识别可用作集合成员的神经网络相当容易。另一个重要的问题是，在 Abbass 和 Chandra 和 Yao 的方法中，从整个 Pareto 正面来看，十个独立运行所得到的解决方案是非常不同的，这意味着这两个方法对随机影响非常敏感。与此相反，当 Jin 等人的方法复杂度较低时，十次独立运行的结果非常稳定。

VII.　总结与展望

基于 Pareto 的机器学习方法为研究机器学习问题提供了一种新的观点。通过基于 Pareto 的优化，我们可以更深入地了解机器学习的不同方面，从而开发新的学习算法。由于演化算法在基于 Pareto 的多目标优化中的成功应用，使得基于 Pareto 的方法更具吸引力。

本文提供了一个最新的，但不一定完整的审查基于 Pareto 的多目标学习算法的现有研究。我们通过三个基准问题来说明如何使用基于 Pareto 的多目标方法来解决机器学习中的重要问题，例如生成可解释模型，用于泛化的模型选择和集成生成。我们表明，没有任何输入的最简单的 Pareto 最优模型选择接近训练数据的均值，而选择一个或两个最重要特征的简单 Paretooptimal 模型捕获数据中的基本知识。此外，我们凭经验证明，通过分析性能和复杂性方面的 Pareto 最优解，以及学习性能 w.r.t.模型的复杂性，我们能够选择最可能在不可见数据上表现出良好性能的模型。最后，我们比较三种基于 Pareto 的方法来生成神经系

统，并指出通过折衷准确性和复杂性的方法可以提供可靠的结果。

许多问题仍有待解决，并且在基于 Pareto 的多目标机器学习领域可能开辟新的领域。一个有趣的问题是基于 Pareto 的机器学习方法如何影响学习行为，例如学习曲线的属性[84]，[85]。在一般优化问题中经验性地披露，通过将多目标单目标问题转化为多目标问题可以减少局部最优解的数量[86]，[87]。如果我们能够证明在机器学习中发生了同样的情况，那么就更有说服力，认为基于 Pareto 的多目标学习能够提高学习成绩。

到目前为止讨论的大多数主题主要关注机器学习中的偏差-方差折衷。机器学习以及人类记忆系统中的另一个重要主题是塑性-稳定性折衷，也称为在线学习[88]，增量学习[89]或灾难性遗忘[90]。在[83]中已经做了一个初步尝试，使用基于帕雷托的方法解决灾难性遗忘问题。已经表明，多目标方法在减轻遗忘方面比单目标方法更有前途。Pareto 最优化的思想还可以扩展到一般网络的连通性和复杂性研究[91,92]，以及尖峰神经网络的结构和功能研究[93,94]。

参考文献

[1] E. Alpaydin, Introduction to Machine Learning. Cambridge, MA: MIT Press, 2004.

[2] T. M. Mitchell, Machine Learning. New York: McGraw-Hill, 1997.

[3] Z. Zoltan, Z. Kalmar, and C. Szepesvari, “Multi-criteria reinforcement learning,” in Proc. Int. Conf. Mach. Learn., 1998, pp. 197–205.

[4] Y. Jin, B. Sendhoff, and E. K¨orner, “Evolutionary multi-objective optimization for simultaneous generation of signal-type and symbol-type representations,” in Proc. 3rd Int. Conf. Evol. Multi-Criterion Optim. Lecture Notes in Computer Science, vol. 3410. New York: Springer- Verlag, 2005, pp. 752–766.

[5] F. Girosi, M. Jones, and T. Poggio, “Regularization theory and neural network architectures,” Neural Comput., vol. 7, pp. 219–269, 1995.

[6] Y. Jin, “Fuzzy modeling of high-dimensional systems: Complexity reduction and interpretability improvement,” IEEE Trans. Fuzzy Syst., vol. 8, no. 2, pp. 212–221, 2000.

[7] Y. Jin, W. V. Seelen, and B. Sendhoff, “On generating FC3 fuzzy rule systems from data using evolution strategies,” IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 29, no. 6, pp. 829–845, Dec. 1999.

[8] Y. Liu and X. Yao, “Ensemble learning via negative correlation,” Neural Netw., vol. 12, pp. 1399–1404, 1999.

[9] C. Cortes and V. Vapnik, “Support vector networks,” Mach. Learn., vol. 20, pp. 273–297, 1995.

[10] B. Olhausen and D. Field, “Sparse coding with an overcomplete basis set: A strategy employed by V1?,” Vis. Res., vol. 37, pp. 3311–3325, 1997.

