基于Bagging的概率神经网络集成分类算法

蒋 芸 陈 娜 明利特 周泽寻 谢国城 陈 珊
(西北师范大学计算机科学与工程学院 兰州 730070)

摘 要 目前的神经网络较多集中在以BP算法为基础的BP神经网络上。针对BP神经网络的不足,在分析研究概率神经网络和机器学习的基础上,结合集成学习的思想,提出了基于Bagging的概率神经网络集成分类算法。理论分析和实验结果都表明,提出的算法能够有效地降低分类误差,提高分类准确率,具有较好的泛化能力以及较快的执行速度,能够取得比传统的BP神经网络分类方法更好和更稳定的分类结果。

关键词 分类,BP神经网络,概率神经网络,集成学习,Bagging
中图法分类号 TP183 **文献标识码** A

Bagging-based Probabilistic Neural Network Ensemble Classification Algorithm

JIANG Yun CHEN Na MING Li-te ZHOU Ze-xun XIE Guo-cheng CHEN Shan
(College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China)

Abstract Neural networks classification algorithm now more concentrates on the BP algorithm which is the representative of the neural networks. Considering the disadvantage of BP neural network, based on the analysis of probabilistic neural networks and machine learning, and combining with the idea of ensemble learning, we proposed a new classification algorithm which is probabilistic neural networks ensemble based on Bagging. Theoretical analysis and experimental results show that the proposed algorithm can effectively reduce the classification error and improve accuracy of classification. The proposed algorithm has good generalization ability and faster speed of execution than the traditional classification methods such as BP neural networks and it can achieve better and more stable classification result.

Keywords Classification, Back propagation neural network, Probabilistic neural networks, Ensemble learning, Bagging

1 概述

分类是数据挖掘、模式识别和机器学习中的一个主要研究领域。目前很多分类算法已经被提出[12],这些分类算法主要涉及决策树、关联规则、贝叶斯、K-邻近、遗传算法、粗糙集、支持向量机以及神经网络等方法[3]。

人工神经网络(Artificial Neural Network)[4]是20世纪80年代后期重新迅速发展起来的人工智能技术。神经网络由一组通过权值相互连接的神经元组成,网络通过调整权值来实现输入样本与其相应类别的对应,只要把数据输入到已经训练好的神经网络就可以直接得到分类结果。神经网络因对噪声数据具有很强的承受能力以及对未经训练的数据进行分类的能力而推动了它在数据挖掘分类方面的应用,BP(Back Propagation)算法作为一种经典有效的神经网络学习算法,在数据挖掘分类中被广泛研究。

神经网络作为分类技术中的重要方法之一[5],其优势在于:(1)神经网络可以任意精度逼近任意函数;(2)神经网络方法本身属于非线形模型,能够适应各种复杂的数据关系;(3)神经网络具备很强的学习能力,使它能够比很多分类算法更好地适应数据空间的变化;(4)神经网络借鉴人脑的物理结构和机理,能够模拟人脑的某些功能,具备"智能"的特点。基于神经网络的分类方法很多,基本是按照神经网络模型的不同而进行区分。用于数据分类常见的神经网络模型包括:BP神经网络、RBF神经网络、SOM神经网络、LVQ神经网络。目前神经网络分类算法研究较多集中在以BP为代表的神经网络上。但BP神经网络存在以下几个缺点:(1)网络结构比较复杂,需调节的参数较多,不容易确定网络的层数及每层的节点数,其隐层单元的选取没有确定性法则,需要根据经验反复试验得到;(2)网络的学习算法收敛速度慢,容易陷入局部极小值,在训练样本较大且要求精度较高时,网络的训练时间较长且常常不收敛;(3)网络扩充性能差,当网络的结构确定后,就难以适应新的环境,训练样本更改后则需要重新进行训练,网络的连接权值全部需要重新确定,相当于重新建立整个网络。

传统的概率神经网络(PNN)[6]是一种建立在径向基函数网络(RBF)基础之上的、简单的、应用范围广[7,8]的分类网络。与传统的BP网络比较,它有3大优点:(1)网络学习过程简单,训练速度快,其训练时间仅仅略大于读取数据的时

