中国科学: 信息科学 2017 年 第 47 卷 第 11 期: 1445–1463
SCIENTIA SINICA Informationis
机器学习专题 · 论文

小时间序列动态完全 Bayesian 集成分类器研究

王双成[1*], 郑飞[1], 高瑞[2]

1. 上海立信会计金融学院信息管理学院, 上海 201620
2. 上海立信会计金融学院统计与数学学院, 上海 201620
* 通信作者. E-mail: wangsc@lixin.edu.cn

收稿日期: 2017–05–23; 接受日期: 2017–06–30; 网络出版日期: 2017–11–15
国家自然科学基金 (批准号: 61272209) 和上海市自然科学基金 (批准号: 15ZR1429700) 资助项目

摘要 提高连续属性小时间序列分类的可靠性重要且具有挑战性. 由于小时间序列所蕴含的信息不充分和时间序列数据具有时序依赖性, 使得优化分类器与数据的拟合程度非常困难, 而且非时间序列数据分类器的许多成熟技术都不具有实用性. 针对这种情况, 本文采用动态完全 Bayesian 分类器来增加属性为类提供的信息量, 以实现时序与非时序信息的融合, 并将基于具有对角平滑参数矩阵的多元 Gaussian 核函数估计属性条件联合密度、平滑参数的区间划分、时序递进分类准确性标准、平滑参数配置树的构建和分类器选择与平均等相结合来建立小时间序列动态完全 Bayesian 集成分类器. 使用宏观经济小时间序列数据集进行实验, 实验的结果显示, 经过优化的动态完全 Bayesian 集成分类器具有良好的分类准确性.

关键词 动态完全 Bayesian 分类器, 多元 Gaussian 核函数, 平滑参数, 分类准确性, 分类器选择与平均

1 引言

小时间序列 (如宏观经济年度时间序列, 以及其他时间序列的局部或一些子序列等) 普遍存在, 对小时间序列的分类预测也有着广泛的需求, 但由于小时间序列所蕴含的信息量不足 (分类器得不到充分的训练), 从而使提高小时间序列分类器的可靠性具有挑战性. 虽然已发展了许多著名的分类器, 如支持向量机、BP 神经网络、决策树和 Bayesian 网络等. 一方面, 这些分类器主要用于非时间序列数据 (要求数据集中记录之间满足独立同分布的假设, 而时间序列数据集不满足这一假设); 另一方面, 它们需要许多例子数据用于训练, 当训练不充分时, 往往很难取得好的分类效果. 动态 Bayesian 分类器是 Bayesian 分类器 [1] 的时序扩展, 由动态 Bayesian 分类器衍生的分类器 (动态 Bayesian 衍生分类器) 可用于时间序列数据分类, 但建立适合于小时间序列的动态 Bayesian 衍生分类器却非常困难, 而且这方面的研究也比较少.

引用格式: 王双成, 郑飞, 高瑞. 小时间序列动态完全 Bayesian 集成分类器研究. 中国科学: 信息科学, 2017, 47: 1445–1463, doi: 10.1360/N112017-00066
Wang S C, Zheng F, Gao R. Dynamic full Bayesian ensemble classifiers for small time series (in Chinese). Sci Sin Inform, 2017, 47: 1445–1463, doi: 10.1360/N112017-00066

目前, 对 Bayesian 衍生分类器 (Bayesian 分类器的衍生分类器) 的研究较多, 可以将这些研究大致分为具有离散属性和连续属性两种情况. 对于离散属性的情况, 如 Chow 和 Liu [2] 的依赖树分类器, Friedman 和 Geiger [3] 的 TAN (tree augmented naive Bayes) 分类器, Domingos 和 Pazzani [4] 在 0-1 损失下对朴素 Bayesian 分类器的优化, Campos 等 [5] 对 TAN 分类器的依赖扩展, Cheng 和 Greiner [6] 基于依赖分析确定结构的 Bayesian 网络分类器, Acid 等 [7] 通过打分搜索发现结构的 Bayesian 网络分类器, Yager [8] 对朴素 Bayesian 分类器的加权平均, Webb 等 [9] 的聚集一依赖估计 (aggregating one-dependence estimators) 分类器, Wang 等 [10, 11] 的 Markov 网络分类器和约束 Bayesian 分类网, Flores 等 [12] 的半朴素 Bayesian 分类器, Daniel 和 Aryeh [13] 关于朴素 Bayesian 分类器的有限样本分析等. 离散属性 (包括对连续属性的离散化) Bayesian 衍生分类器研究的核心是分类器结构, 常用的结构有变量星型结构 (朴素 Bayesian 分类器等)、Clique (不考虑变量之间条件独立性的变量集) 星型结构 (半朴素 Bayesian 分类器等)、树形结构 (TAN 分类器等)、有向网络结构 (Bayesian 网络分类器等)、无向网络结构 (Markov 网络分类器等)、汇聚一依赖结构 (汇聚一依赖估计分类器等) 和完全图结构 (完全 Bayesian 网络和完全 Markov 网络分类器等). 这些结构各具特色, 可分别用于满足不同的分类需要. 关于连续属性的情况, 20 世纪末, John 和 Langley [14] 使用经典的 Gaussian 函数和 Gaussian 核函数 (没有引入平滑参数) 估计属性边缘密度建立了两种朴素 Bayesian 分类器 (不需要分类器结构学习), 奠定了基于密度估计研究连续属性 Bayesian 衍生分类器的基础. Pérez 等 [15, 16] 在 John 和 Langley [14] 工作的基础上, 对两种朴素 Bayesian 分类器进行了依赖扩展. He 等 [17] 和 Luis 等 [18] 分别依据 Gaussian 函数与 Gaussian 核函数估计属性密度建立朴素 Bayesian 与完全 Bayesian 分类器, 以及它们在故障诊断和光谱分析方面的应用. Xiang 等 [19] 研究了基于 Gaussian 核函数估计属性边缘密度的属性加权朴素 Bayesian 分类器. Wang 等 [20~22] 分别基于 Gaussian 函数、Gaussian 核函数和 Gaussian Copula 函数估计属性密度, 并结合分类准确性标准与属性父结点的贪婪选择建立 Bayesian 网络分类器 (需要分类器结构学习). Wang 等 [1] 和 Dong 等 [23] 对连续属性完全 Bayesian 分类器 (不需要分类器结构学习) 进行了探索. 连续属性 (不进行离散化, 但需要估计属性密度) Bayesian 衍生分类器研究需要关注两个方面: 一个是分类器结构, 另一个是属性密度估计. 连续属性 Bayesian 衍生分类器的结构与离散属性的情况类似. 在属性密度估计方面, 目前主要采用 Gaussian 函数、Gaussian 核函数和 Copula 函数来估计属性密度, 它们各有优势与不足, 需要根据具体情况和需求来确定属性密度的估计方法. 虽然这些 Bayesian 衍生分类器不能直接用于时间序列 (尤其是小时间序列) 的分类, 但为动态 Bayesian 衍生分类器的研究奠定了基础.

对动态 Bayesian 衍生分类器的研究主要集中在离散属性的情况, 如 Martínez 和 Sucar [24]、Palacios-Alonso 等 [25]、Arriaga 等 [26] 和 Wang 等 [27] 探索了动态朴素 Bayesian 分类器 (最简单的动态 Bayesian 衍生分类器), 以及 Alkhateeb 等 [28]、Yu 等 [29]、Kafai 等 [30] 和 Wang 等 [31] 的动态 Bayesian 网络分类器的理论、方法和应用研究. 关于连续属性的动态 Bayesian 衍生分类器, Wang 等 [32, 33] 基于 Gaussian 函数和 Gaussian 核函数估计属性边缘密度建立动态朴素 Bayesian 分类器, 并将其用于小时间序列数据的分类, 收到了较好的效果, 但动态朴素 Bayesian 分类器同样也蕴含属性条件独立性的假设, 这使得分类器的属性为类提供的信息不够充分. 依据 Bayesian 衍生分类器中属性为类提供的信息构成理论 [22], 属性可为类提供 3 种依赖信息, 它们分别是传递依赖信息、直接导出依赖信息和间接导出依赖信息, 而朴素 Bayesian 分类器中的属性只能为类提供传递依赖信息, 从而会影响分类器的分类准确性. 这一结论同样适合于动态 Bayesian 衍生分类器, 也就是动态朴素 Bayesian 分类器中的属性同样只能为类提供传递依赖信息, 而动态完全 Bayesian 分类器中的属性可为类提供所有的 3 种依赖信息.

本文的主要贡献如下:

(1) 以动态朴素 Bayesian 分类器、动态完全 Bayesian 分类器和动态 Bayesian 网络分类器等为基础分类器, 分别对它们进行类时序、属性时序和混合时序依赖扩展形成时序同步的动态 Bayesian 衍生分类器, 再通过错位对应来建立时序非同步的动态 Bayesian 衍生分类器 (不同阶的动态 Bayesian 衍生分类器), 最终形成动态 Bayesian 衍生分类器体系架构, 为进一步深入研究动态 Bayesian 衍生分类器奠定了基础.

(2) 将基于具有对角平滑参数矩阵的多元 Gaussian 核函数估计属性条件联合密度, 时序递进分类准确性标准、平滑参数配置树的构建 (核心) 和分类器选择与平均等相结合, 给出了适合于小时间序列 (多时间序列) 的具有连续属性动态完全 Bayesian 集成分类器.

(3) 选择具有重要实际意义的宏观经济问题和真实小时间序列数据, 分别从平滑参数配置树的构建、分类准确性比较和平滑参数对分类准确性的影响 3 个方面进行实验与分析, 以及对动态完全 Bayesian 集成分类器可靠性的验证.

2 动态 Bayesian 衍生分类器的构成

首先介绍 Bayesian 网络 (Bayesian 和动态 Bayesian 分类器的基础), 然后给出 Bayesian 和动态 Bayesian 分类器的定义, 最后建立动态 Bayesian 衍生分类器的体系架构.

