基于特征提取的选择性神经网络集成方法*)

朱帮助

(北京航空航天大学经济管理学院 北京 100083) (五邑大学系统科学与技术研究所 江门 529020)

摘 要 针对现有神经网络集成研究方法在输入属性、集成方式和集成形式上的不足,提出一种基于特征提取的选择性神经网络集成模型——NSNNEIPCABag。该模型通过 Bagging 算法产生若干训练子集;利用改进的主成分分析(IPCA)提取主成分作为输入来训练个体网络;采用 IPCA 从所有个体网络中选择出部分线性无关的个体网络;采用神经网络对选择出来的个体网络进行非线性集成。为检验该模型的有效性,将其用于时间序列预测,结果表明本文提出的方法的泛化能力优于流行的其它集成方法。

关键词 集成学习,神经网络集成,非线性集成,选择性集成,特征提取

Feature Extraction-based Selective Neural Network Ensemble Method

ZHU Bang-Zhu

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100083)

(Institute of System Science and Engineering, Wuyi University, Jiangmen 529020)

Abstract In order to solve the problems of the existing research methods in input attribute, ensemble fashion and ensemble form, a novel selective & nonlinear neural network ensemble method, i. e. NSNNEIPCABag, is proposed in this paper. In this model, some different training subsets are first generated by Bagging algorithm. Then the feature extraction technique, improved principal component analysis (IPCA), is used to extract the data feature for training individual networks, and to select the appropriate number of ensemble members from the available networks. Finally, the selected members are aggregated into an ensemble model by neural network. For illustration and testing purposes, the proposed ensemble model is applied to time series forecasting with the favor results obtained, which shows that the generalization ability of the proposed method can be superior to those of other neural network ensembles.

Keywords Ensemble learning, Neural network ensemble, Nonlinear ensemble, Selective ensemble, Feature extraction

1 前言

1990 年,Hansen 和 Salamon[1]开创性地提出了神经网络集成(Neural Network Ensemble)方法。神经网络集成通过训练多个神经网络并将其结果进行合成,可以显著地提高神经网络系统的泛化能力。1995 年,Krogh 和 Vedelsby[2]指出,神经网络集成的泛化误差等于集成中个体网络的平均泛化误差和平均差异度之差。因此,要增强神经网络集成的泛化能力,一方面应尽可能提高个体网络的泛化能力,另一方面应尽可能地增大集成中各网络之间的差异。

目前,常见的神经网络集成主要是通过扰动训练数据来获得差异度较大的个体网络,例如 Boosting、Bagging 算法。然而,现有的神经网络集成存在着三个主要问题:(1)多数研究将所有的输入属性都用来训练个体网络。实际上,这些输入属性在对未知数据进行预测或分类时并非都是有用的,或者并非都是十分有用的。从特征选择的角度,预测或分类在很大程度上是由最能体现回归或类别特征的少数关键属性决定的[3]。由此,通过特征选择出对预测或分类最有用的属性作为神经网络的输入可能取得更好的结果。(2)多数研究将所有被训练的个体网络都进行集成。然而,"越多,越好"原则并不适用于任何情况,从中选择若干个体网络进行集成有可能取得更好的效果[4~6]。(3)多数研究采用简单平均[7,8]或加权平均[9]对个体网络输出进行线性集成。然而,现实中个体网络之间更多地表现为非线性关系[5],线性集成很难抓住这种非线性特征。

基于上述分析,本文将特征选择技术——改进的主成分分析(IPCA)引入神经网络集成的构建中,以相对较少的主成分去发现和保留绝大部分样本信息,并将这些主成分作为输入来创造和训练个体神经网络;在此基础上,采用 IPCA 法从所有被训练的个体网络中选择出部分线性无关的个体网络;随后,利用神经网络对选择出来的部分个体网络进行非线性集成,形成一种新的选择性神经网络集成模型——NSNNEIPCABag;最后,为检验该模型的有效性,将其应用于时间序列预测。结果表明,本文提出的方法的泛化能力优于流行的其它集成方法。

2 IPCA

特征提取的基本任务是如何从众多特征中提取出那些最有效的特征,即研究如何把高维特征空间压缩到低维特征空间,同时保留住绝大部分样本信息[3]。特征提取的方法很多,其中主成分分析(PCA)被认为是一种有效的约简方法。然而,传统的 PCA 只保留了各维指标间相互影响的信息,忽视了各维指标变异程度上的信息,因此,有必要加以改进。本文采用均值标准化方法[7,10]来改进传统的 PCA,该方法能够使