到稿日期:2012-07-22 返修日期:2012-10-16 本文受国家自然科学基金项目(61163036,61163039),甘肃省科技计划(甘肃省自然科学基金项目1010RJZA022,1107RJZA112),2012年度甘肃省高校基本科研业务费专项资金项目,甘肃省高校研究生导师项目(1201-16),西北师范大学第三期知识与创新工程科研骨干项目(nwnu-kjcxgc-03-67)资助。

蒋 芸(1970—),女,博士,副教授,硕士生导师,主要研究方向为数据挖掘、粗糙集理论及应用,E-mail:jiangyun@nwnu.edu.cn。

间，无需反复训练网络，因而速度大约比 BP 神经网络快 5 个数量级[6]；(2)网络的分类能力强，收敛性较好，且不存在不收敛或陷入局部极小值的情况，只要有足够多的训练样本，概率神经网络就能够保证获得贝叶斯准则下的最优解；(3)网络结构设计方便灵活，容错性及扩充性能好，允许增加或减少训练数据而无需重新进行长时间的训练。

针对 BP 神经网络的不足，在分析研究概率神经网络和机器学习的基础上，结合集成学习的思想，提出了基于 Bagging 的概率神经网络集成分类算法。该算法通过构造多个分类器来增强系统的泛化能力，能有效地利用集成学习技术来提高系统的分类性能。本文第 1 节介绍所用到的概率神经网络，并简单介绍机器学习中集成学习的相关知识；第 2 节具体介绍基于 Bagging 的概率神经网络集成分类算法；第 3 节通过实验对该算法进行性能测试并对实验结果进行分析；最后对本文的工作进行总结。

2 基本理论

2.1 贝叶斯决策理论[9]

贝叶斯分类方法是概率统计学中的一种决策方法，可描述为：假设有两种已知的分类模式 θ_A、θ_B，对于要判断的分类特征样本 $X=(x_1,x_2,\cdots,x_n)$，

若 $h_A l_A f_A(X)>h_B l_B f_B(X)$，则 $X\in\theta_A$；

若 $h_A l_A f_A(X)<h_B l_B f_B(X)$，则 $X\in\theta_B$。

上式中，h_A、h_B 为分类模式 θ_A、θ_B 的先验概率($h_A=N_A/N$，$H_B=N_B/N$)；N_A、N_B 为分类模式 θ_A、θ_B 的训练样本数；N 为训练样本总数；l_A 为将属于 θ_A 的分类特征样本 X 错误地划分到模式 θ_B 的损失；l_B 为将属于 θ_B 的分类特征样本 X 错误地划分到模式 θ_A 的损失；f_A、f_B 为分类模式 θ_A、θ_B 的概率密度函数(Probability Density Function)，通常概率密度函数不能精确地获得，只能根据现有的分类特征样本求其统计值。

针对贝叶斯分类的弱点，Parzen 在 1962 年提出了一种从已知随机样本中估计概率密度函数的方法，只要样本数目足够多，该方法所获得的函数就可以连续平滑地逼近真实概率密度函数。由 Parzen 方法得到的概率密度函数估计式如下：

$$f_A(X)=\frac{1}{(2\pi)^{P/2}\sigma^p}\frac{1}{m}\sum_{i=1}^{m}\exp\left[-\frac{(X-X_{ai})^{\mathrm{T}}(X-X_{ai})}{2\sigma^2}\right] \tag{1}$$

式中，X_{ai} 为分类模式 θ_A 的第 i 个训练向量；m 为分类模式 θ_A 的训练样本数量；p 为待分类样本及训练样本的维数；σ 为平滑参数，参数 σ 的不同取值对 PNN 分类器的误差估计有很大的影响，因此需要以实验的方法在参数 σ 的不同取值情况下根据所获得的分类准确率的比较来判定参数 σ 的最佳取值。

2.2 概率神经网络[6]