2.1 Bayesian 网络

Bayesian 网络 (Bayesian network) 是描述随机变量 (简称为变量) 之间有向依赖关系的图模型, 其他的图模型还包括 Markov 网络 (描述无向依赖关系) 和链图 (描述混合依赖关系) 等. Bayesian 网络由结构 (有向无环图) 和参数 (条件概率或密度) 两部分构成, 它的基本功能是分解联合概率 (或密度), 以提高联合概率 (或密度) 计算的效率和可靠性. Bayesian 分类器、Bayesian 衍生分类器、动态 Bayesian 分类器和动态 Bayesian 衍生分类器都是 Bayesian 网络 (或与时间有关的动态 Bayesian 网络) 的表现形式. 用 $Z_1, Z_2, \ldots, Z_n$ 表示变量, $z_1, z_2, \ldots, z_n$ 是它们的值, G 表示 Bayesian 网络, 则

$$\begin{aligned} p(z_1, z_2, \ldots, z_n) &= p(z_1)p(z_2|z_1)\cdots p(z_n|z_1, z_2, \ldots, z_{n-1}) \\ &= \prod_{i=1}^{n} p(z_i|z_1, z_2, \ldots, z_{i-1}) = \prod_{i=1}^{n} p(z_i|\pi_i, G), \end{aligned}$$

其中 π_i 是变量 Z_i 在 Bayesian 网络 G 中父结点集 Π_i 的配置, $\Pi_i \subseteq \{Z_1, Z_2, \ldots, Z_{i-1}\}$, 当给定 Π_i 时, Z_i 与 $\{Z_1, Z_2, \ldots, Z_{i-1}\} - \Pi_i$ 条件独立. 图 1 是 Heckerman 给出的一个影响学生进入高校因素的 Bayesian 网络 [34], 其中 H, SEX, SES, PE, IQ 和 CP 分别表示隐藏变量、性别、社会地位、父母的鼓励、智商和学院计划.

基于图 1 的 Bayesian 网络, 6 个变量的联合概率能够按 Bayesian 网络进行如下的分解:

$$\begin{aligned} &p(\mathrm{H}, \mathrm{SEX}, \mathrm{SES}, \mathrm{PE}, \mathrm{IQ}, \mathrm{CP}) \\ &= p(\mathrm{H})p(\mathrm{SEX}|\mathrm{H})p(\mathrm{SES}|\mathrm{H}, \mathrm{SEX})p(\mathrm{PE}|\mathrm{H}, \mathrm{SEX}, \mathrm{SES})p(\mathrm{IQ}|\mathrm{H}, \mathrm{SEX}, \mathrm{SES}, \mathrm{PE})p(\mathrm{CP}|\mathrm{H}, \mathrm{SEX}, \mathrm{SES}, \mathrm{PE}, \mathrm{IQ}) \\ &= p(\mathrm{H})p(\mathrm{SEX})p(\mathrm{SES}|\mathrm{H})p(\mathrm{PE}|\mathrm{SEX}, \mathrm{SES})p(\mathrm{IQ}|\mathrm{H}, \mathrm{PE})p(\mathrm{CP}|\mathrm{SES}, \mathrm{PE}, \mathrm{IQ}). \end{aligned}$$

一般联合概率计算的复杂性随变量增加指数增长, 基于 Bayesian 网络的联合概率分解计算可解决这一问题, 从而使解决以多变量联合概率计算为基础的实际问题成为可能. Bayesian 网络可以大致

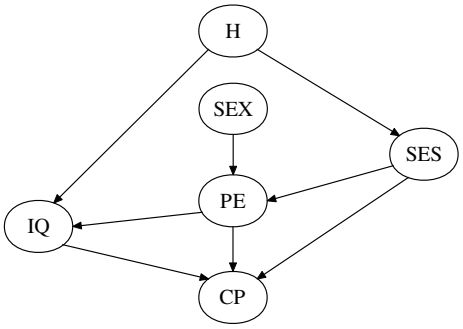

图 1　影响学生进入高校因素的 Bayesian 网络 (修改自文献 [34])

Figure 1　Bayesian network on the influencing factors for students into universities (modified from [34])

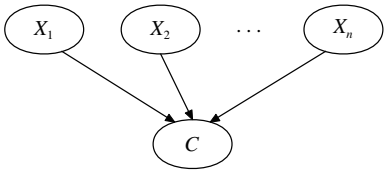

图 2　Bayesian 分类器的结构

Figure 2　Structure of BC

分成两种: 一种是弧的方向具有因果语义, 主要用于变量之间的因果分析 (也称为因果 Bayesian 网络), 需要考虑所有变量的条件或边缘概率; 另一种是弧的方向没有因果含义, 只表示信息的流动情况, 所突出的是类 (一个特殊的决策变量) 与属性 (除类之外的其他变量) 之间的映射关系 (也称为 Bayesian 网络分类器), 其核心是类的满条件概率计算, 对于动态 Bayesian 网络也是如此.

2.2　Bayesian 和动态 Bayesian 分类器的定义与表示形式

分别用 $X_1,\ldots,X_n$ 和 C 表示非时序属性和类, $x_1,\ldots,x_n,c$ 是具体的取值, D 是具有 N 个记录的非时序数据集.

定义1　称基于满条件概率 $p(c|x_1,\ldots,x_n)$ 进行分类的分类器为 Bayesian 分类器 (Bayesian classifier, BC).

BC 可以表示为

$$\arg\max_{c(x_1,\ldots,x_n)} \{p(c|x_1,\ldots,x_n)\}, \tag{1}$$

其中 $p(\cdot)$ 表示概率, max 表示对所有可能的 c 取最大的 $p(c|x_1,\ldots,x_n)$, arg 则表示具有最大概率的 c. 依据 Bayesian 网络理论, Bayesian 分类器的结构如图 2 所示.

分别用 $X_i[1],\ldots,X_i[T]$ $(1\leqslant i\leqslant n)$ 和 $C[1],\ldots,C[T]$ 表示时间序列属性变量 (简称为属性) 和时间序列类变量 (简称为类), $x_i[1],\ldots,x_i[T]$ 和 $c[1],\ldots,c[T]$ 是具体的取值; $D[1],\ldots,D[T]$ 是累计时间段数据集序列, $D[1]\subset D[2]\subset\cdots\subset D[T]$, $N[1],\ldots,N[T]$ 是对应时间段数据集中的例子数量. 动态 Bayesian 分类器是 Bayesian 分类器的时序扩展 (本文的动态是指与时间 t 有关, 来源于文献 [35]), 它可以有多种形式的定义, 我们给出下面的定义.

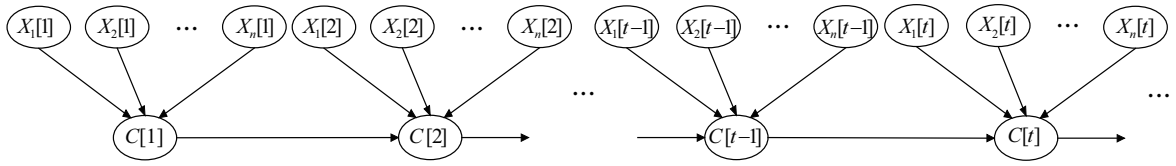

图 3 动态 Bayesian 分类器结构

Figure 3 Structure of DBC

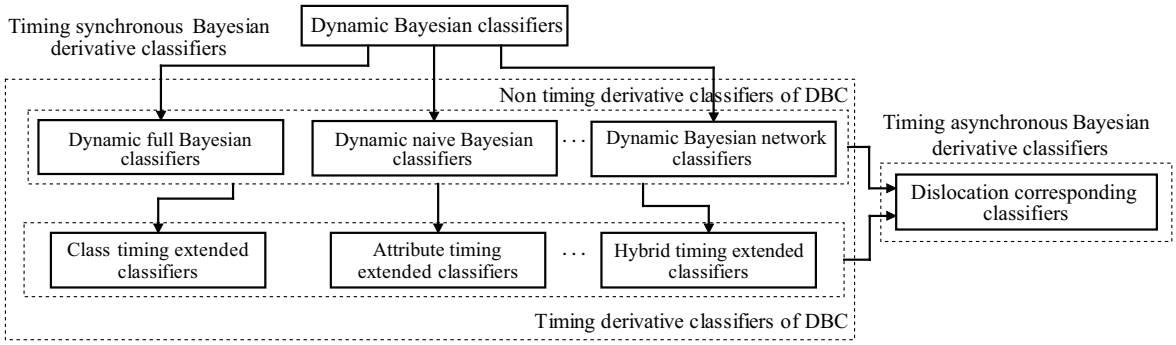

图 4 DBDC 的体系架构

Figure 4 Architecture of DBDC

定义2 称具有图 3 结构 (用 $G[t]$ 表示) 的分类器为动态 Bayesian 分类器 (dynamic Bayesian classifier, DBC).

依据 Bayesian 网络理论和图 3 中所蕴含的条件独立性关系 (给定 $C[t-1]$, $C[t]$ 与 $C[1],\ldots,C[t-2],X_1[1],\ldots,X_n[1],\ldots,X_1[t-1],\ldots,X_n[t-1]$ 之间条件独立), 可以得到

$$p\left(c[t]|c[1],\ldots,c[t-1],x_1[1],\ldots,x_n[1],\ldots,x_1[t],\ldots,x_n[t],G[t]\right)=p\left(c[t]|c[t-1],x_1[t],\ldots,x_n[t]\right).$$

DBC 可以表示为

$$\underset{c[t](c[t-1],x_1[t],\ldots,x_n[t])}{\arg\max}\left\{p\left(c[t]|c[t-1],x_1[t],\ldots,x_n[t]\right)\right\}. \tag{2}$$

2.3 动态 Bayesian 衍生分类器体系架构

由 DBC 可衍生出一系列的分类器, 本文将这些分类器统称为动态 Bayesian 衍生分类器 (dynamic Bayesian derivative classifier, DBDC), DBDC 的体系架构如图 4 所示.