*)国家自然科学基金(70471074);广东省科技计划(2004B36001051)。朱帮助 讲师,博士生,主要研究方向:复杂经济社会系统建模与仿真。

得标准化后的相关矩阵不仅消除了指标量纲与数量级的影响，还体现出了各维指标变异程度上差异的影响。

设输入空间 R^p 中的 n 个样本数据向量 $x_k(k=1,2,\cdots,n)$，$x_k \in R^p$，构建 $n\times p$ 原始数据矩阵。

(1)将原始数据进行均值标准化处理：

$$x_{ij}^{'}=x_{ij}/\bar{x}_j$$

其中 $\bar{x}_j=\sum_{i=1}^{n}x_{ij}/n$。

(2)标准化变换后，计算相关矩阵 $R=(r_{ij})_{p\times p}$，其中

$$r_{ij}=\frac{\mathrm{var}(x_j)}{\bar{x}_j^2},i=j$$

其中 $\mathrm{var}(x_j)=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$。

(3)令 $|R-\lambda I|=0$，求解相关矩阵 R 的特征根 λ_j 且使得 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$，其对应的特征向量为 $u_1\geqslant u_2\geqslant\cdots u_p$。

(4)引入新的变量：

$$\begin{cases}F_{i1}=u_1^{'}x_i=u_{11}x_{i1}+u_{12}x_{i2}+\cdots+u_{1p}x_{ip}\\F_{i2}=u_2^{'}x_i=u_{21}x_{i1}+u_{22}x_{i2}+\cdots+u_{2p}x_{ip}\\F_{ip}=u_p^{'}x_i=u_{p1}x_{i1}+u_{p2}x_{i2}+\cdots+u_{pp}x_{ip}\end{cases}$$

其中，$F_{ij}(j=1,2,\cdots,p)$为第 j 主成分。

(5)计算累计方差贡献率：

$$E=\sum_{k=1}^{m}\lambda_k/\sum_{k=1}^{p}\lambda_k$$

当 $E\geqslant\theta$(θ 通常为 85%)时 m 的最小整数作为 m 的值，即主成分的个数为 m。

3 NSNNEIPCABag 算法

假设训练集 S 由属性 $A_1,A_2,\cdots,A_{n+1}$ 组成，其中 $A_1,A_2,\cdots,A_n$ 为输入属性，而 A_{n+1} 为期望输出属性。目前存在的大多数集成方法在训练个体网络时是将所有的输入属性都用上，即希望建立从 $A_1,A_2,\cdots,A_n$ 到 A_{n+1} 的映射。由于预测或分类在很大程度上是由最能体现回归或类别特征的少数关键属性决定的[3]，因此只需要提取出一些对预测或分类最有用的属性作为神经网络的输入。

NSNNEIPCABag 算法的具体做法如下：给定训练集，通过 Bagging 算法获得 m 个训练子集，并在各训练子集上用 IPCA 提取出对预测或分类起重要作用的属性作为输入训练出个体网络；然后，利用 IPCA 选择出部分线性无关的个体网络；最后，利用神经网络对选择出的部分个体网络进行非线性集成。NNEIPCABag 算法的基本过程如图 1 所示。

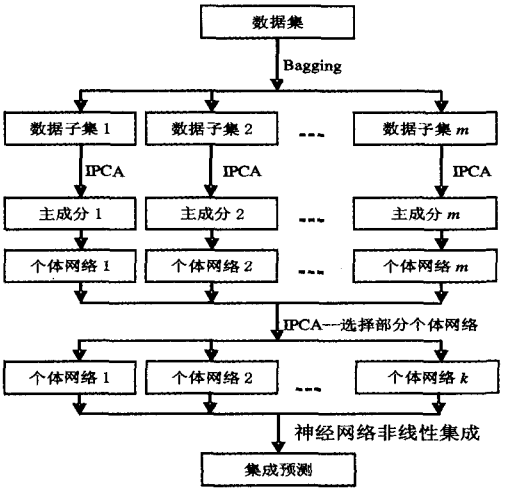

图 1 NSNNEIPCABag 算法的流程

4 试验测试

为验证上述 NSNNEIPCABag 方法的有效性，将其用于广东省江门市 GDP 增长率的预测。本文采用 BP 神经网络作为个体网络和非线性集成工具，并采用 MATLAB 7.01 编程实现。

4.1 数据描述

本文选择江门市 1982～2004 年的 8 个指标构成 GDP 预测指标体系[7]：江门市 GDP(x_1)、国家 GDP(x_2)、广东省 GDP(x_3)、外贸出口总额(x_4)、财政支出(x_5)、社会消费品零售总额(x_6)、固定资产投资(x_7)和实际利用外资(x_8)。