概率神经网络(Probabilistic Neural Networks，PNN)是 Specht 在 1990 首先提出的，它是一种基于贝叶斯分类规则与 Parzen 窗口的概率密度函数估计方法发展而来的并行算法。在实际应用中，尤其是在解决分类问题的应用中，它的优势在于用线性学习算法来完成非线性学习算法所做的工作，同时保持非线性算法的高精度等特征；这种网络对应的权值就是模式样本的分布，网络不需要训练，因而能够满足训练上实时处理的要求。

PNN 网络是由径向基函数网络发展而来的一种前馈神经网络，其理论依据是贝叶斯最小风险准则(即贝叶斯决策理论)，PNN 作为径向基网络的一种，适合于模式分类。PNN 的层次模型由输入层、模式层、求和层、输出层共 4 层组成，其基本结构如图 1 所示。

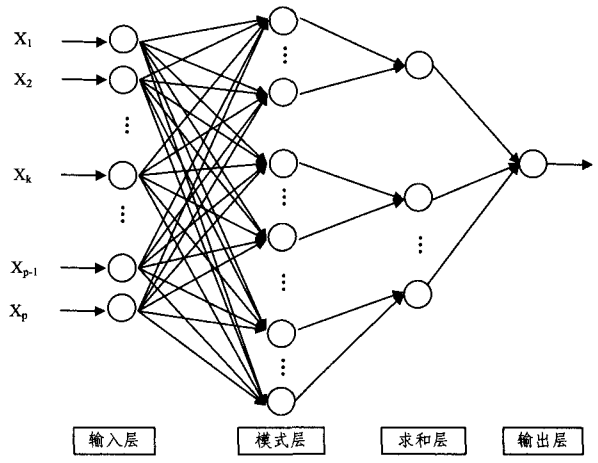

图 1 概率神经网络基本结构

输入层接受来自训练样本的值，将特征向量传递给网络，其神经元数目和训练样本的维数相等。

模式层计算输入特征向量与训练样本集中各个模式的匹配关系，模式层神经元的个数等于各个类别训练样本数之和，该层每个模式单元的输出为：

$$f(X,W_i)=\exp\left[-\frac{(W_i-X)^{\mathrm{T}}(W_i-X)}{2\sigma^2}\right] \tag{2}$$

式中，W_i 为输入层到模式层连接的权值；σ 为平滑参数，它对分类起着很重要的作用。

求和层将属于某类的概率累计，按式(2)计算，从而得到分类模式的估计概率密度函数。每一类只有一个求和层单元，求和层单元与只属于自己类的模式层单元相连接，而与模式层中的其他单元没有连接。因此求和层单元简单地将属于自己类的模式层单元的输出相加，而与属于其他类别的模式层单元的输出无关。求和层单元的输出与基于内核的各类密度的估计成比例，通过输出层的归一化处理，就能得到各类的概率估计。

输出层在各个分类模式的估计概率密度中选择一个具有最大后验概率密度的神经元作为整个系统的输出。输出层神经元是一种竞争神经元，每个神经元分别对应于一个数据类型即分类模式，输出层神经元个数等于训练样本数据的种类个数，它接收从求和层输出的各类概率密度模式，概率密度函数最大的那个神经元输出为 1，所对应的那一类即为待识别的样本类别，其他神经元的输出全为 0。

2.3 集成学习方法

机器学习(Machine Learning)[10]是研究计算机如何模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构并使之不断改善自身的性能。

在机器学习领域，Valiant[11]曾对概率近似正确(probable approximate correct)学习给出如下定义：对任意的概率分布 P 和任意 $0<\varepsilon,\delta\leqslant 1$，对所有概念 $c\in C$，如果存在算法 A，通过使用 c 产生的样本能在 $1/\varepsilon$，$1/\delta$ 和 N 的多项式时间内输出假设 h，使得 $P[\Delta(h,c)<\varepsilon]\geqslant 1-\delta$ 或 $P[\Delta(h,c)>\varepsilon]<\delta$，则称概念 c 是可概率近似正确学习的。进一步，如果学习结果的