图 4 给出的体系架构适合于具有离散、连续和混合属性的动态 Bayesian 衍生分类器, 这些分类器可用于各种情况的时间序列数据分类.

根据概率公式, 可以得到

$$p\left(c[t]|c[t-1],x_1[t],\ldots,x_n[t]\right)=\frac{p(c[t],c[t-1],x_1[t],\ldots,x_n[t])}{p(c[t-1],x_1[t],\ldots,x_n[t])}=\alpha p(c[t]|c[t-1])f(x_1[t],\ldots,x_n[t]|c[t]),$$

其中 α 是与 $C[t]$ 无关的量, $p(c[t]|c[t-1])$ 是类转移概率, $f(\cdot)$ 表示属性密度. 由 $f(x_1[t],\ldots,x_n[t]|c[t])$ 的不同计算方式可得到一系列的动态 Bayesian 衍生分类器, 本文将这些分类器称为 DBC 的非时序衍生

$x_1[1]$	...	$x_n[1]$	$\rightarrow$	$c[1]$
$x_1[2]$	...	$x_n[2]$	$\rightarrow$	$c[2]$
...	...	...	...	...
$x_1[t-1]$	...	$x_n[t-1]$	$\rightarrow$	$c[t-1]$
$x_1[t]$	...	$x_n[t]$	$\rightarrow$	$c[t]$
...	...	...	...	...

(a)

$x_1[1]$	...	$x_n[1]$	$\rightarrow$	$c[1+q]$
$x_1[2]$	...	$x_n[2]$	$\rightarrow$	$c[2+q]$
...	...	...	...	...
$x_1[t-1]$	...	$x_n[t-1]$	$\rightarrow$	$c[t-1+q]$
$x_1[t]$	...	$x_n[t]$	$\rightarrow$	$c[t+q]$
...	...	...	...	...

(b)

图 5　时序同步和时序非同步对应关系

Figure 5　Corresponding relationship of (a) timing synchronization and (b) timing asynchronous

分类器 (只在时间点或时间片内的结构发生变化), 其中具有代表性的 3 种分类器是动态完全 Bayesian 分类器、动态朴素 Bayesian 分类器和动态 Bayesian 网络分类器. DBC 的时序衍生分类器 (只在时间点或时间片之间的结构发生变化) 是指对 DBC 的非时序衍生分类器进行时序依赖扩展而得到的分类器, 包括类时序依赖扩展分类器、属性时序依赖扩展分类器和混合时序依赖扩展分类器等. 错位对应衍生分类器 (简称为错位对应分类器) 是指在时间序列 q ($q \geqslant 1$) 阶错位对应的基础上, 建立的 DBC 的非时序和时序衍生分类器. 非错位对应分类器是将 $x_1[t], \ldots, x_n[t], c[t-1]$ 作为输入, 对 $c[t]$ 进行分类预测, 而 q 阶错位对应分类器则是以 $x_1[t], \ldots, x_n[t], c[t]$ 作为输入对 $c[t+q]$ 进行分类预测, 时序同步和时序非同步的对应关系如图 5 所示.

3　动态完全 Bayesian 集成分类器

动态完全 Bayesian 分类器是一种 DBC 的非时序衍生分类器, 这种分类器不需要结构学习 (小时间序列分类器结构学习的可靠性无法得到保障), 而且能够充分利用小时间序列数据集中所蕴含的分类信息来提高分类准确性.

3.1　动态完全 Bayesian 分类器的定义和表示形式

定义3　称具有图 6(a) 结构 (用 $G_F[t]$ 表示) 的分类器为动态完全 Bayesian 分类器 (dynamic full Bayesian classifier, DFBC).

DFBC 可以表示为

$$\underset{c[t](c[t-1],x_1[t],\ldots,x_n[t])}{\arg\max} \{p(c[t]|c[t-1])f(x_1[t],\ldots,x_n[t]|c[t],G_F[t])\}. \tag{3}$$

在 DFBC 中, 属性可为类提供传递、直接导出和间接导出 3 种依赖信息. 传递依赖信息是由与类直接相连的属性所提供的信息, 是分类的主要信息, 图 6(b) 中的 $X_1[t], X_2[t], \ldots, X_n[t], C[t-1]$ 为 $C[t]$ 提供传递依赖信息; 直接导出依赖信息是由属性和类形成的 V 结构 [10] 直接诱导出的信息, 也是分类的重要信息; 间接导出依赖信息是由属性和属性形成的 V 结构所间接诱导出的信息, 往往也起到不可忽视的作用, 图 6(b) 中的 $X_1[t], X_2[t], \ldots, X_{n-1}[t]$ 还为 $C[t]$ 提供直接导出和间接导出两种依赖信息. 动态朴素 Bayesian 分类器中的属性只为类提供传递依赖一种信息, 因此, 相对于动态朴素 Bayesian 分类器, DFBC 应该具有更好的分类准确性. 从 DFBC 的定义和表示形式可知, 建立连续属性 DFBC 的核心是属性条件联合密度估计.

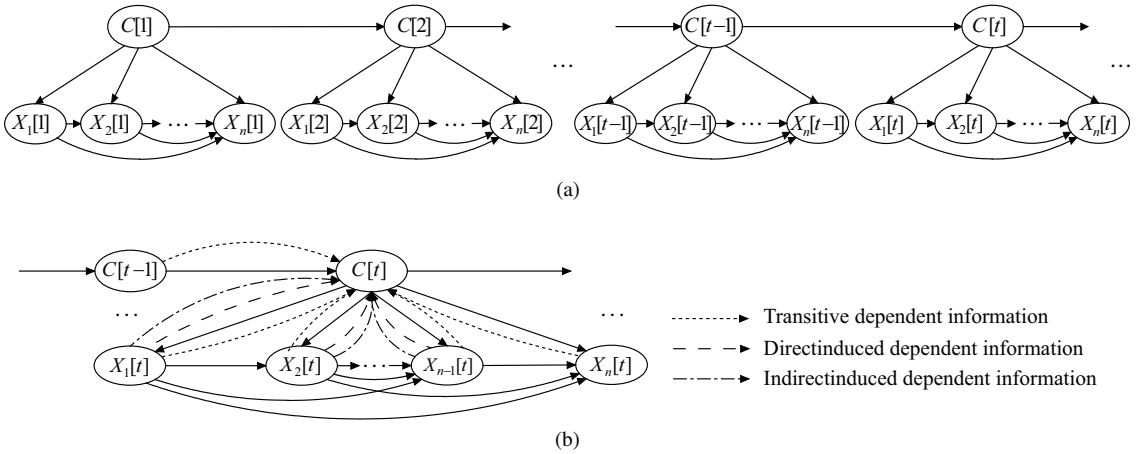

图 6 DFBC 的结构

Figure 6 Structure of DFBC. (a) Whole structure; (b) local structure

3.2 属性条件联合密度估计

建立小时间序列的分类器, 必须满足两个条件: 一个是分类器要与数据充分的拟合, 另一个是分类器不能与数据过度拟合. 我们将动态完全 Bayesian 分类器与基于多元 Gaussian 核函数的属性条件联合密度估计相结合来满足第 1 个条件, 为多元 Gaussian 核函数选取最简单的对角平滑参数矩阵和建立平滑参数配置树来满足第 2 个条件. 基于数据集 D 的具有对角平滑参数矩阵的多元核函数一般形式为

$$\varphi(x_1\cdots x_n|D)=\frac{1}{N\rho_1\cdots\rho_n}\sum_{m=1}^{N}\prod_{i=1}^{n}K_i\left(\frac{x_i-x_{im}}{\rho_i}\right),\tag{4}$$

其中 N 是数据集 D 中记录的数量, $\frac{1}{\rho_1\cdots\rho_n}\prod_{i=1}^{n}K_i(\frac{x_i-x_{im}}{\rho_i})$ 是核函数叠加中的第 m 项 (每个叠加项都是一个密度函数, 经过 N 次叠加后再平均得到的仍是一个密度函数), $K_i(\cdot)$ 是 X_i 的核函数, $\rho_1,\ldots,\rho_n$ 是平滑参数 (也称为窗宽或带宽), x_{im} $(1\leqslant i\leqslant n,\ 1\leqslant m\leqslant N)$ 表示 X_i 在数据集 D 中第 m 个记录的观测值.

设置 $K_i(\cdot)$ 为 Gaussian 函数 (也可以是其他的密度函数, 如均匀核函数、三角核函数和指数核函数等), 可以得到

$$K_i\left(\frac{x_i-x_{im}}{\rho_i}\right)=\frac{1}{\sqrt{2\pi}\rho_i}\exp\left[-\frac{(x_i-x_{im})^2}{2\rho_i^2}\right].$$

用 $\hat{p}(c[t]|c[t-1],D[t])$ 和 $\hat{f}(x_1[t],\ldots,x_n[t]|c[t],D[t])$ 表示 $p(c[t]|c[t-1])$ 和 $f(x_1[t],\ldots,x_n[t]|c[t])$ 的估计, 那么

$$\hat{p}(c[t]|c[t-1],D[t])=\frac{\hat{p}(c[t],c[t-1],D[t])}{\hat{p}(c[t-1],D[t])}=\frac{N(c[t],c[t-1])}{N(t)}\Big/\frac{N(c[t-1])}{N(t)}=\frac{N(c[t],c[t-1])}{N(c[t-1])},$$

其中 $N(t)$, $N(c[t-1])$ 和 $N(c[t],c[t-1])$ 分别是 $D[t]$ 中类时间序列所有数据的数量, $C[t-1]=c[t-1]$ 的情况数量和 $C[t]=c[t]$ 且 $C[t-1]=c[t-1]$ 的情况数量.

$$\hat{f}(x_1[t],\ldots,x_n[t]|c[t],D[t])=\frac{1}{(2\pi)^{n/2}N(c[t])\rho_1^2\cdots\rho_n^2}\sum_{v=1}^{t}\mathrm{signa}(c[v])\prod_{i=1}^{n}\exp\left[-\frac{(x_i[t]-x_i[v])^2}{2\rho_i^2}\right],\tag{5}$$

其中 ρ_i 也是标准差, $N(c[t])$ 是 $D[t]$ 中 $C[t]=c[t]$ 的情况数量, $\text{signa}(c[v])=\left\{\begin{array}{l}1,\ c[v]=c[t]\\0,\ c[v]\neq c[t]\end{array}\right.$.

