[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/30 (2006.01)
G06F 15/18 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200710193876.2

[43] 公开日 2008 年 5 月 28 日

[11] 公开号 CN 101187944A

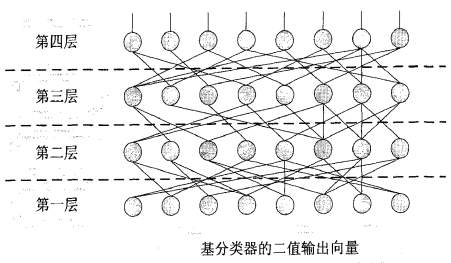
[22] 申请日 2007.11.30
[21] 申请号 200710193876.2
[71] 申请人 中国科学院合肥物质科学研究院
地址 230031 安徽省合肥市西郊董铺 1130 号信箱智能所
[72] 发明人 黄德双 章 军
[74] 专利代理机构 安徽省合肥新安专利代理有限责任公司
代理人 赵晓薇
权利要求书 3 页 说明书 7 页 附图 2 页

[54] 发明名称

基于小生境粒子群优化算法的分类器集成的多层选择方法

[57] 摘要

本方法涉及基于小生境粒子群优化算法的分类器集成的多层选择方法，属于模式识别领域。本方法包括：(1)对构成集成的基分类器构建二值输出向量；(2)制定集成选择标准；(3)选定集成的选择表示方法；(4)使用小生境粒子群优化算法构建多层集成选择模型。首先，本方法对训练出来的参加集成的基分类器使用验证集和训练集构建二值输出向量，然后通过小生境 PSO 算法，选择上一层中不同的二值输出向量参加本层中不同的集成，从而构建了一个多层选择模型。其优点是：通过构建一个多层选择模型，能够充分利用基分类器的有用信息，因而能获得比原始的多分类器集成更高的精度和可靠性。



1、 一种基于小生境粒子群优化算法的分类器集成的多层选择方法，其特征是:

（1）对基分类器构建一个二值输出向量(Oracle Output)

设共有 L 个分类器参加集成，$D=\{D_1, \ldots. D_L\}$，$y_i=[y_{i1}, \ldots., y_{iL}]^T$为 L 个分类器对输入样本 x_i的识别输出。每个分类器均为标签分类器，其中 y_{ij}是指第 j 个分类器对第 i 个训练样本的识别输出。每个分类器输出中能够正确识别该样本，则对应的 $y_{ij}=0$，不能正确识别，则对应 $y_{ij}=1$，通过这种方式，构成了一个分类器多层选择模型的基础。

（2）制定集成选择标准

在集成选择中，选择标准是非常重要的。在实际应用中，最常用的选择标准即基于验证集的大多数投票错误(majority voting error, MVE)。在本专利中，所有的输出都是基于前一层中的二值输出向量的。对于本专利所提出的多层集成选择模型，如果提供足够多和差异度大的二值输出(oracle output)，则很可能会出现过拟合现象。因此，本专利提出了一种新的选择标准，使用的是平均大多数投票错误率 $\overline{MVE}$，$\overline{MVE}$ 的计算不仅根据验证集，而且根据训练集，它主要希望训练集和验证之间的误差尽可能是接近。因而能最大可能减小系统的泛化错误。

（3）集成的选择表示方法

在多层集成选择模型中，每一个集成可以通过发现一个选择向量(pruning vector)来表示对应的选择结果，假定 $H_i=[h_{i1}, \ldots\ldots, h_{iL}]^T$为第 i 层的选择向量，其中 h_{ij}表示对应的第 j 个分类器或二值输出是否

被加入到集成中，当 $h_{ij}=1$ 时表示对应的输出或分类器被包括在集成中，如果为零则将该分类器或输出排除在集成之外。因此，在多层集成选择模型中，对每一个集成需要发现这样一个合适的二值向量以选择对应的分类器或输出。在这里，使用小生境粒子群优化算法来发现这样的二值向量。

（4）小生境粒子群优化算法

使用一种多子群并行小生境PSO算法，算法能够有效地模拟一个自然的生态系统，其中不同的子群之间互相竞争。在竞争以后，胜利者将继续探索原来的区域(小生境)，而失败者将被迫探索其他的区域(小生境)。在算法中，具体的实现是通过小生境识别技术来判断探索中的两点是否位于同一小生境，其他子群中粒子进入胜利者所拥有的小生境，则通过罚函数方式，迫使该粒子离开其他子群所拥有的小生境。罚函数如下式所示：

$$eval(x_n^i)=\begin{cases} f(x_n^i) & if \quad hill\_valley(x_n^i, x_k^{best})=1 \\ f(x_n^i)-p(x_n^i) & otherwise \end{cases} \qquad (1)$$