正确率($1-\delta$)很大，则称 c 为强可学习的。如果正确率只比随机猜测略好(即略高于 0.5)，则称 c 为弱可学习的。Kearns 和 Valiant 还提出了弱学习算法与强学习算法的等价性问题，如果两者等价，那么在学习概念时，只需找到一个比随机猜测略好的弱学习算法，就可以将其提升为强学习算法，而不必直接去寻找通常情况下很难获得的强学习算法[12]。1990 年 Shapire 通过一个构造性方法，对该问题做出了肯定的证明[13]。

集成学习(Ensemble learning)[14]将数个简单的、精度比随机猜测略好的弱分类器以某种方式组合在一起对新样本进行分类，构成一个高精度的估计，从而有效克服过学习，提高分类准确率。

对于个体分类器的生成策略，常用的技术包括 Boosting 和 Bagging 等[15]。研究表明：这两种方法大大开发了弱学习的能力，在精确性和对不同领域数据的计算可行性等方面的表现都是比较突出的。

Boosting 算法首先对训练集样本赋一初始权重，随后对训练集采用学习分类器进行多次训练，对训练失败的样本赋以较大的权重，在后续学习中更重视对这些样本的学习，从而得到评价函数序列，最后根据某种策略进行综合。

Bagging(Bootstrap aggregating)[16]是一种把多个不同的个体学习器集成为一个学习器的集成学习方法，其理论基础是通过可重复取样得到不同的数据子集，使得在不同数据子集上训练得到的个体学习器具有较高的泛化性能及有较大的差异度。该算法用从原始训练集中随机抽取的若干样例来训练模型，得到的预测集合体在预测一个类标时，采取投票方式，取多个预测类标中出现次数最多的那个类标为该样例的最后类标。由于 Bagging 算法的个体分类器之间不存在强的依赖关系，因此算法可以并行。利用现有网络的分布式计算可以进一步提高算法的时间效率，并且 Bagging 总是可以改善学习器的性能[17]。

3 基于 Bagging 的概率神经网络集成分类算法 Bagging-PNN

Bagging 技术的主要思想是采用重采样技术，从原始数据集中分别独立随机地选取数据，并且将此过程独立进行多次，直到产生很多个独立的数据集。给定一个弱学习算法，可以通过该弱学习算法对产生的多个训练样本集进行学习，得出预测函数序列，将结果进行投票，得票最多的作为最后的结果。

为了提高概率神经网络的分类准确率和泛化能力，本文采用 Bagging 思想来生成集成所需的个体概率神经网络分类器。基于 Bagging 的概率神经网络集成分类算法具体如下：每次从大小为 n 的训练样本集中随机抽取 n 个样本，用概率神经网络分类算法进行训练，得到一个概率神经网络分类器，利用相同的方法生成多个概率神经网络分类器，训练之后可得到一个分类函数序列 $c_1(x), c_2(x), \cdots, c_T(x)$，最终的分类函数 $C(X)$ 对分类问题采用投票方式，得票最多的分类结果即为分类函数 $C(X)$ 的最终类别。

Bagging 方法通过重新选取训练集增加了分类器集成的差异度，从而提高了泛化能力。这样最后可以获得稳定的和更高准确率的结果。Bagging-PNN 具体流程如图 2 所示。

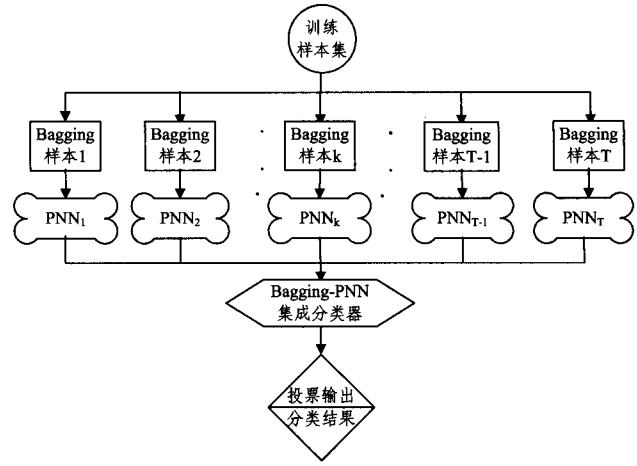

图 2 Bagging-PNN 分类算法流程

Bagging-PNN 算法具体描述为：

输入：D：训练样本集；T：创建个体分类器的数量；A：概率神经网络分类算法。

输出：集成分类器 C(X)