DFBC 可以具体表示为

$$\begin{aligned}&\underset{c[t](c[t-1],x_1[t],\ldots,x_n[t])}{\arg\max}\{p(c[t]|c[t-1])f(x_1[t],\ldots,x_n[t]|c[t],D[t],G_F[t])\}\\&=\underset{c[t](c[t-1],x_1[t],\ldots,x_n[t])}{\arg\max}\left\{\frac{N(c[t],c[t-1])}{(2\pi)^{n/2}N(c[t])N(c[t-1])\rho_1^2\cdots\rho_n^2}\sum_{v=1}^{t}\text{signa}(c[v])\prod_{i=1}^{n}\exp\left[-\frac{(x_i[t]-x_i[v])^2}{2\rho_i^2}\right]\right\}.\end{aligned}\tag{6}$$

3.3 时序递进分类准确性标准

对于时间序列数据集 $D[T]$, 选择一个阈值 T_0, T_0 的值可依据时间序列的大小 T, 类转移概率与条件密度估计的有效性, 或实际需要来确定. 用 $\text{accuracy}(\text{DFBC},\boldsymbol{\rho},D[T],T_0)$ 表示 DFBC 的分类准确率, 其中 $\boldsymbol{\rho}=(\rho_1,\ldots,\rho_n)$ 是平滑参数向量, $c_{\text{prediction}}[t]$ 是使用 $D[t-1]$ 进行训练, 并依据 $x_1[t],\ldots,x_n[t],c[t-1]$ 对 $c[t]$ 的分类预测结果, $c_{\text{true}}[t]$ 是真正的结果, 那么

$$\text{accuracy}(\text{DFBC},\boldsymbol{\rho},D[T],T_0)=\frac{1}{T-T_0+1}\sum_{t=T_0}^{T}\text{signb}(c_{\text{prediction}}[t],c_{\text{true}}[t]),\tag{7}$$

其中 $\text{signb}(c_{\text{prediction}}[t],c_{\text{true}}[t])=\left\{\begin{array}{l}1,\ c_{\text{prediction}}[t]=c_{\text{true}}[t]\\0,\ c_{\text{prediction}}[t]\neq c_{\text{true}}[t]\end{array}\right.$.

$D[T]$ 中的数据记录之间具有时序影响, 因此不能采用非时间序列数据分类器的分类准确性评价标准 (已有许多评价标准). 对于时间序列数据分类器同样可以建立一系列的分类准确性评价标准, 本文给出的时序递进分类准确性评价标准是适合于小时间序列的标准.

3.4 平滑参数的优化

在对类时间序列进行分类预测时, 由于被预测的时序类值与时序邻近的类值往往有比较密切的联系 (已经通过大量的实验进行了验证), 因此, 能够准确预测邻近的连续时序类值越多的分类器应该更加可靠, 基于这一思想来建立小时间序列 DFBC 的平滑参数配置树. 通过实验发现, 平滑参数对 DFBC 分类准确性影响的核心区域是区间 $(0,1]$. 在 $(0,1]$ 中, 平滑参数越小 (Gaussian 函数曲线变陡), 分类器与数据的拟合程度越好, 但泛化性能变差; 随着平滑参数的增大 (Gaussian 函数曲线变缓), 分类器与数据的拟合程度会下降. 可见平滑参数的变化直接影响分类器与数据的拟合程度, 因此需要对平滑参数进行优化. 通过建立平滑参数配置树 (算法 1) 来优化平滑参数的配置, 将突出对被预测类值的时序邻近类值预测的可靠性, 因此适合于小时间序列数据集的分类. 如果是长时间序列, 只关注被预测类值的时序邻近类值预测 (局部) 的可靠性是不够的, 要根据实际情况和需求进行扩展 (也要放宽约束条件), 应该是局部与整体的结合. 将 $H=\left\{\rho^1,\rho^2,\ldots,\rho^L\right\}$ 作为每一个平滑参数的取值集合, 用 ρ_i^j $(1\leqslant i\leqslant n,1\leqslant j\leqslant L)$ 表示属性 $X_i[t]$ 的平滑参数 ρ_i 的第 j 个取值, 下面给出建立平滑参数配置树的算法.

L 是一个与 n 和 T 无关的量, 相对于分类准确性估计, 建立平滑参数配置树运算的时间复杂度是 $O(nT^2)$.

3.5 DFBC 的集成

遍历平滑参数配置树, 设具有最小阈值的平滑参数配置向量依次是 $\boldsymbol{\rho}^1,\ldots,\boldsymbol{\rho}^U$, 具有 $\boldsymbol{\rho}^u$ $(1\leqslant u\leqslant U)$ 参数配置的动态完全 Bayesian 分类器用 DFBC_u 表示, 选择这些分类器, 通过分类器平均建立动态

算法 1. 建立平滑参数配置树算法

输入: 时间序列数据集 $D[T]$、平滑参数初始化阈值 T_0 和平滑参数取值集合 H
输出: 平滑参数配置树
计算 $\rho^* = \arg\max_{\rho=\rho_1=,\ldots,=\rho_n\in H}\{\text{accuracy}(\text{DFBC}, \boldsymbol{\rho}, D[T], T_0)\}$ //确定初始平滑参数
初始化平滑参数向量和平滑参数搜索数组 smoothing_parameter[] $= \rho^*$, search[] $= 0$;
for $t = T$ to 1
  if accuracy(DFBC, smoothing_parameter[], $D[T], t) < 1$ then
    保存 $T^* = t$;
    exit for
  end if
end for
创建一个结点 N (根结点), 将 T^* 和 ρ^* 存入结点 N 的数据域, $T' = T^*$;
for $t = T'$ to 1 //按层次建树
  深度优先遍历树;
  if 一条路径的终端结点是非叶子结点 then
    按照从根结点到终端结点的路径重新配置平滑参数向量 smoothing_parameter[] 和
    平滑参数搜索数组 search[], 并将路径终端结点设置为当前结点, sign=0;
    for i=1 to n
      if search[i] $= 0$ then //属性 X_i 的平滑参数 ρ_i 没有进行过优化设置
        for $j = L$ to 1
          if accuracy(DFBC, $(\ldots, \rho_i^j, \ldots), D[T], t) = 1$ then //设置 smoothing_parameter[i] $= \rho_i^j$
            search[i] $= 1$, 更新平滑参数向量 smoothing_parameter[], sign $= 1$, mark $= 0$;
            for $u = t$ to 1 //发现临界值 T^*
              if accuracy(DFBC, smoothing_parameter[], $D[T], t) < 1$ then
                保存 $T^* = t$, 创建一个新结点 N, 把 T^* 和 ρ_i^j 存入结点, mark=1;
                exit for
              end if
            end for
            if mark=0 then
              设置最新创建的结点为叶子结点;
            end if
          end if
        end for
      end if
    end for
    if sign=0 then
      设置当前结点为叶子结点;
    end if
  end if
end for

完全 Bayesian 集成分类器 (dynamic full Bayesian ensemble classifier, DFBEC), DFBEC 的集成结构如图 7 所示.

DFBEC 可以表示为:

$$\arg\max_{c[t](c[t-1],x_1[t],\ldots,x_n[t])}\left\{\frac{1}{U}\sum_{u=1}^{U}\frac{N(c[t],c[t-1])}{(2\pi)^{n/2}N(c[t])N(c[t-1])\left(\rho_1^u\right)^2\cdots\left(\rho_n^u\right)^2}\right.$$

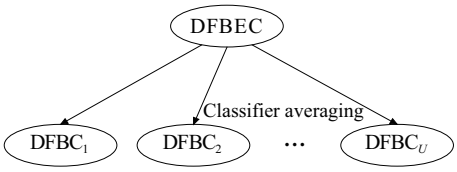

图 7 DFBEC 的集成结构

Figure 7 Ensemble structure of DFBEC

$$\cdot \sum_{v=1}^{t} \operatorname{signa}(c[v]) \prod_{i=1}^{n} \exp\left[-\frac{(x_i[t]-x_i[v])^2}{2\left(\rho_i^u\right)^2}\right]\Bigg\}, \tag{8}$$

其中 ρ_i^u 是 $\boldsymbol{\rho}^u$ 的第 i 个分量. U 是一个与 n 和 T 无关的量, 相对于 Gaussian 函数计算, DFBEC 分类运算的时间复杂度是 $O(nT)$.