上式中，x^i_n表示第n个子群中第i个粒子位置，$f(x^i_n)$为该粒子原始的适应度值。x_k^{best}是第k个子群中最优粒子的位置，k不等于n，$p(x^i_n)$是罚函数，在这里可以取一个较大的常数。

原始的粒子群优化是一种基于实值的算法，在本专利中，Kennedy等所提出的一种简单的二进制版本的PSO算法被采用。在该算法中，

粒子的速度被用来作为一个概率来决定对应的位为零还是一。整个数学描述如下所示:

$$S(v)=\frac{1}{1+\exp^{-v}} \tag{2}$$

$$x_{id}=\begin{cases}1 & if \quad rand(*)\le S(v)\\ 0 & otherwise\end{cases} \tag{3}$$

在上面的公式中，v表示对应粒子的速度，rand(*)是一个均匀分布在[0,1]之间的随机数。粒子的状态只有两种情况1或0，而速度V与某个概率阈值相关，速度越大，则粒子位置取1的可能性越大，反之越小。

（5）多层集成选择模型的构建

在多层模型中，每一层中将有多种不同的选择集成。底层是由基分类器所产生的二值输出向量(Oracle output vector)所组成。整个多层选择模型以第一层中的二值输出向量为基础，每层将依赖上一层而选择出多种不同的、性能优秀的集成，而新的集成的结果又组成一个二值输出向量,这样就有机地构成了一个新颖的集成多层选择模型。每一层的多个集成选择问题被看成是一个多模优化问题，小生境 PSO 算法被用来在每一层中发现多组最优的选择集成，从而在实践中最终实现了分类器多层选择模型。

基于小生境粒子群优化算法的分类器集成的多层选择方法

所属领域 本发明涉及一种利用小生境粒子群优化算法来构建分类器集成的多层选择模型，从而能够提高系统集成的性能，本发明可以广泛应用于所有需要模式识别的场合。

背景技术 近年来，分类器集成方法在机器学习和数据挖掘领域吸引了越来越多的研究者的重视。研究者们通常认为，一个集成的性能在很大程度上依赖基分类器的两个方面：一个是基分类器的精确度，另一个是基分类器之间的差异度。一个具有差异度大并且准确的基分类器的集成将肯定具有比单个分类器更好的性能[Kittler, J., M. Hatef, R. Duin, and J. Matas, On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. 20(3): p. 226-239.]。近十年来，一些研究者已经开发了一些建立差异度大的基分类器方法，其中著名的有随机子空间方法[Ho, T.K., The Random Space Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. 20(8): p. 832-844.], Bagging 和 Boosting 方法[Breiman, L., Bagging Predictors. Machine Learning, 1996. 24(2): p. 123-140.; Kuncheva, L.I., M. Skurichina, and R.P.W. Duin, An experimental Study on Diversity for Bagging and Boosting with Linear Classifiers. Information Fusion, 2002. 3(2): p. 245-258.]。随机子空间方法是随机选择不同的特征子集来做训练集以训练参加集成的基分类器。而 Bagging 算法则是随机选择不同的样本来做训练集训练基分类器。Boosting 算法也是使用不同的样本来构建训练集，但和 Bagging 算法不同的是，在 Boosting 算法中，

难以识别的样本将比容易识别的样本有更多的机会构成训练集以训练参加集成的下一个基分类器。当差异度大的基分类器形成以后，如何选择不同的基分类器参加集成将变得非常关键。一般认为好的组合方法不仅应该具有精度高的基分类器，同时应该具有较高的差异度。通常，这种分类器的选择称为集成选择技术(ensemble pruning)[Margineantu, D. D. and T. G. Dietterich. Pruning Adaptive Boosting in 14th International Conference on Machine Learning. 1997]。使用分类器集成选择技术主要有两方面的原因：首先，选择部分基分类器集成其性能可能比全部参加集成要好[Zhou, Z. H., J.X. Wu, and W. Tang, Ensembling Neural Networks: Many Could be Better than All. Artificial Intelligence, 2002. 137(1-2): p. 239-263.]。另一方面，普通集成方法需要大量的存储器以存储基分类器信息，而集成选择技术则能够在很大程度上减少这种存储资源的消耗，这对实际应该是非常有帮助的，能够进一步提高集成的效率。