[11] T. Fawcett, “ROC graphs: Notes and practical considerations for data mining researchers,” HP Labs., Palo Alto, CA, Tech. Rep. HPL-2003-4, 2003.

[12] J. Banfield and A. Raftery, “Model-based Gaussian and non-Gaussian clustering,” Biometrics,

vol. 49, pp. 803–821, 1993.
[13] M. H. C. Law, A. P. Topchy, and A. K. Jain, “Multiobjective data clustering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Piscataway, NJ: IEEE Press, 2004, vol. 2, pp. 424–430.
[14] I. Das and J. Dennis, “A closer look at drawbacks of minimizing weighted sum of objectives for Pareto set generation in multicriteria optimization problems,” Struct. Optim., vol. 14, no. 1, pp. 63–69, 1997.
[15] Y. Jin, M. Olhofer, and B. Sendhoff, “Dynamic weighted aggregation for evolutionary multi-objective optimization: Why does it work and how?,” in Proc. Genetic Evol. Comput. Conf. San Mateo, CA: Morgan Kaufmann, 2001, pp. 1042–1049.
[16] H. Abbass, “Speeding up back-propagation using multi-objective evolutionary algorithms,” Neural Comput., vol. 15, no. 11, pp. 2705–2726, 2003.
[17] R. de A. Teixeira, A. Braga, R. Takahashi, and R. Saldanha, “Improving generalization ofMLP with multi-objective optimization,” Neurocomputing, vol. 35, pp. 189–194, 2000.
[18] Y. Jin, Ed., Multi-Objective Machine Learning. New York: Springer- Verlag, 2006.
[19] G. Liu and V. Kadirkamanathan, “Learning with multi-objective criteria,” in Proc. Inst. Electr. Eng. Conf. Artif. Neural Netw., 1995, pp. 53– 58.
[20] Y. Matsuyama, “Harmonic competition: A self-organizing multiple criteria optimization,” IEEE Trans. Neural Netw., vol. 7, no. 3, pp. 652–668, May 1996.
[21] K. Kottathra and Y. Attikiouzel, “A novel multicriteria optimization algorithm for the structure determination of multilayer feedforward neural networks,” J. Netw. Comput. Appl., vol. 19, pp. 135–147, 1996.
[22] K. Deb, Multi-Objective Optimization Using Evolutionary Algorithms. Chichester, U.K.: Wiley, 2001.
[23] T. Furukawa, “Parameter identification with weightless regularization,” Int. J. Numer. Methods Eng., vol. 52, pp. 219–238, 2001.
[24] H. Abbass, “A memetic Pareto approach to artificial neural networks,” in Proc. 14th Aust. Joint Conf. Artif. Intell., 2001, pp. 1–12.
[25] Y. Jin, T. Okabe, and B. Sendhoff, “Evolutionary multi-objective approach to constructing neural network ensembles for regression,” in Applications of EvolutionaryMulti-Objective Optimization, C. Coello Coello, Ed. Singapore: World Scientific, 2004, pp. 653–672.
[26] Y. Jin, T. Okabe, and B. Sendhoff, “Neural network regularization and ensembling using multi-objective evolutionary algorithms,” in Proc. Congr. Evol. Comput. Piscataway, NJ: IEEE Press, Jun. 2004, pp. 1– 8.
[27] J. Fieldsend and S. Singh, “Pareto multi-objective non-linear regression modelling to aid CAPM analogous forecasting,” in Proc. Int. Joint Conf. Neural Netw., 2002, vol. 1, pp. 388–393.
[28] J. Fieldsend and S. Singh, “Pareto evolutionary neural networks,” IEEE Trans. Neural Netw., vol. 16, no. 2, pp. 338–354, Mar. 2005.
[29] N. Garcia-Pedrajas, C. Hervas-Martinez, and J. Munoz-Perez, “Multiobjective cooperative coevolution of artificial neural networks,” Neural Netw., vol. 15, no. 10, pp. 1255–1274, 2002.
[30] G. G. Yen and H. Lu, “Hierarchical rank density genetic algorithm for radial-basis function neural network design,” in Proc. Congr. Evol. Comput., 2002, vol. 1, pp. 25–30.
[31] T. Hatanaka, N. Kondo, and K. Uosaki, “Multi-objective structure selection for radial basis function networks based on genetic algorithm,” in Proc. Congr. Evol. Comput., 2003, pp. 1095–1100.