4 实验与分析

分别选择国内生产总值 (gross domestic product, GDP), 属性是第一产业同比增长率、第二产业同比增长率、第三产业同比增长率、M0 同比增长率、M1 同比增长率、M2 同比增长率、储蓄存款同比增长率、各项贷款总额同比增长率、全社会固定资产投资总额同比增长率、社会消费品零售总额同比增长率、居民消费价格总指数、全国零售商品价格总指数、出口商品总额同比增长率、进口商品总额同比增长率、财政预算收入完成额同比增长率、财政预算支出完成额同比增长率和 GDP 同比增长率 17 个指标, 类是 GDP 增长率是否为转折点; 外贸进出口 (import and export, IAE), 属性是美元汇率、欧元汇率和原油价格等 6 个指标, 类是 IAE 增长率是否为转折点; 失业率 (unemployment rate, UR), 属性是城镇登记失业率、规模以上工业企业利润总额增长率和规模以上工业企业所得税增长率等 9 个指标, 类是 UR 是否为转折点; 生产者物价指数 (producer price index，PPI), 属性是工业增加值同比增长、工业生产者购进价格指数和城镇单位就业人员平均货币工资指数等 20 个指标, 类是 PPI 是否为转折点; 居民消费价格指数 (consumer price index, CPI), 属性是平均房价变化指数、人均收入变化率和人均支出变化率等 15 个指标, 类是 CPI 是否为转折点; 工业生产指数 (industrial production index, IPI), 属性是工业增加值同比增长、工业用电消费总量变化率和工业煤炭消费总量变化率等 11 个指标, 类是 IPI 是否为转折点; 固定资产投资 (fixed asset investment, FAI), 属性是国家财政收入累计增长率、国家财政支出累计增长率和房地产投资累计增长率等 26 个指标, 类是 FAI 完成额累计增长率是否为转折点; 通货膨胀率 (inflation rate, IR), 属性是产出缺口、利率和汇率等 9 个指标, 类是 IR 是否为转折点; 税收收入 (tax revenue, TR), 属性是关税收入同比增长率、外贸企业出口退税同比增长率和进出口规模同比增长率等 14 个指标, 类是 TR 同比增长率是否为转折点. 按照是否为时序转折点 (时序变化的上下局部极值点为转折点) 对这 9 个指标进行二值离散化得到 9 个类变量 (简称为类), 将影响它们的相关因素作为属性指标 (简称为属性, 所有的影响因素都取实数值). 分别从平滑参数配置树的构建、分类准确率比较和平滑参数变化对分类准确性的影响 3 个方面进行实验与分析, 用于实验的时间序列数据来源于 Wind 数据库和国家统计局数据库, 基本情况如表 1 所示.

表 1 时间序列数据集情况

Table 1 Time series data sets

Data set	Size	Classes	Attributes
GDP	29	2	17
IAE	37	2	6
UR	31	2	9
PPI	35	2	20
CPI	45	2	15
IPI	47	2	11
FAI	118	2	26
IR	60	2	9
TR	166	2	14

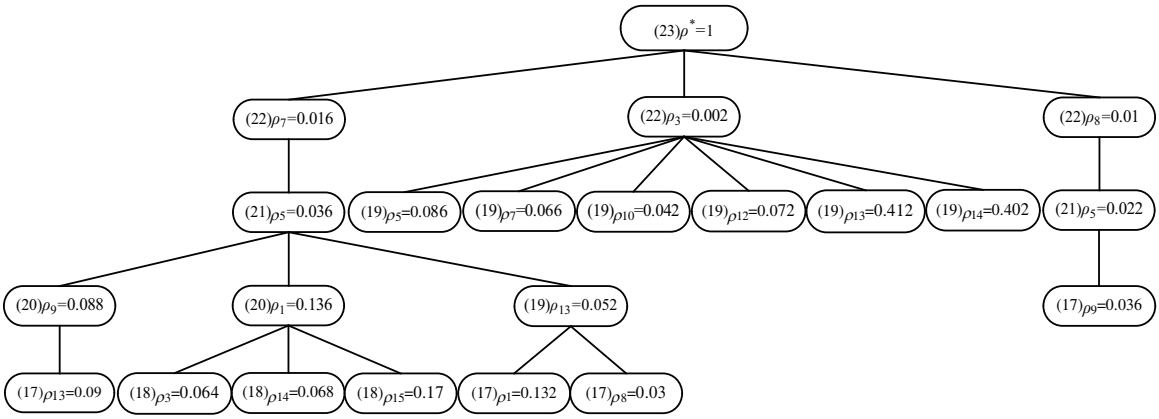

图 8 GDP 的平滑参数配置树

Figure 8 Smoothing parameter configuration tree of GDP

4.1 平滑参数配置树的构建

选择 GDP 年度数据, 取 $T_0 = 15$ (确定初始平滑参数的阈值) 和 $H = \{0.002k\}$ $(1 \leqslant k \leqslant 500)$ (平滑参数的取值集合), 得到平滑参数的初始配置和准确分类的临界值是 $\rho^* = 1$ (所有平滑参数都取 1) 和 $T^* = 23$ ($T_0 > 23$ 时, $\text{accuracy}(\text{DFBC}, \rho^*, D[T], T_0) = 1$; $T_0 = 23$ 时, $\text{accuracy}(\text{DFBC}, \rho^*, D[T], T_0) < 1$), 将 $T^* = 23$ 和 $\rho^* = 1$ 存入一个结点的数据域, 并将这个结点作为平滑参数配置树的根结点, 依据算法 1 建立的平滑参数配置树如图 8 所示.

在平滑参数配置树中, 除根结点外, 每一个结点的数据域需要存储准确分类的临界值和一个平滑参数值, 从根结点到叶子结点的每一支都对应一个平滑参数配置向量, 而由一个平滑参数配置向量又可得到一个 DFBC. 遍历平滑参数配置树, 得到最小临界值是 17, 具有最小临界值的平滑参数配置向量为

$$(1, 1, 1, 1, 0.036, 1, 0.016, 1, 0.088, 1, 1, 1, 0.09, 1, 1, 1, 1),$$
$$(0.132, 1, 1, 1, 0.036, 1, 0.016, 1, 1, 1, 1, 1, 0.052, 1, 1, 1, 1),$$

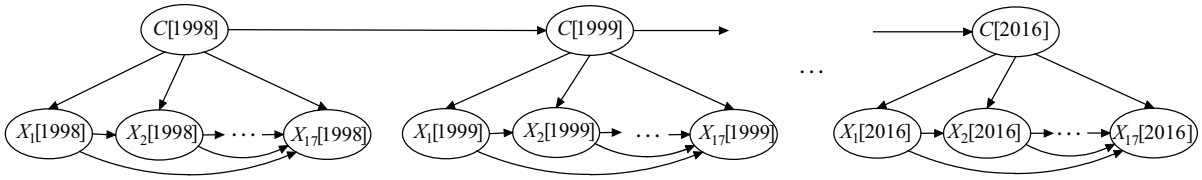

图 9 GDP 的 DFBC 的结构

Figure 9 Structure of DFBC for GDP

$$(1,1,1,1,0.036,1,0.016,0.03,1,1,1,1,0.052,1,1,1,1)\,,$$
$$(1,1,1,1,0.022,1,1,0.01,0.036,1,1,1,1,1,1,1,1)\,.$$

由上面的 4 个平滑参数配置向量可获得 4 个 DFBC (在平滑参数配置树中所进行的平滑参数配置向量选择, 实质上就是 DFBC 选择), 通过分类器平均得到 GDP 的 DFBEC. 也可以放宽最小阈值的限制, 如在 GDP 的平滑参数配置树中取 19 作为分类器选择的临界值, 可得到 13 个平滑参数配置向量, 以及它们对应的 13 个 DFBC.

用于对 $C[2016]$ 进行分类预测的 DFBC 的结构如图 9 所示. 根据图 9 中蕴含的变量之间条件独立性关系, 可以得到

$$\begin{aligned}
&p(c[2016]|c[1998],\ldots,c[2015],x_1[1998],\ldots,x_{17}[1998],\ldots,x_1[2016],\ldots,x_{17}[2016],G_F[2016])\\
&\quad=p(c[2016]|c[2015],x_1[2016],\ldots,x_{17}[2016])\\
&\quad=\frac{p(c[2016],c[2015],x_1[2016],\ldots,x_{17}[2016])}{p(c[2015],x_1[2016],\ldots,x_{17}[2016])}\\
&\quad=\alpha p(c[2016]|c[2015])f(x_1[2016],\ldots,x_{17}[2016]|c[2016]).
\end{aligned}$$

具有图 9 结构和用于对 $C[2016]$ 进行分类预测的 DFBC 能够被表示为

$$\underset{c[2016](c[2015],x_1[2016],\ldots,x_{17}[2016])}{\arg\max}\{p(c[2016]|c[2015])f(x_1[2016],\ldots,x_{17}[2016]|c[2016],G_F[t])\}\,.$$

4.2 分类准确性比较

使用 8 个分类器进行分类准确性比较实验与分析, 它们是条件随机场 (conditional random fields, CRF) 分类器, 在文献 [32] (使用 Gaussian 函数估计属性条件边缘密度) 和文献 [33] (采用 Gaussian 核函数估计属性条件边缘密度) 中给出的两个动态朴素 Bayesian 分类器 (记为 GDNB 和 KDNB), 平滑参数随机变化的 DFBC 投票集成分类器 (记为 RDFBC, 随机产生 20 个 DFBC, 并进行投票集成), 在 DFBC 中分别采用 Gaussian 函数和 Gaussian Copula 函数估计属性条件联合密度的两个分类器 (记为 GDFB 和 KDFB), 通过分类器选择得到的 DFBSC (对具有相同最小阈值情况的分类器进行随机选择) 和采用分类器平均得到的 DFBEC (对具有相同最小阈值情况的分类器进行平均), 其中 T_0 的值依次选取后 13 个时间点 (或时间片), 实验结果如表 2∼10 所示.