方法：

(1)for i=1 to T

(2)从训练样本集 D 中随机有放回的抽取 n 份样本，创建样本集 D_i；

(3)用样本集 D_i 和概率神经网络分类算法 A 训练，得到概率神经网络分类器 $c_i(x)$；

(4)end for

(5)输出集成分类器

$$C(X)=\operatorname{argmax}\sum_{i}^{T}c_i(x)$$

使用集成分类器 C(X)对未知样本 x 分类：

(1) 未知样本 x 分类时，每个分类器 $c_i(x)$得出一个分类结果，T 个分类器投票，得票数最多的类别即为未知样本 x 的分类结果；

(2) 投票输出分类结果

$$C(x)=\operatorname{argmax}\sum_{i}^{T}c_i(x)$$

4 实验结果及分析

本节主要通过实验研究集成分类算法 Bagging-PNN 的性能。为此，从 UCI 标准机器学习库[18]中选择了 7 个数据集，这些数据集来源于不同的领域：模式识别(iris，zoo，glass)、医学诊断(breast-WisConsin，lymphography)、控制应用(balance-scale)、统计分析(haberman)。其主要特征如表 1 所列。

表 1 UCI 标准数据集的特征描述

数据集	样本数目	连续属性	离散属性	类别数目
balance-scale	625	0	4	3
breast-Wisconsin	699	9	0	2
glass	214	9	0	6
haberman	306	0	3	2
iris	150	4	0	3
lymphography	148	0	18	4
zoo	101	16	1	7

采用 10 层交叉(10-fold Cross Validation)的方法在数据集上做分类测试，将数据集随机分成 10 份，依次选择其中 9 份作为训练集，余下的 1 份作为测试集，分别记录其分类准确率，最后求得平均值，即算法的分类准确率。为了验证算法的

性能，对比了概率神经网络分类算法 PNN，同时还与著名分类算法 C4.5 以及 Bagging-C4.5[19,20]进行了对比，4 种方法的详细测试结果如表 2 所列，其中第 1 列为数据集的名称，第 2 列为采用决策树 C4.5 算法得到的分类准确率，第 3 列为采用 Bagging-C4.5 算法得到的分类准确率，第 4 列为采用概率神经网络 PNN 算法得到的分类结果，第 5 列为采用本文提出的 Bagging-PNN 算法得到的分类准确率。实验计算机配置：CPU 为 Intel Pentium 4 3.06GHZ，内存 DDR 512MB，操作系统为 Windows XP SP3，测试工具为 MATLAB R2007b。

表 2　4 种方法的分类准确率比较(%)

数据集	C4.5	Bagging-C4.5	PNN	Bagging-PNN
balance-scale	77.82	82.33	89.36	89.6
breast-Wisconsin	94.72	95.77	95.65	95.97
glass	67.52	72.99	72.86	73.39
haberman	71.05	72.06	74.66	74.97
iris	95.2	94.87	96	96
lymphography	78.31	79.59	78.57	78.93
zoo	92.61	93.29	97	97.2