当前，如何选择合适的基分类器参加集成并没有一定的标准。个体基分类器的性能和他们之间的差异度都不能被用来直接选择合适的基分类器，以参加集成[Zeuobi, G. and P. Cunningham. Using Diversity in Preparing Ensembles of Classifiers based on Different Feature Subsets to Minimise Generalisation Error. in Proceedings of the 12th European Conference on Machine Learning. 2001]。一般使用基于验证集的大多数投票错误率(majority voting err, MVE)作为基分类器的选择标准。当一个集成具有较小的 MVE 时，该集成被认为具有较好的性能。通常所使用的贪婪方法(Greedy Algorithm)和聚类方法通常只能发现局部最优解。

而遗传算法(GA)作为一种全局优化算法，经常被用来选择合适的基分类器。然而，即使是使用遗传算法通常也只能得到一个集成结果，一些具有有用或互补信息的基分类器将可能从集成中丢弃[Mukherjee, S. and T.L. Fine. Ensemble Pruning Algorithms for Accelerated Training. in IEEE International Conference on Neural Networks. 1996.]。本发明通过对上述不足之处的详细分析，提出了创新的解决方案，提出的方法能尽可能充分利用每一个基分类器的有用信息，并且最终集成的识别率也有大幅度的提高。

发明内容 本发明的目的是克服现有多分类器集成选择技术的不足，提供了一种新颖的多层选择集成的方法。

本发明的目的是这样实现的:

（1）对每个基分类器构建一个二值输出向量(Oracle Output)

设共有 L 个分类器参加集成，$D=\{D_1, \ldots. D_L\}$，$y_i=[y_{i1}, \ldots., y_{iL}]^T$为 L 个分类器对输入样本 x_i的识别输出。每个分类器均为标签分类器，其中 y_{ij}是指第 j 个分类器对第 i 个训练样本的识别输出。每个分类器输出中能够正确识别该样本，则对应的 $y_{ij}=0$，不能正确识别，则对应 $y_{ij}=1$，通过这种方式，构成了一个分类器多层选择模型的基础。

（2）制定集成选择标准

在集成选择中，选择标准是非常重要的。在实际应用中，最常用的选择标准即基于验证集的大多数投票错误(majority voting error, MVE)。在本文中，所有的输出都是基于前一层中的二值输出向量的。假定 L 个分类器对输入样本 x_i进行投票，则相应的大多数投票输出可以按

下式得到:

$$y_i^{MV}=\begin{cases}1, & if\sum_{j=1}^{L}y_{ij}\geq\frac{L}{2}\\ 0, & if\sum_{j=1}^{L}y_{ij}<\frac{L}{2}\end{cases} \quad (1)$$

则一般基于验证集的MVE可按下式计算:

$$MVE=\frac{1}{N}\sum_{i=1}^{N}y_i^{MV} \quad (2)$$

在公式(2)中，N为验证集中的样本数，然而，对于本文所提出的多层集成选择模型，如果提供足够多和差异度大的二值输出(oracle output)，则很可能会出现过拟合现象。因此，本文提出了一种新的选择标准，平均大多数投票错误率$\overline{MVE}$，$\overline{MVE}$的计算不仅根据验证集，而且根据训练集，它主要希望训练集和验证之间的误差尽可能是接近。$\overline{MVE}$可按下式定义:

$$\overline{MVE}=\frac{1}{2}(MVE^{T}+MVE^{V}) \quad (3)$$

(3)集成的选择表示方法

在多层集成选择模型中，每一个集成可以通过发现一个选择向量(pruning vector)来表示对应的选择结果，假定$H_i=[h_{i1},\ldots\ldots,h_{iL}]^T$为第i层的选择向量，其中$h_{ij}$表示对应的第j个分类器或二值输出是否被加入到集成中，当$h_{ij}=1$时表示对应的输出或分类器被包括在集成中，如果为零则将该分类器或输出排除在集成之外。因此，在多层集成选择模型中，对每一个集成需要发现这样一个合适的二值向量以选择对应的分类器或输出。在这里，使用小生境粒子群优化算法来发现这样的二值向量。

(4)小生境粒子群优化算法

使用一种多子群并行小生境 PSO 算法，算法能够有效地模拟一个自然的生态系统，其中不同的子群之间互相竞争。在竞争以后，胜利者将继续探索原来的区域(小生境)，而失败者将被迫探索其他的区域(小生境)。在算法中，具体的实现是通过小生境识别技术来判断探索中的两点是否位于同一小生境，其他子群中粒子进入胜利者所拥有的小生境，则通过罚函数方式，迫使该粒子离开其他子群所拥有的小生境。罚函数如下式所示:

$$eval(x_n^i)=\begin{cases} f(x_n^i) & if \quad hill\_valley(x_n^i, x_k^{best})=1 \\ f(x_n^i)-p(x_n^i) & otherwise \end{cases} \tag{4}$$