从表 2∼10 能够发现, 在分类准确性方面 DFBEC 具有明显的优势. DFBEC 与 CRF 的比较: 在 CRF 中, 需要将属性离散化, 这样会导致一部分信息的丢失; 另外 CRF 还需要计算高阶条件概率, 其可靠性也得不到保障, 而 DFBEC 不存在这些的问题. DFBEC 与 GDNB 和 KDNB 的比较: 根

表 2 GDP 波动转折点分类结果

Table 2 Classification results of GDP fluctuation turning point

Classifier	T_0=17	T_0=18	T_0=19	T_0=20	T_0=21	T_0=22	T_0=23	T_0=24	T_0=25	T_0=26	T_0=27	T_0=28	T_0=29	Ave.
CRF	69.23	75.00	72.72	70.00	66.66	75.00	85.71	83.33	100.00	100.00	100.00	100.00	100.00	84.43
GDNB	69.23	66.66	63.63	70.00	77.77	75.00	71.42	83.33	80.00	100.00	100.00	100.00	100.00	81.31
KDNB	76.92	83.33	81.81	90.00	88.88	87.50	100.00	100.00	100.00	100.00	100.00	100.00	100.00	92.96
RDFBC	69.23	75.00	72.72	70.00	66.66	62.50	57.14	50.00	40.00	50.00	33.33	50.00	100.00	61.27
GDFB	46.15	50.50	45.45	50.50	55.55	62.50	57.14	66.66	40.00	50.00	66.66	50.00	0.00	49.35
KDFB	69.23	66.66	72.72	70.00	66.66	62.50	57.14	50.00	60.00	75.00	66.66	100.00	100.00	70.51
DFBSC	92.30	91.66	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	98.22
DFBEC	92.30	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.41

表 3 IAE 波动转折点分类结果

Table 3 Classification results of IAE fluctuation turning point

Classifier	T_0=25	T_0=26	T_0=27	T_0=28	T_0=29	T_0=30	T_0=31	T_0=32	T_0=33	T_0=34	T_0=35	T_0=36	T_0=37	Ave.
CRF	61.53	58.33	63.63	70.00	66.66	62.50	71.42	83.33	80.00	75.00	66.66	50.00	100.00	69.93
GDNB	53.84	50.00	54.54	60.00	55.55	50.00	42.85	50.00	40.00	50.00	66.66	50.00	100.00	55.65
KDNB	69.23	66.66	72.72	70.00	66.66	62.50	57.14	66.66	80.00	75.00	100.00	100.00	100.00	75.89
RDFBC	61.53	58.33	63.63	70.00	66.66	62.50	57.14	66.66	80.00	75.00	66.66	100.00	100.00	71.39
GDFB	38.45	33.33	36.36	40.40	44.44	50.00	57.14	66.66	60.00	50.00	33.33	50.00	0.00	43.09
KDFB	53.84	58.33	54.54	50.00	55.55	50.00	42.85	33.33	40.00	50.00	33.33	50.00	0.00	43.98
DFBSC	76.92	83.33	72.72	80.00	77.77	75.00	85.71	83.33	80.00	100.00	100.00	100.00	100.00	85.75
DFBEC	76.92	83.33	81.81	80.00	88.88	87.50	85.71	100.00	100.00	100.00	100.00	100.00	100.00	91.09

表 4 UR 波动转折点分类结果

Table 4 Classification results of UR fluctuation turning point

Classifier	T_0=19	T_0=20	T_0=21	T_0=22	T_0=23	T_0=24	T_0=25	T_0=26	T_0=27	T_0=28	T_0=29	T_0=30	T_0=31	Ave.
CRF	61.53	58.33	63.63	60.00	55.55	50.00	57.14	50.00	60.00	50.00	33.33	00.00	0.00	46.12
GDNB	46.15	41.66	45.45	50.00	44.44	50.00	42.85	50.00	40.00	50.00	66.66	50.00	100.00	52.09
KDNB	61.53	66.66	63.63	60.00	55.55	62.50	71.42	66.66	60.00	75.00	100.00	100.00	100.00	72.53
RDFBC	61.53	66.66	72.72	70.00	66.66	62.50	57.14	66.66	60.00	50.00	33.33	0.00	0.00	51.32
GDFB	38.46	41.66	36.36	40.00	33.33	37.50	42.85	33.33	40.00	25.00	33.33	50.00	100.00	42.45
KDFB	61.53	58.33	63.63	60.00	66.66	75.00	71.42	66.66	60.00	75.00	66.66	50.00	100.00	67.30
DFBSC	76.92	75.00	72.72	70.00	66.66	75.00	71.42	83.33	80.00	75.00	100.00	100.00	100.00	80.47
DFBEC	92.30	91.66	90.90	90.00	88.88	87.50	100.00	100.00	100.00	100.00	100.00	100.00	100.00	95.48

据 Bayesian 衍生分类器中属性为类提供的信息构成理论, GDNB 和 KDNB 中的属性只能为类提供一种依赖信息 (传递依赖信息), 而 DFBEC 中的属性却能够为类提供包括传递依赖信息在内的 3 种依赖信息. DFBEC 与 RDFBC 的比较: RDFBC 是对平滑参数随机变化而产生的分类器所进行的集成, 这些分类器之间的分类准确性差异较大, 从而会降低 RDFBC 的可靠性, DFBEC 不存在这一问题. DFBEC 与 GDFB 和 KDFB 的比较: 一方面, GDFB 和 KDFB 易于导致对数据的过度拟合, 从而会

表 5 PPI 波动转折点分类结果

Table 5 Classification results of PPI fluctuation turning point

Classifier	T_0=23	T_0=24	T_0=25	T_0=26	T_0=27	T_0=28	T_0=29	T_0=30	T_0=31	T_0=32	T_0=33	T_0=34	T_0=35	Ave.
CRF	53.84	58.33	63.63	60.00	55.55	62.50	57.14	66.66	80.00	100.00	100.00	100.00	100.00	73.67
GDNB	46.15	41.66	45.45	40.00	33.33	37.50	28.57	33.33	40.00	50.00	66.66	100.00	100.00	50.97
KDNB	61.53	66.66	63.63	70.00	77.77	75.00	71.42	66.66	80.00	100.00	100.00	100.00	100.00	79.43
RDFBC	69.23	75.00	72.72	70.00	77.77	75.00	85.71	83.33	80.00	75.00	66.66	50.00	0.00	67.72
GDFB	38.46	33.33	36.36	40.00	33.33	37.50	42.85	50.00	40.00	50.00	66.66	100.00	100.00	51.42
KDFB	53.84	50.00	54.54	50.00	44.44	37.50	42,85	50.00	60.00	50.00	33.33	50.00	100.00	52.80
DFBSC	76.92	75.00	72.72	80.00	77.77	85.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	89.80
DFBEC	84.61	83.33	90.90	90.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	96.06

表 6 CPI 波动转折点分类结果

Table 6 Classification results of CPI fluctuation turning point

Classifier	T_0=33	T_0=34	T_0=35	T_0=36	T_0=37	T_0=38	T_0=39	T_0=40	T_0=41	T_0=42	T_0=43	T_0=44	T_0=45	Ave.
CRF	61.53	58.33	63.63	60.00	55.55	62.50	57.14	66.66	80.00	100.00	100.00	100.00	100.00	74.26
GDNB	38.46	33.33	36.36	40.00	33.33	37.50	42.85	50.00	60.00	75.00	100.00	100.00	100.00	57.45
KDNB	53.84	58.33	63.63	70.00	66.66	75.00	71.42	83.33	100.00	100.00	100.00	100.00	100.00	80.17
RDFBC	69.23	66.66	63.63	70.00	77.77	75.00	85.71	83.33	80.00	75.00	100.00	100.00	100.00	80.48
GDFB	46.15	50.00	54.54	60.00	55.55	50.00	42.85	50.00	60.00	75.00	100.00	100.00	100.00	64.93
KDFB	46.15	50.00	45.45	50.00	44.44	50.00	42.85	33.33	40.00	50.00	33.33	0.00	0.00	37.35
DFBSC	76.92	75.00	72.72	80.00	77.77	87.50	100.00	100.00	100.00	100.00	100.00	100.00	100.00	89.99
DFBEC	76.92	83.33	81.81	90.00	88.88	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	93.92

表 7 IPI 波动转折点分类结果

Table 7 Classification results of IPI fluctuation turning point

Classifier	T_0=35	T_0=36	T_0=37	T_0=38	T_0=39	T_0=40	T_0=41	T_0=42	T_0=43	T_0=44	T_0=45	T_0=46	T_0=47	Ave.
CRF	38.46	41.66	45.45	40.00	44.44	50.00	42.85	33.33	20.00	25.00	0.00	0.00	0.00	29.32
GNBC	53.84	50.00	45.45	50.00	44.44	37.50	28.57	33.33	40.00	50.00	33.33	50.00	0.00	39.73
KNBC	53.83	58.33	54.54	60.00	66.66	75.00	57.14	50.00	60.00	50.00	66.66	50.00	100.00	61.70
RDFBC	46.15	41.66	45.45	50.00	44.44	50.00	57.14	66.66	60.00	50.00	66.66	100.00	100.00	59.85
GDFB	30.76	33.33	36.35	30.00	33.33	37.50	42.85	33.33	40.00	25.00	33.33	50.00	0.00	32.75
KDFB	46.15	50.00	54.54	60.00	66.66	50.00	42.85	33.33	20.00	0.00	0.00	0.00	0.00	32.58
DFBSC	53.84	58.33	54.54	60.00	66.66	75.00	57.14	50.00	60.00	75.00	66.66	100.00	100.00	67.47
DFBEC	69.23	66.66	63.63	70.00	66.66	62.50	71.42	83.33	100.00	100.00	100.00	100.00	100.00	81.03

降低分类器的泛化性能; 另一方面, GDFB 和 KDFB 都需要协方差矩阵的计算, 而小时间序列所蕴含的信息不充分, 无法可靠地估计协方差矩阵和它的逆矩阵. DFBEC 与 DFBSC 的比较: DFBSC 所进行的是分类器选择, 即选择最好的分类器, 但在实际中很难做到, 而且每一个 DFBC 也都有它的优势和不足; DFBEC 是将所选择的分类器进行平均, 可使这些分类器之间形成互补, 因此会有更好的分类效果.