[32] J. Bi, “Multi-objective programming in SVMs,” in Proc. 20th Int. Conf. Mach. Learn., Washington, DC, 2003, pp. 35–42.
[33] C. Igel, “Multi-objective model selection for support vector machines,” in Evolution Multi-Criterion Optimization Lecture Notes in Computer Science, vol. 3410. New York: Springer-Verlag, 2005, pp. 534–546.
[34] H. Nakayama and T. Asada, “Support vector machines formulated as multi-objective linear programming,” in Proc. ICOTA, 2001, pp. 1171– 1178.
[35] D. Kim, “Structural riskminimization on decision trees using an evolutionary multiobjective optimization,” in Proc. Eur. Conf. Genetic Program. Lecture Notes in Computer Science, vol. 3003. New York: Springer- Verlag, 2004, pp. 338–348.
[36] E. Bernado-Manssilla and J. Garrell-Guii, “MOLeCS: Using multiobjective evolutionary algorithms for learning,” in Proc. EMO 2001 Lecture Notes in Computer Science, vol. 1993. New York: Springer- Verlag, pp. 696–710.
[37] S. Wiegand, C. Igel, and U. Handmann, “Evolutionary multiobjective optimization of neural networks fore face detection,” Int. J. Comput. Intell., vol. 4, no. 3, pp. 237–253, 2005.
[38] Y. Zhang and P. I. Rockett, “Evolving optimal feature extraction using multi-objective genetic programming: A methodology and preliminary study on edge detection,” in Proc. Genetic Evol. Comput. Conf., 2005, pp. 795–802.
[39] J. Teo and H. Abbass, “Automatic generation of controllers for embodied legged organisms:APareto evolutionary multi-objective approach,” Evol. Comput., vol. 12, no. 3, pp. 355–394, 2004.
[40] M. Luque, O. Cordon, and E. Herrera-Viedma, “A multi-objective genetic algorithm for learning linguistic persistent quries in text retrieval environments,” in Multi-Objective Machine Learning, Y. Jin, Ed. New York: Springer-Verlag, 2006, ch. 25, pp. 585–600.
[41] Y. Jin, Advanced Fuzzy Systems Design and Applications. Heidelberg, Germany: Physica Verlag, 2003.
[42] H. Ishibuchi, T. Murata, and I. T¨urksen, “Single-objective and twoobjective genetic algorithms for selecting linguistic rules for pattern recognition,” Fuzzy Sets Syst., vol. 89, pp. 135–150, 1997.
[43] A. Gomez-Skarleta, F. Jimenez, and J. Ibanez, “Pareto-optimality in fuzzy modeling,” in Proc. 6th Eur. Congr. Int. Tech. Soft Comput., 1997, pp. 694– 700.
[44] T. Suzuki, T. Furuhashi, S. Matsushima, and H. Tsutsui, “Efficient fuzzy modeling under multiple criteria by using genetic algorithms,” in Proc. IEEE Int. Conf. Syst., Man, Cybern., 1999, vol. 5, pp. 314–319.
[45] H. Ishibuchi, T. Nakashima, and T. Murata, “Three-objective geneticsbased machine learning for linguistic rule extraction,” Inf. Sci., vol. 136, no. 1–4, pp. 109–133, 2001.
[46] K. Tachibana and T. Furuhashi, “A structure identification method of submodels for hierarchical fuzzy modeling using the multiple objective genetic algorithm,” Int. J. Intell. Syst., vol. 17, pp. 495–513, 2001.
[47] F. Jimenez, G. Sanchez, A. F. Gomez-Skarmeta, H. Roubos, and R. Babuska, “Fuzzy modeling with multi-objective neuro-evolutionary algorithms,” in Proc. IEEE Int. Conf. Syst., Man, Cybern, 2002, vol. 3, pp. 253–258.
[48] H. Wang, S. Kwong, Y. Jin, W. Wei, and K. Man, “Agent-based evolutionary approach to interpretable rule-based knowledge extraction,” IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 35, no. 2, pp. 143–155, May 2005.