上式中，x^i_n表示第 n 个子群中第 i 个粒子位置，$f(x^i_n)$为该粒子原始的适应度值。x_k^{best}是第 k 个子群中最优粒子的位置，k 不等于 n，$p(x^i_n)$是罚函数，在这里可以取一个较大的常数。

原始的粒子群优化是一种基于实值的算法，在本专利中，Kennedy 等所提出的一种简单的二进制版本的 PSO 算法被采用。在该算法中，粒子的速度被用来作为一个概率来决定对应的位为零还是一。整个数学描述如下所示:

$$S(v)=\frac{1}{1+\exp^{-v}} \tag{5}$$

$$x_{id}=\begin{cases} 1 & if \quad rand(*)\le S(v) \\ 0 & otherwise \end{cases} \tag{6}$$

在上面的公式中，v表示对应粒子的速度，rand(*)是一个均匀分布在[0,1]之间的随机数。粒子的状态只有两种情况1或0，而速度V与某个概率阈值相关，速度越大，则粒子位置取1的可能性越大，反之越小。

（5）多层集成选择模型的构建

在多层模型中，每一层中将有多种不同的选择集成。底层是由基分类器所产生的二值输出向量(Oracle output vector)所组成。整个多层选择模型以第一层中的二值输出向量为基础，每层将依赖上一层而选择出多种不同的、性能优秀的集成，而新的集成的结果又组成一个二值输出向量，这样就有机地构成了一个新颖的集成多层选择模型。每一层的多个集成选择问题被看成是一个多模优化问题，小生境 PSO 算法被用来在每一层中发现多组最优的选择集成，从而在实践中最终实现了分类器多层选择模型。

本发明的创新之处在于:

1、 使用了一种全新的自适应小生境技术。

在小生境技术中首次提出一种显式的探索信息交换机制，并根据这种机制首先实现了一种自适应小生境粒子群优化算法。为了实现探索信息交换，算法使用了小生境识别技术，并从受限优化问题中引入了罚函数技术，将多模态优化问题视为一种受限优化问题，由于这种探索信息交换，算法能自动识别所探索的小生境，因此它不需要调整任何额外的小生境参数，甚至包括小生境半径，即算法不需要依赖任何先验知识，因而能适应任何实际的应用。

2、 构建了一种集成多层选择模型。

普通的集成分类器选择方法，通常只能获得一个最优集成，在这种情况下，一些具有有用信息的基分类器将可能丢失。而在多层模型中，每一层中将有多种不同的选择集成。底层是由基分类器所产生的二值输

出向量所组成。整个多层选择模型以第一层中的二值输出向量为基础，每层将依赖上一层而选择出多种不同的、性能优秀的集成，而新的集成的结果又组成一个二值输出向量，这样就有机地构成了一个新颖的集成多层选择模型。能进一步的提高集成的性能。

附图说明 下面结合附图对本发明作进一步的说明。

图 1 每一层中不同选择的示意图。

图 2 多层选择模型示意图。

图 3 小生境识别技术示意图。

图 4 小生境 PSO 算法效果示意图。

图 5 UCI 数据集比较图。

图 6 UCI 数据集比较图。

具体实施方式 本专利首先使用一些传统的能够增加基分类器差异度的集成算法，如 Bagging 和 Adaboost 算法产生基分类器，在差异度不同的基分类器形成以后，再根据基分类器的输出结果为每个基分类器构建一个二值输出向量，然后根据所构建的二值输出向量，使用本专利中所提出的新颖的集成选择标准，并运用多子群粒子群优化算法来对每一层中发现多个不同的最优集成选择，在每一层中的输出构成了下一层中的输入，从而最终构建了一个集成多层选择模型，而该模型最终的输出结果作为整个系统的识别输出。本专利采用了目前国际上标准的公共机器学习数据集: UCI 机器学习数据集。在测试中基分类器由 Bagging 和 Adaboost 算法分别形成，最后的测试结果显示了该模型能够有效地提高整个分类器集成的识别性能。