表 8 FAI 波动转折点分类结果

Table 8 Classification results of FAI fluctuation turning point

Classifier	T_0=106	T_0=107	T_0=108	T_0=109	T_0=110	T_0=111	T_0=112	T_0=113	T_0=114	T_0=115	T_0=116	T_0=117	T_0=118	Ave.
CRF	61.53	58.33	63.63	70.00	77.77	75.00	71.42	66.66	80.00	100.00	100.00	100.00	100.00	78.80
GNBC	38.46	41.66	45.45	50.00	44.44	37.50	42.85	50.00	60.00	75.00	66.66	50.00	0.00	46.31
KNBC	69.23	66.66	63.63	70.00	77.77	87.50	85.71	83.33	80.00	75.00	66.66	100.00	100.00	78.88
RDFBC	53.84	58.33	54.54	50.00	55.55	62.50	71.42	66.66	60.00	75.00	66.66	50.00	100.00	63.42
GDFB	53.84	50.00	45.45	40.00	44.44	50.00	57.14	50.00	40.00	25.00	0.00	0.00	0.00	35.07
KDFB	61.53	58.33	54.54	50.00	44.44	50.00	57.14	66.66	60.00	75.00	100.00	100.00	100.00	67.51
DFBSC	69.23	66.66	72.72	80.00	77.77	87.50	85.71	83.33	80.00	100.00	100.00	100.00	100.00	84.84
DFBEC	76.92	83.33	81.81	80.00	77.77	75.00	85.71	100.00	100.00	100.00	100.00	100.00	100.00	89.27

表 9 IR 波动转折点分类结果

Table 9 Classification results of IR fluctuation turning point

Classifier	T_0=48	T_0=49	T_0=50	T_0=51	T_0=52	T_0=53	T_0=54	T_0=55	T_0=56	T_0=57	T_0=58	T_0=59	T_0=60	Ave.
CRF	53.84	58.33	63.63	60.00	55.55	62.50	57.14	50.00	60.00	50.00	66.66	100.00	100.00	64.43
GNBC	46.15	41.66	45.45	40.00	44.44	50.00	42.85	50.00	40.00	25.00	0.00	0.00	0.00	32.73
KNBC	69.23	66.66	63.63	60.00	66.66	75.00	85.71	100.00	100.00	100.00	100.00	100.00	100.00	83.61
RDFBC	69.23	66.66	63.63	60.00	55.55	50.00	42.85	33.33	20.00	25.00	33.33	0.00	0.00	39.97
GDFB	53.83	50.00	45.45	50.00	55.55	62.50	57.14	50.00	60.00	50.00	66.66	50.00	100.00	57.78
KDFB	61.53	58.33	54.54	50.00	44.44	37.50	42.85	50.00	40.00	50.00	66.66	100.00	100.00	58.14
DFBSC	69.23	75.00	72.72	70.00	77.77	87.50	100.00	100.00	100.00	100.00	100.00	100.00	100.00	88.63
DFBEC	84.61	83.33	90.90	90.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	96.06

表 10 TR 波动转折点分类结果

Table 10 Classification results of TR fluctuation turning point

Classifier	T_0=154	T_0=155	T_0=156	T_0=157	T_0=158	T_0=159	T_0=160	T_0=161	T_0=162	T_0=163	T_0=164	T_0=165	T_0=166	Ave.
CRF	46.15	41.66	45.45	50.00	55.55	62.50	71.42	83.33	80.00	75.00	66.66	50.00	100.00	63.67
GNBC	46.15	41.66	36.35	40.00	33.33	37.50	42.85	33.33	20.00	25.00	33.33	0.00	0.00	29.96
KNBC	69.23	66.66	72.72	70.00	66.66	75.00	85.71	83.33	100.00	100.00	100.00	100.00	100.00	83.79
RDFBC	61.53	58.33	63.63	70.00	66.66	62.50	57.14	50.00	80.00	75.00	66.66	100.00	100.00	70.11
GDFB	46.15	50.00	45.45	50.00	55.55	50.00	42.85	33.33	20.00	25.00	0.00	0.00	0.00	32.18
KDFB	61.53	58.33	54.54	50.00	44.44	37.50	42.85	50.00	60.00	75.00	100.00	100.00	100.00	64.17
DFBSC	76.92	75.00	81.81	80.00	88.88	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	92.51
DFBEC	92.30	91.66	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	98.77

4.3 平滑参数变化对分类准确性的影响

平滑参数决定着 Gaussian 函数曲线的形状, 因此, 平滑参数的变化将影响分类器与数据的拟合程度使用 GDP, IAE, UR, PPI, CPI 和 IPI 6 个时间序列数据集, T_0 分别取 9, 17, 11, 15, 25 和 27, 用 $s1,\ldots,s9,s10,\ldots,s18,s19,\ldots,s27,s28$ 表示 $0.001,\ldots,0.009,0.01,\ldots,0.09,0,1,\ldots,0.9,1$, 分别从平滑参数的同步变化 (所有平滑参数同步取值, 相当于只有一个平滑参数) 和异步变化 (平滑参数之间可以

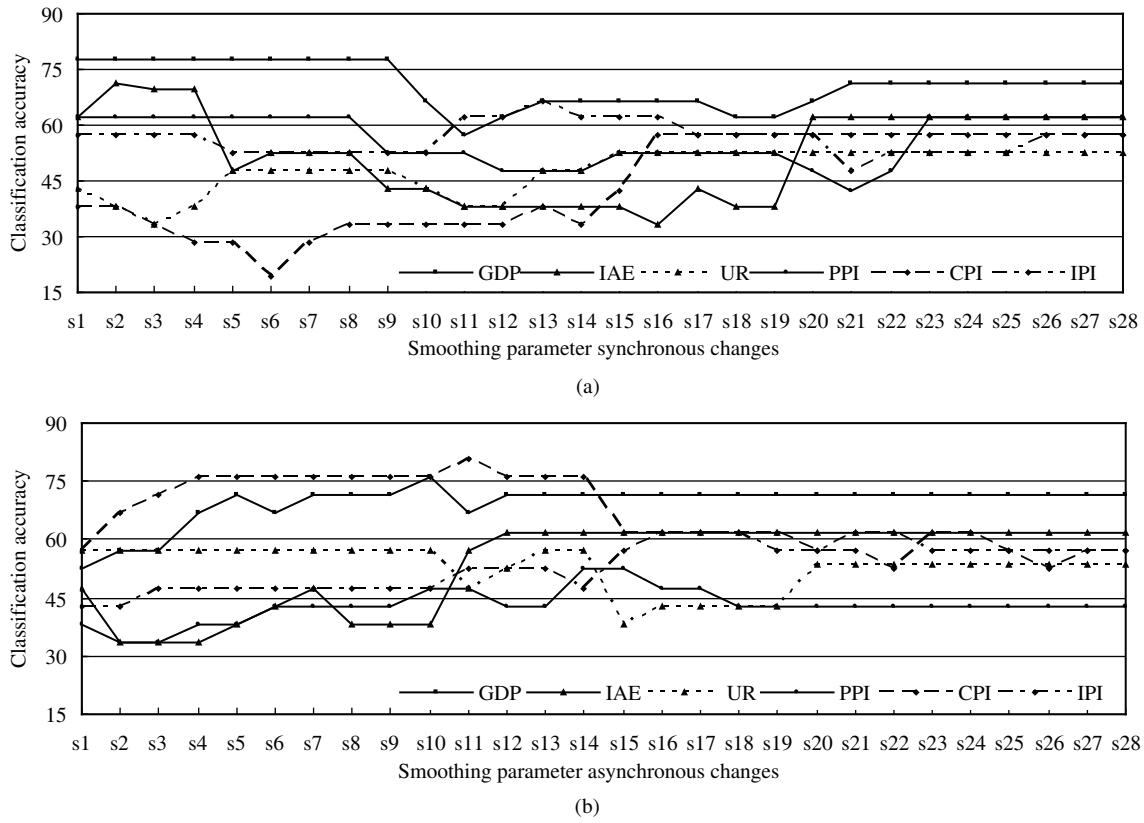

图 10　平滑参数变化对分类准确性的影响

Figure 10 Influence of smoothing parameter changes on classification accuracy. Influence of (a) synchronous and (b) asynchronous changes

取不同的值, 但变化的平滑参数只有一个, 其他的平滑参数需要控制) 两方面, 进行平滑参数变化对分类准确性的影响实验与分析, 6 个时间序列数据集的平滑参数变化对分类准确性的影响如图 10 所示. 为弱化初始配置对平滑参数变化的影响, 对 6 个时间序列数据集平滑参数初始配置都取 1, 异步变化的平滑参数按数据集的顺序依次选取 ρ_7, ρ_1, ρ_2, ρ_{13}, ρ_5 和 ρ_5.

从图 10 中可以看到, 无论是同步变化还是异步变化, 6 个时间序列数据集的平滑参数变化对 DFBC 的分类准确性都有较大的影响. 在同步变化方面, 最大分类准确性差异依次是 14.28%, 38.09%, 19.05%, 19.05%, 38.09% 和 19.05% (平均值是 24.60%); 在异步变化方面, 最大分类准确性差异依次是 23.79%, 28.57%, 19.05%, 19.05%, 28.57% 和 19.05% (平均值是 23.01%), 因此, 需要对平滑参数进行优化. 再有, 一些平滑参数的变化对 DFBC 的分类准确性有影响, 而另一些平滑参数的变化却没有影响, 在 6 个时间序列数据集中, 对 DFBC 的分类准确性有影响的参数数量分别是 11, 5, 7, 13, 10 和 7.

5　结论和进一步的工作

在动态 Bayesian 分类器的基础上, 分别从非时序衍生分类器、时序衍生分类器和错位对应衍生分类器 3 个方面构建了动态 Bayesian 衍生分类器体系框架, 并结合具有对角平滑参数矩阵的多元 Gaussian 核函数、时序递进分类准确性标准、平滑参数配置树和分类器选择与平均等建立了适合于小

时间序列 (多时间序列) 的动态完全 Bayesian 集成分类器. 动态完全 Bayesian 分类器能够更有效地提取小时间序列数据集中的时序与非时序分类信息; 采用多元 Gaussian 核函数估计属性联合密度和时序递进分类准确性评价标准能使 DFBEC 与小时间序列数据集充分拟合, 而结合为多元 Gaussian 核函数选择对角平滑参数矩阵和平滑参数配置树的构建可避免 DFBEC 与数据的过度拟合 (对小时间序列数据分类至关重要); 通过分类器选择与平均又使 DFBEC 具有良好的泛化性能.