[49] H. Wang, S. Kwong, Y. Jin, W. Wei, and K. Man, “Multi-objective hierarchical genetic algorithm for interpretable fuzzy rule absed knowledge extraction,” Fuzzy Sets Syst., vol. 149, no. 1, pp. 149–186, 2005.
[50] U. Markowska-Kaczmar and P. Wnuk-Liplnski, “Rule extraction from neural networks with Pareto optimization,” in Artificial Intelligence and Soft Computing. Berlin, Germany: Springer-Verlag, 2004, pp. 450– 455.
[51] L. Hansen and P. Salammon, “Neural network ensembles,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, no. 10, pp. 993–1101, Oct. 1990.
[52] A. Chandra and X. Yao, “DIVACE: Diverse and accurate ensemble learning algorithm,” in Proc. 5th Int. Conf. Intell. Data Eng. Autom. Learn., 2004, pp. 619–625.
[53] D. Opitz and J. Shavlik, “Generating accurate and diverse members of a neural network ensemble,” in Advances in Neural Information Processing Systems 8. Cambridge, MA: MIT Press, 1996, pp. 535–541.
[54] G. Brown, J. Wyatt, R. Harris, and X. Yao, “Diversity creation methods: A survey and categorisation,” J. Inf. Fusion, vol. 6, no. 1, pp. 5–20, 2005.
[55] B. Rosen, “Ensemble learning using decorrelated neural networks,” Connection Sci., vol. 8, pp. 373–384, 1996.
[56] A. Chandra and X. Yao, “Evolving hybrid ensembles of learning machines for better generalisation,” Neurocomputing, vol. 69, pp. 686–700, 2006.
[57] H. Abbass, “Pareto neuro-evolution: Constructing ensemble of neural networks using multi-objective optimization,” in Proc. Congr. Evol. Comput, Dec. 2003, pp. 2074–2080.
[58] N. Garcia-Pedrajas, C. Hervas-Martinez, and J.Munoz-Perez, “COVNET: A cooperative coevolutionary model for evolving artificial neural networks,” IEEE Trans. Neural Netw., vol. 14, no. 3, pp. 575–596, May 2003.
[59] L. Kuncheva and C.Whitaker, “Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy,” Mach. Learn., vol. 51, no. 2, pp. 181–207, 2003.
[60] T. Hatanaka, N. Kondo, and K. Uosaki, “Multiobjective structure selection for RBF networks and its application to nonlinear system design,” in Multi-Objective Machine Learning, Y. Jin, Ed. New York: Springer- Verlag, 2006, ch. 21, pp. 491–505.
[61] H. Ishibuchi and T. Yamamoto, “Evolutionary multi-objective optimization for generating an ensemble of fuzzy rule-based classifiers,” in Proc. Genetic Evol. Comput. Conf. Lecture Notes in Computer Science, vol. 2723, 2003, pp. 1077–1088.
[62] M. Kupinski and M. Anastasio, “Multiobjective genetic optimization of diagnostic classifiers with implementations for generating receiver operating characteristic curves,” IEEE Trans. Med. Imag., vol. 18, no. 8, pp. 675–685, Aug. 1999.
[63] J. Horn and N. Nafpliotis, “Multiobjective optimization using the niched Pareto genetic algorithms,” IlliGAL, Univ. Illinois at Urbana-Champaign, Urbana-Champaign, Tech. Rep. 93005, 1993.
[64] L. Gr¨aning, Y. Jin, and B. Sendhoff, “Generalization improvement in multi-objective learning,” in Proc. Int. Joint Conf. Neural Netw., 2006, pp. 9893–9900.
[65] R. Everson and J. Fieldsend, “Multi-class ROC analysis from a multiobjective optimisation perspective,” Pattern Recognit. Lett., vol. 27, pp. 918–927, 2006.
[66] S. Park, D. Nam, and C. H. Park, “Design of a neural controller using multi-objective optimization for nonminimum phase systems,” in Proc. IEEE Int. Conf. Fuzzy Sets Syst., 1999, vol.