通过以上结果可以看出，本文提出的分类算法 Bagging-PNN 在 7 个数据集上都表现出了优越的分类性能。对比这 4 种分类算法，Bagging-PNN 分类算法在 6 个数据集上获得了最高的分类准确率，仅仅在 lymphography 数据集上的分类准确率略低于 Bagging-C4.5 算法 0.66%，但仍然比单独的 C4.5 以及 PNN 要好。虽然在 iris 数据集上 Bagging-PNN 和 PNN 分类准确率相当，但泛化能力上 Bagging-PNN 要强于 PNN，造成这种结果的因素很多，但有两个因素值得考虑：(1)分布密度 SPREAD 的值：当分布密度 SPREAD 的值接近于 0 时，它趋近最近邻分类器；当 SPREAD 的值较大时，它构成对几个训练样本的邻近分类器；当分布密度 SPREAD 的值接近于∞时，它趋近线性分类器。最佳分布密度 SPREAD 的值需要通过实验来获得。(2)分类器数量：使用 Bagging 时应该选用多少个体分类器才最合适这个问题，Breiman[16]指出个体分类器的数目应当随着分类种数的增多而增加。一些研究成果[21-23]表明，当集成中的个体学习器差异较大时，集成的效果较好。但如何获得差异较大的个体学习器、如何获得最佳分类器数量以及如何评价多个学习器之间的差异度，目前仍没有特别好的方法。如果能找到这样的方法，将极大地促进集成学习技术在应用领域的发展。本文在这些数据集上通过实验得出，一般集成 10～50 个就可以得到非常好的分类准确率，图 3 展示的是在 breast-Wisconsin 数据集上集成不同数量的分类器的分类误差率。可以看出，算法在集成 15 个个体分类器时获得了最低的分类误差，并且在集成数量达到 50 时仍然可以获得非常快的分类速度。表 3 给出的是在 breast-Wisconsin 数据集上集成不同数量的分类器时的训练以及测试用时，其中训练样本的数量为 620 个，测试样本的数量为 70 个。可以看出，Bagging-PNN 算法运行速度是非常快的。

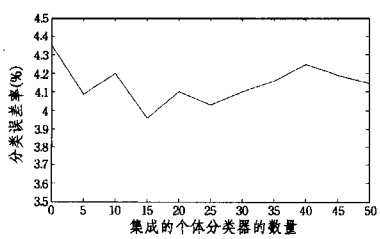

图 3　breast-Wisconsin 数据集上集成不同数量的分类器的分类误差率

表 3　breast-Wisconsin 数据集上集成不同数量的分类器所用的具体时间

时间 \ 分类器	5	10	15	20	25	30	35	40	45	50
620 个样本训练时间(S)	0.72	1.05	1.35	1.68	1.98	2.32	2.62	2.95	3.27	3.6
70 个样本测试时间(S)	0.39	0.61	0.89	1.13	1.39	1.61	1.88	2.12	2.36	2.57

结束语　集成学习方法通过构造多个学习器来增强学习系统的泛化能力，本文针对分类问题，提出了基于 Bagging 的 PNN 集成分类算法。采用这种方法能有效地利用集成学习技术来提高系统的分类性能。理论分析和实验结果都表明，本文提出的方法能够有效地降低分类误差，提高分类准确率，具有较好的泛化能力以及较快的执行速度，能够取得比传统的 BP 神经网络和决策树分类方法更好和更稳定的分类结果，是有效和实用的。

参 考 文 献

[1] Phyu T N. Survey of Classification Techniques in Data Mining [C]// Proceedings of the International MultiConference of Engineers and Computer Scientists. Hong Kong, 2009(1)

[2] 罗可，林睦纲，郗东妹. 数据挖掘中分类算法综述[J]. 计算机工程，2009，31(5)：3-5

[3] Han Jia-wei, Micheline K. 数据挖掘概念与技术(第 2 版)[M]. 范明，孟小峰，译. 北京：机械工业出版社，2007

[4] El-shafie A, Muklisin M, Najah Ali A, et al. Performance of artificial neural network and regression techniques for rainfall-runoff prediction[J]. International Journal of the Physical Sciences, 2011, 6(8): 1997-2003

[5] Selles M A, Schmid S R, Sánchez-Caballero S, et al. Theoretical Model of a Multi-Layered Polymer Coated Steel-Strip Ironing Process Using a Neural Network[C]// Materials Science Forum. Switzerland, 2012: 139-144

[6] Jiten P, Choi S-K. Classification approach for reliability-based topology optimization using probabilistic neural networks[J]. Structural and Multidisciplinary Optimization, 2012, 45(4): 529-543