考虑到诸年数据的可比性，将原始数据环比化处理：

$$x_i(t)^{'}=\frac{x_i(t)/w(t)}{x_i(t-1)/w(t-1)}-1,i=1,2,\cdots,8;t=1982,1983,\cdots,2004$$

式中，$x_i(t)^{'}$为指标 i 的环比值；$x_i(t)$为指标 i 的原始数据值；w 为累积物价指数；t 和 $t-1$ 分别代表当前年度和上一年度。原始数据限于篇幅不便给出，具体数据可参考江门市历年统计年鉴或与作者联系。

采用单步预测，即用上一年度的经济数据作为输入，下一年的 GDP 增长率作为输出。设 1983～1999 年的经济数据及其对应的 1984～2000 年 GDP 增长率为训练样本集；设 2000～2003 年的经济数据及其对应的 2001～2004 年 GDP 增长率为测试样本集。此外，本文还利用 2004 年的经济数据来预测 2005 年的 GDP 增长率。

进一步地，利用 Bagging 算法形成 5 个训练子集，大小分别为 17,16,15,14 和 13。将 IPCA 对各训练子集进行特征提取，分别得到三个主成分 F_1，F_2 和 F_3，其累积贡献分别为 92.34%，91.65%，92.01%，92.31% 和 92.07%。因此，在一定程度上，主成分可以被用来作为个体网络输入。此外，本文还利用 IPCA 对 2000～2004 年的经济数据进行了特征提取。

4.2 试验结果

利用 MATLAB 7.01 神经网络工具箱，创造 5 个个体网络，分别用各子集主成分训练个体网络。个体网络均为 BP 神经网络，网络结构为 3-5-1。使用 BP 算法，最大学习次数 5000 次，学习率为 0.001，训练误差为 0.00001，网络初始权值设置为[−1,1]间的随机数。当所有的训练结果都达到要求，即认为个体网络被训练好了。

当所有个体网络被训练完成后，利用 IPCA 选择出 3 个线性无关的个体网络，并采用 BP 神经网络进行非线性集成。为了便于比较，本文还利用相应的经济数据，构建出了以下四种集成模型：(1)NNEBag——不进行特征提取，直接将所有属性数据训练个体网络，并采用简单平均集成所有个体网络；(2)NNEIPCABag——进行特征提取，利用特征数据训练个体网络，并采用简单平均集成所有个体网络；(3)SNNEIPCABag——进行特征提取，利用特征数据训练个体网络，并采用简单平均集成部分个体网络；(4)NSNNEIPCABag——进行特征提取，利用特征数据训练隔热

(下转第 172 页)

3.3 实验比较

我们用某地区1985年到1998年以来有关林业方面的统计数据作为数据集来验证算法2的有效性。为了简单起见，我们只选择了其中的部分属性：人数、平均工资、投资额、原木产量、等外材产量、小规格树产量和总销量。{人数，平均工资，投资额，原木产量，等外材产量，小规格树产量}为条件属性集合，总销量为决策属性。我们还对表中的数据进行了离散化处理：采用相对比较的方法，将数据转化为前一年的增幅表，然后通过离散编码的方式，再将其转化为Boolean表。相对于前一年，数据增加的记为1，数据减少的记为0。这样，我们可以较方便地用KNN算法来分析总销量的变化与其他因素之间的关系。同时，为了提高分类结果的可靠性，在这里我们采用交叉验证的方法对数据集进行分类。

为便于理解，在这里我们只给出几个简单的实验结果。按传统的KNN算法进行分类的一些结果如下：当K值为3、训练数据为9个(1985年到1993年数据)、测试数据为5个(1994年到1998年数据)时，我们得到的分类准确度为60%。当K值为3、训练数据为9个(1990年到1998年数据)、测试数据为5个(1985年到1989年数据)时，我们得到的分类准确度为40%。

采用算法2对与上面相同的数据集进行分类的结果如下：先采用粗糙集理论对布尔表进行属性约简，得到的属性约简结果为：{人数、投资额、原木产量、等外材产量}。删除不相容以及冗余的属性值{平均工资，小规格树产量}后，再用KNN算法对其进行分类。当K的值为3、训练数据为9个(1985年到1993年数据)、测试数据为5个(1994年到1998年数据)时，我们得到的分类准确度为80%。当K值为3、训练数据为9个(1990年到1998年数据)、测试数据为5个(1985年到1989年数据)时，我们得到的分类准确度为60%。