I, pp. 533–537.
[67] O. Cordon, F. Herrera, M. del-Jesus, and P. Villar, “A multi-objective genetic algorithm for feature selection and granularity learning in fuzzyrule based classification systems,” in Proc. Joint 9th IFSA World Congr. 20th NAFIPS Int. Conf., 2001, vol. 3, pp. 1253–1258.
[68] C. Emmanouilidis, A. Hunter, J. MacIntyre, and C. Cox, “Selecting features in neurofuzzy modelling by multiobjective genetic algorithms,” in Proc. Int. Joint Conf. Neural Netw., 1999, pp. 749–754.
[69] L. Oliveira, R. Sabourin, F. Bortolozzi, and C. Suen, “Feature seelction for ensembles: A hierarchical multi-objective genetic algorithm approach,” in Proc. 7th Int. Conf. Anal. Recognit., 2003, pp. 676–680.
[70] Y. Kim, W. Street, and F. Menczer, “Evolutionary model selection in unsupervised learning,” Intell. Data Anal., vol. 6, pp. 531–556, 2002.
[71] J. Handl and J. Knowles, “Exploiting the tradeoff—The benefits of multiple objectives in data clustering,” in Evolutionary Multi-Criterion Optimization Lecture Notes in Computer Science, vol. 3410. New York: Springer-Verlag, 2005, pp. 547–560.
[72] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a dataset via the gap statistics,” J. R. Stat. Soc. B, vol. 63, pp. 411– 423, 2001.
[73] X. Yao, “Evolving artificial neural networks,” Proc. IEEE, vol. 87, no. 9, pp. 1423–1447, Sep. 1999.
[74] C. Igel and M. H¨usken, “Empirical evaluation of the improved Rprop learning algorithm,” Neurocomputing, vol. 55, no. C, pp. 105–123, 2003.
[75] M. Riedmiller and H. Braun, “A direct adaptive method for faster backpropgation learning: The RPROP algorithm,” in Proc. IEEE Int. Conf. Neural Netw., 1993, vol. 1, pp. 586–591.
[76] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan, “A fast elitist nondominated sorting genetic algorithm for multi-objective optimization: NSGA-II,” in Proc. Parallel Problem Solving Nat., 2000, vol. VI, pp. 849– 858.
[77] L. Prechelt, “PROBEN1—A set of neural network benchmark problems and benchmarking rules,” Fakult¨at Inf., Univ. Karlsruhe, Karlsruhe, Germany, Tech. Rep., 1994.
[78] S. Thrun, “Extracting rules from artificial neural networks with distributed representation,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 1994, pp. 505–512.
[79] S. Kamruzzaman and M. Islam, “Extraction of symbolic rules from artificial neural networks,” Trans. Eng., Comput. Technol., vol. 10, pp. 271– 277, 2005.
[80] R. Setiono, “Generating concise and accurate classification rules for breast cancer disgnosis,” Artif. Intell. Med., vol. 18, pp. 205–219, 2000.
[81] K. Burnham and D. Anderson, Model Selection and Multimodel Inference. New York: Springer-Verlag, 2002.
[82] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. New York: Springer-Verlag, 2001.
[83] Y. Jin and B. Sendhoff, “Alleviating catastrophic forgetting via multiobjective learning,” in Proc. Int. Joint Conf. Neural Netw., 2006, pp. 6367–6374.
[84] S. Amari, “A universal theorem on learning curves,” Neural Netw., vol. 6, no. 2, pp. 161–166, 1993.
[85] T. Fine and S. Mukherjee, “Parameter convergence and learning curves for neural networks,” Neural Comput., vol. 11, pp. 747–769, 1999.

[86] S. Louis and G. Rawlins, “Pareto optimality, GA-easiness and deception,” in Proc. Int. Conf. Genetic Algorithms, 1993, pp. 118–123.
[87] J. Knowles, R. Watson, and D. Corne, “Reducing local optima in singleobjective problems bymulti-objectivization,” in Proc. 1st Int. Conf. Evol. Multi-Criterion Optim. Lecture Notes in Computer Science, vol. 1993. New York: Springer-Verlag, 2001, pp. 269–283.
[88] V. de Angulo and C. Torras, “On-line learning with minimal degradation in feedforward networks,” IEEE Trans. Neural Netw., vol. 6, no. 3, pp. 657– 668, May 1995.
[89] S.Wan and L. Banta, “Parameter incremental learning algorithm for neural networks,” IEEE Trans. Neural Netw., vol. 17, no. 6, pp. 1424–1438, Nov. 2006.
[90] M. Frean and A. Robins, “Catastrophic forgetting in simple networks: An analysis of the pseudo-rehearsal solution,” Netw.: Comput. Neural Syst., vol. 10, pp. 227–236, 1999.
[91] R. Adams, L. Calcraft, andN.Davey, “Connectivity in real and evolved associative memories,” in Proc. Brain Inspired Cognit. Syst., 2006, pp. 153– 159.
[92] O. Sporns andG. Tononi, “Classes of network connectivity and dynamics,” Complexity, vol. 7, pp. 28–38, 2002.
[93] Y. Jin, R. Wen, and B. Sendhoff, “Evolutionary multi-objective optimization of spiking neural networks,” in Proc. Int. Conf. Artif. Neural Netw. (ICANN) Part I, Lecture Notes in Computer Science, vol. 4668. New York: Springer-Verlag, 2007, pp. 370–379.
[94] W. Maass and C. Bishop, Pulsed Neural Networks. Cambridge, MA: MIT Press, 1999.