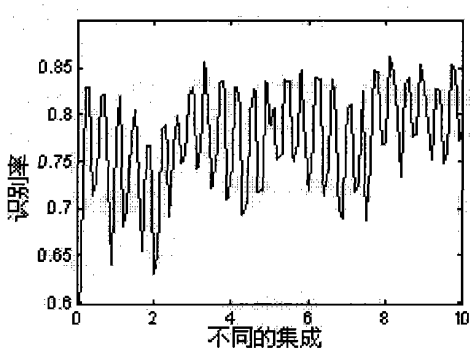

0.85
0.8
0.75
0.7
0.65
0.6
0
2
4
6
8
10
不同的集成

图 1

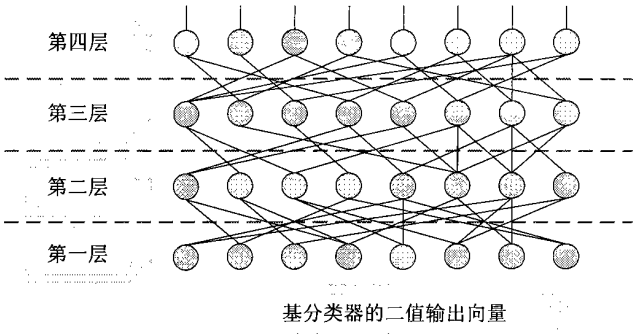

第四层
第三层
第二层
第一层
基分类器的二值输出向量

图 2

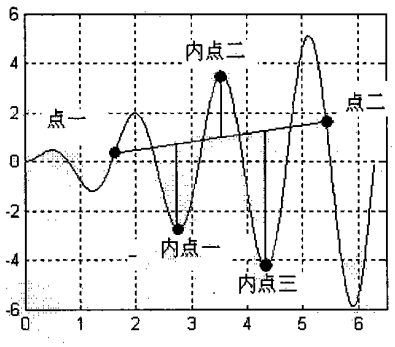

6
4
2
0
-2
-4
-6
0
1
2
3
4
5
6
内点二
点一
点二
内点一
内点三

图 3

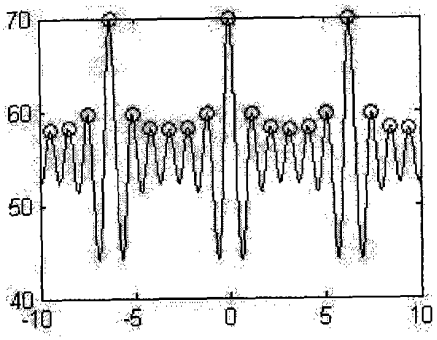

70
60
50
40
-10
-5
0
5
10

图 4

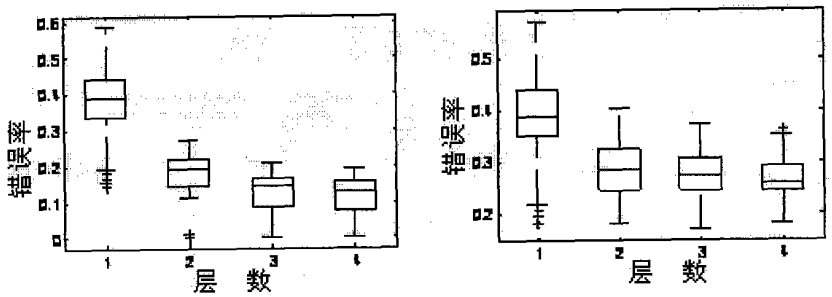

错误率
层 数
错误率
层 数

图 5

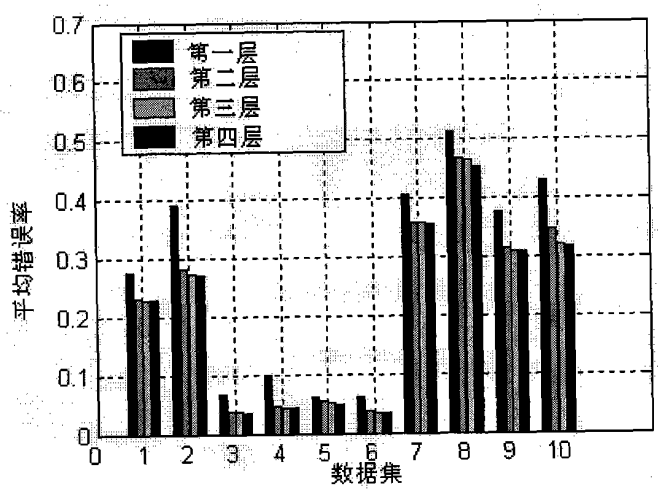

第一层
第二层
第三层
第四层
平均错误率
数据集
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9 10

图 6