[7] El-Emary I M, Ramakrishnan S. On the Application of Various Probabilistic Neural Networks in Solving Different Pattern Classification Problems[J]. World Applied Sciences Journal, 2008, 4(6): 772-780

[8] Al-Timemy A H, Al-Naima F M, Qaeeb N H. Probabilistic Neural Network for Breast Biopsy Classification[C]// International Conference on Developments in eSystems Engineering. 2009: 101-106

[9] Brian P, Stephen M S, Kennedy David N, et al. A Bayesian model of shape and appearance for subcortical brain segmentation[J]. NeuroImage, 2011, 56(3): 907-922

[10] Etherm A. 机器学习导论[M]. 范明，昝红英，牛常勇，译. 北京：机械工业出版社，2009

[11] Valiant L G. A theory of the learnable[J]. Communications of the ACM, 1984, 27(11): 1134-1142

[12] 毕华，梁洪力，王珏. 重采样方法与机器学习[J]. 计算机学报，2009，32(5)：862-877

[13] Schapire R E. The strength of weak learnability[J]. Machine

Learning,1990,5(2):197-227
[14] Robi P. Ensemble learning [EB/OL]. http://www.scholarpedia.org/article/Ensemble_learning,2012-12-11
[15] Buhlmann P. Bagging, Boosting and Ensemble Methods [M]. Berlin:Springer Berlin Heidelberg,2012
[16] Hamid P, Sajad P, Zahra R, et al. CDEBMTE: Creation of Diverse Ensemble Based on Manipulation of Training Example [J]. Pattern Recognition,2012,7329:197-206
[17] 周志华,陈世福. 神经网络集成[J]. 计算机学报,2002,25(1):1-8
[18] Frank A, Asuncion A. UCI Machine Learning Repository [DB/OL]. http://archive.ics.uci.edu/ml. Irvine,CA: University of California,School of Information and Computer Science,2010
[19] Martis R J,Acharya U R,Tan J H,et al. Application of empirical mode decomposition for automated detection of epilepsy using EEG signals[J]. Int J Neural Syst,2012,22(6):1250027
[20] Jayakishan M,Ram B C,Madhab P R,et al. Cascaded Factor Analysis and Wavelet Transform Method for Tumor Classification Using Gene Expression Data[J]. International Journal of Information Technology and Computer Science,2012,4:73-79
[21] Adhvaryu P S,Panchal Mahesh P. A Review on Diverse Ensemble Methods for Classification[J]. IOSR Journal of Computer Engineering,2012,1(4):27-32
[22] Ye Ren,Suganthan P N. Empirical comparison of bagging-based ensemble classifiers [C]// Information Fusion,2012 15th International Conference. 2012:917-924
[23] Tian Jin, Li Ming-qiang, Chen Fu-zan, et al. Coevolutionary learning of neural network ensemble for complex classification tasks[J]. Pattern Recognition,2012,45(4):1373-1385

(上接第 212 页)

2.7387%,4.4747%],最后对两个部分的预测数值进行累加运算,利用第 5 节中第(8)步中的误差进行分析,则得到的结果和误差分析如表 3 和图 7 所示。

表 3 支持度计数向量 SV 实际值与误差值对比表

时间	实际值	预测值	绝对误差	相对误差
1	6	6.00004	0	0%
2	6	5.89972	0.10028	1.67133%
3	4	3.93856	0.061444	1.5361%
4	1	1.15868	−0.15868	15.868%
5	2	1.82787	0.17213	8.6065%
6	7	6.70806	0.29194	4.17057%
7	3	2.90299	0.09701	3.23366%
8	3	2.92089	0.07911	2.637%
9	5	4.96322	0.03678	0.7356%
10	6	6.03413	−0.03413	0.56883%
11	7	6.82468	0.17532	2.50457%
12	4	3.84754	0.152458	3.81145%
13	5	4.81497	0.18503	3.7006%
14	4	3.78885	0.21115	5.27875%
15	6	5.80844	0.19156	3.19266%