通过上述实验比较，我们可以看出，当样本数据的特征属性较多以及样本的容量较大的时，用算法2得到的结果比用传统KNN分类算法要好。

结束语 KNN算法是目前数据挖掘领域一种比较常见的分类算法，由于其实现的简单性，在许多领域有着广泛的应用。由于KNN算法不需要构建分类模型，所有的有关分类的计算都是在对新样本数据分类的时候进行的，因此当样本数据的特征属性的数量较多、样本的容量较大时，分类的时间代价很大，分类的效果不是很好，这会对实际应用产生很大的影响。本文提出的算法2是在对新的样本数据进行分类之前，先对它们用算法1进行属性约简，删除那些对样本的决策影响很小或者是根本没有影响的冗余属性，从而使KNN分类能够比较顺利地进行；提高了分类的效率，扩大了KNN算法的应用范围，同时保证了分类的准确性。但是，在现实中，样本数据集中可能会包括一些噪声样本数据，同时也可能会包括一些属性缺失的样本数据，这将会对KNN分类造成很大的影响。如何用粗糙集理论对这些噪声样本数据以及属性缺失的样本数据进行处理，从而可以使KNN分类能够顺利地进行，将是我们下一步要研究的问题。

参考文献

1 Pawlak Z D. Rough set theory and its application to data analysis [J]. Cybernetics and Systems, 1998,29(9):611～6685

2 Pawlak Z D. Rough set theory and its applications[J]. Journal of Telecommunications and Information Technology,2002(3)

3 陈安，陈宁，周龙骤，等. 数据挖掘技术及应用. 北京：科学出版社，2006

4 胡学刚，郭光亚. 一种基于粗糙集的朴素贝叶斯分类算法. 合肥工业大学学报，2006(2)

5 张冬玲. 基于粗糙集理论的属性约简算法的实现. 计算机应用，2006(2)

(上接第133页)

体网络，并采用非线性集成所有个体网络。此外，本文选择均方差(RMSE)作为比较标准，具体比较结果如表1所示。

表1 预测结果比较

集成方法	RMSE	名次	2005
NNEBag	0.0087	5	0.11180
SNNEBag	0.0071	3	0.10947
NNEIPCABag	0.0081	4	0.11235
SNNEIPCABag	0.0053	2	0.11249
NSNNEIPCABag	0.0052	1	0.11245

从试验结果可以发现：(1)尽管只有17个训练样本，所有集成方法的测试结果仍都非常接近于实际值；(2)集成预测效果明显好于任何一个个体网络；(3)总体上，本文提出的NSNNEIPCABag方法优于其它所有集成方法；(4)江门市2005年GDP的增长率稍高于11%；(5) NSNNEIPCABag方法适用于经济预测。

结束语 本文试图提供一种新的基于特征提取的选择性神经网络集成方法——NSNNEIPCABag，该方法综合集成了Bagging算法、IPCA特征提取方法、选择性集成技术以及非线性集成技术，充分发挥它们的协同优势，在一定程度上解决了目前神经网络集成研究中存在的主要问题。虽然仅用了17个学习样本，本文提出的NSNNEIPCABag方法的泛化能力总体上比其它集成方法更好。本文的研究为小样本经济预测提供了一种新的有效途径。

参考文献

1 Hansen L K, Salamon P. Neural Network Ensembles [J]. IEEE Trans Pattern Analysis and Machine Intelligence, 1990, 12(10): 993～1001

2 Krogh A, Vedelsby J. Neural Network Ensembles: Cross Validation, and Active Learning [J]. Advances in Neural Information Processing System, 1995(7):231～238

3 凌锦江，陈兆乾，周至华. 基于特征选择的神经网络集成方法[J]. 复旦学报(自然科学版)[J]，2004,43(5):685～688

4 Zhou Z H, Wu J X, Tang W. Ensemble Neural Networks: Many Could Be Batter Than All [J]. Artificial Intelligence, 2002, 137 (1-2):239～263

5 Yu L A, Wang S Y, Lai K K. A Novel Nonlinear Ensemble Forecasting Model Incorporating GLAR and ANN for Foreign Exchange Rates [J]. Computer & Operation Research, 2005 (32):2523～2541

6 Yu L A, Wang S Y, Lai K K, et al. A Bias-variance-complexity Trade-off Framework for Complex System Modeling [J]. Lecture Notes in Computer Science, 2006, 3980: 518～527

7 Lin J, Zhu B Z. Improved Principal Component Analysis and Neural Network Ensemble Based Economic Forecasting [J]. Lecture Notes in Computer Science, 2006, 4113: 135～145

8 林健，彭敏晶. 基于神经网络集成的GDP预测模型[J]. 管理学报，2005,2(4):434～436

9 Lin J, Zhu B Z. Neural Network Ensemble Based on Forecasting Effective Measure and Its Application [J]. Journal of Computational Information Systems, 2005, 1 (4):781～787

10 程其云，王有元，陈伟根. 基于改进主成分分析的短期负荷预测方法[J]. 电网技术，2005,29(3):64～67