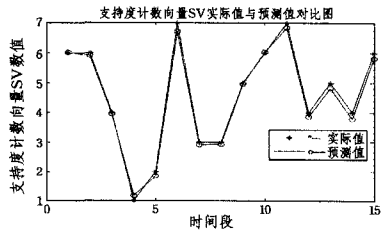

图 7 支持度 SV 实际值与预测值对比图

从表 3 和图 7 可看出预测的精准度比较高。

结束语 利用小波变换多分辨率的特点,把小波变换和灰色模型应用到动态关联规则元规则挖掘中,该方法首先利用小波变换对动态关联规则元规则支持度计数进行处理,这样不仅能保持支持度计数向量在时间上的变化特征,而且能保持在时间上的细节部分,其次利用灰色模型来进行预测,其预测结果从表 3 和图 7 中可以看出预测精准度比较高。然而本文也存在不足之处,如对于时间粒度的划分,现有的动态关联规则的时间段的划分都是等时间段划分,是静态的而动态关联规则的时间粒度的动态划分也是有着变化规律的,同样在动态关联规则元规则挖掘的趋势变化这一块也有待考虑。在文献[15]中论述的趋势度的概念是在支持度和置信度的基础上提出的,因此也可以考虑把趋势度添加进来进行规则的挖掘,这样可以在支持度和置信度的基础上去除无用的关联规则,以对实际应用数据库的关联规则的挖掘做出更好的指导作用。

参 考 文 献

[1] Liu Jin-feng,Rong gang. Mining dynamic association rules in databases[C]// Xi'an Proceedings of International Conference on Computtional Intelligences and Security 2005. Xi'an,2005:688-955
[2] 荣冈,刘进锋,顾海杰. 数据库中动态关联规则的挖掘[J]. 控制理论与应用,2007,24(1):129-133
[3] 沈斌,姚敏. 一种新的动态关联规则及其挖掘算法[J]. 控制与决策,2009,24(9):1310-1315
[4] 刘俊,张忠林,谢彦峰,等. 基于时间序列模型的关联规则元规则挖掘[J]. 计算机工程,2009,15(35):94-96
[5] 胡俊胡,玉清,肖忠卿. 基于小波变换的网络流量预测模型[J]. 计算机工程,2008,34(19):112-114,129
[6] 吴朝阳. 小波变换和 GM-ARMA 组合模型的股指预测[J]. 智能系统学报,2011,6(3):279-282
[7] 白翔宇,叶新铭,蒋海. 基于小波变换与自回归模型的网络流量预测[J]. 计算机科学,2007,34(7):47-54
[8] 佟伟明,李一军,单永正. 基于小波分析的时间序列数据挖掘[J]. 计算机工程,2008,34(1):26-28
[9] Zhang Yi, Wei Yong, Zhou Ping. Improved Approach of Gray Derivative in GM(1,1) Model [J]. The Journal of Grey System, 2006,116(10):160-162
[10] Sun Yan-na. Optimization of Grey Derivative in GM(1,1) Based on the Discrete Exponential Sequence[C]// Proceeding of the 2nd International Symposium on Information Processing(ISTP 2009). Huangshan,P. R. China,2009:313-315
[11] Yang Jiang-tian. Multivariable trend analysis using grey model for machinery condition monitoring[C]// Eleventh World Congress in Mechanism and Machine Science. 2004:2188-2191
[12] 张华,任若恩. 基于小波分解和残差 GM(1,1)—AR 的非平稳时间序列预测[J]. 系统工程理论与实践,2010,30(6):1016-1020
[13] 刘思峰,党耀国,方志耕,等. 灰色系统理论及其应用[M]. 北京:科学出版社,2004:142-146
[14] 刘思峰,邓聚龙. GM(1.1)模型的适用范围[J]. 系统工程理论及实践,2000,20(5):121-124
[15] 张忠林,曾庆飞,许凡. 动态关联规则的趋势度挖掘方法[J]. 计算机应用,2012,32(1):196-198