使用宏观经济时间序列进行实验的结果显示, 基于 DFBEC 的小时间序列转折点预测具有良好的准确性. 进一步的研究是将类变量的时序影响、属性变量的时序影响与平滑参数的调整相结合建立平滑参数配置树, 以及探索基于可重复搜索的平滑参数配置森林, 来提高小时间序列 (多时间序列) 分类的可靠性.

参考文献

1 Wang S C, Du R J, Liu Y. The learning and optimization of full Bayes classifiers with continuous attributes. Chinese J Comput, 2012, 35: 2129–2138 [王双成, 杜瑞杰, 刘颖. 连续属性完全贝叶斯分类器的学习与优化. 计算机学报, 2012, 35: 2129–2138]

2 Chow C K, Liu C N. Approximating discrete probability distributions with dependence trees. IEEE Trans Inf Theory, 1968, 14: 462–467

3 Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn, 1997, 29: 131–161

4 Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn, 1997, 29: 103–130

5 Campos C P D, Corani G, Scanagatta M, et al. Learning extended tree augmented naive structures. Int J Approx Reason, 2016, 68: 153–163

6 Cheng J, Greiner R. Comparing Bayesian network classifiers. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI-99), San Francisco, 1999. 101–108

7 Acid S, de Campos L M, Castellano J G. Learning Bayesian network classifiers: searching in a space of partially directed acyclic graphs. Mach Learn, 2005, 59: 213–235

8 Yager R R. An extension of the naive Bayesian classifier. Inf Sci, 2006, 176: 577–588

9 Webb G I, Boughton J, Wang Z. Not so naïve Bayes: aggregating one-dependence estimators. Mach Learn, 2005, 58: 5–24

10 Wang S C, Liu X H, Tang H Y. The learning and optimizing of Markov network classifiers based on dependency analysis. Pattern Recogn Artif Inteligence, 2006, 19: 485–490 [王双成, 刘喜华, 唐海燕. 基于依赖分析的马尔科夫网络分类器学习与优化. 模式识别与人工智能, 2006, 19: 485–490]

11 Wang S C, Xu G L, Du R J. Restricted Bayesian classification networks. Sci China Inf Sci, 2013, 56: 078105

12 Flores M J, Gámez J A, Martínez A M. Domains of competence of the semi-naive Bayesian network classifiers. Inf Sci, 2014, 260: 120–148

13 Daniel B, Aryeh K. A finite sample analysis of the naive Bayes classifier. J Mach Learn Res, 2015, 16: 1519–1545

14 John G H, Langley P. Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI-1995), San Francisco, 1995. 338–345

15 Pérez A, Larrañagaa P, Inzaa I. Supervised classification with conditional Gaussian networks: increasing the structure complexity from naive Bayes. Int J Approx Reason, 2006, 43: 1–25

16 Pérez A, Larranga P, Inza I. Bayesian classifiers based on kernel density estimation: flexible classifiers. Int J Approx Reason, 2009, 50: 341–362

17 He Y L, Wang R, Kwong S, et al. Bayesian classifiers based on probability density estimation and their applications to simultaneous fault diagnosis. Inf Sci, 2014, 259: 252–268

18 Luis G, Eduardo G P, Ramsés H M. Bayesian nonparametric classification for spectroscopy data. Comput Stat Data Anal, 2014, 78: 56–68

19 Xiang Z L, Yu X R, Kang D K. Experimental analysis of naïve Bayes classifier based on an attribute weighting framework with smooth kernel density estimations. Appl Intel, 2016, 44: 611–620

20 Wang S C, Gao R, Wang L M. Bayesian network classifiers based on Gaussian kernel density. Expert Syst Appl, 2016, 51: 207–217
21 Wang S C, Gao R, Du R J. Restricted Bayesian network classifier based on Gaussian Copula. Chinese J Comput, 2016, 39: 1612–1625 [王双成, 高瑞, 杜瑞杰. 基于高斯 Copula 的约束贝叶斯网络分类器研究. 计算机学报, 2016, 39: 1612–1625]
22 Wang S C, Gao R, Du R J. Restricted Gaussian classification network. Acta Autom Sin, 2015, 41: 2128–2140 [王双成, 高瑞, 杜瑞杰. 约束高斯分类网研究. 自动化学报, 2015, 41: 2128–2140]
23 Dong W Y, Zhou M C. Gaussian classifier-based evolutionary strategy for multimodal optimization. IEEE Trans Neural Netw Learn Syst, 2014, 25: 1200–1216
24 Martínez M, Sucar L E. Learning dynamic naive bayesian classifier. In: Proceedings of the 21st International Florida Artificial Intelligence Research Symposium (FLAIRS-21), San Francisco, 2008. 655–659
25 Palacios-Alonso M A, Brizuela C A, Sucar L E. Evolutionary learning of dynamic naive Bayesian classifiers. J Autom Reason, 2009, 45: 21–37
26 Arriaga A, Sucarsuccar H H, Mendozadurán L E, et al. A comparison of dynamic naive Bayesian classifiers and hidden Markov models for gesture recognition. J Appl Res Technol, 2011, 9: 81–102
27 Wang S C, Zhang J F, Wang H. The method of dynamic naive Bayesian classifier for impact analysis of China's economic growth. ICIC Express Lett Part B Appl, 2013, 4: 7–12
28 Alkhateeb J H, Pauplin O, Ren J, et al. Performance of hidden Markov model and dynamic Bayesian network classifiers on handwritten Arabic word recognition. Knowl-Based Syst, 2011, 24: 680–688
29 Yu B, Mark T. Learning gene regulations from multiple knockout data via an efficient dynamic Bayesian network reconstruction. Bioph J, 2011, 100: 311–322
30 Kafai M, Bhanu B. Dynamic Bayesian networks for vehicle classification in video. IEEE Trans Ind Inf, 2012, 8: 100–109
31 Wang S C, Bi Y J, Pei Z. The method of dynamic Bayesian network classifiers for impact analysis on China import and export of goods. Chinese J Manage Sci, 2011, 19: 625–629 [王双成, 毕玉江, 裴瑱. 商品进出口影响分析的动态贝叶斯网络分类器方法. 中国管理科学, 2011, 19: 625–629]
32 Wang S C, Pei Z, Bi Y J. Dynamic Bayesian network classifier model for predicting the cyclical turning points of economic fluctuation. J Ind Eng Eng Manage, 2011, 25: 173–177 [王双成, 裴瑱, 毕玉江. 经济周期转折点预测的动态贝叶斯网络分类器模型. 管理工程学报, 2011, 25: 173–177]
33 Wang S C, Gao R, Du R J. Learning and optimization of dynamic naive Bayesian classifiers for small time series. Control Decis, 2017, 32: 163–166 [王双成, 高瑞, 杜瑞杰. 小时间序列动态朴素贝叶斯分类器学习与优化. 控制与决策, 2017, 32: 163–166]
34 Heckerman D. Bayesian networks for data mining. Data Min Knowl Discov, 1997, 1: 79–119
35 Friedman N, Murphy K P, Russell S. Learning the structure of dynamic probabilistic networks. In: Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence, Madison, 1998. 139–147

Dynamic full Bayesian ensemble classifiers for small time series

Shuangcheng WANG[1*], Fei ZHENG[1] & Rui GAO[2]

1. *School of Information Management, Shanghai Lixin University of Accounting and Finance, Shanghai,* 201620*, China*;
2. *School of Statistic and Mathematics, Shanghai Lixin University of Accounting and Finance, Shanghai,* 201620*, China*
* Corresponding author. E-mail: wangsc@lixin.edu.cn

Abstract Improving the reliability of small time series classifiers with continuous attributes is an important and challenging task. The information contained in small time series is not sufficient and a temporal dependency exists between data records, which makes it very difficult to optimize the fitting degree between the classifier and the data, and many mature techniques of non-time series data classifiers are not practical. We use a dynamic full Bayesian classifier to increase the amount of information provided by the attribute to the class, and realize the fusion of temporal and nonsequential information. By combining the conditional joint density estimation of attributes based on the multivariate Gaussian kernel function with a diagonal smoothing parameter matrix, the interval division of smoothing parameter values, the timing progressive classification accuracy criterion, the construction of the smoothing parameter configuration tree, classifier selection and averaging, etc, a dynamic full Bayesian ensemble classifier was established for small time series. Experiments were performed using small time series in macroeconomic analysis. The results show that the optimized dynamic full Bayesian ensemble classifiers have very good classification accuracy.

Keywords dynamic full Bayesian classifiers, multivariate Gaussian kernel function, smoothing parameter, classification accuracy, classifier selection and averaging

Shuangcheng WANG received his B.S. and M.S. degrees in mathematics from Northeast Normal University, Changchun, China, in 1983 and 1994, respectively, and Ph.D. degree in communication and information system from Jilin University, Changchun, China, in 2004. He has been with the School of Information Management, Shanghai Lixin University of Accounting and Finance, Shanghai, China, since 2004, where he is currently a Professor of computer applications. His current research interests include artificial intelligence, machine learning, and data mining.

Rui GAO received his B.S. degree in mathematics from Taiyuan Normal University, Taiyuan, China; M.S. degree in statistics from University of Shanghai for Science and Technology, Shanghai, China; and Ph.D. degree in statistics from Shanghai University of Finance and Economics, China, in 2003, 2006, and 2015, respectively. She is currently a lecturer at the School of statistics and mathematics, Shanghai Lixin University of Accounting and Finance, Shanghai, China. Her current research interests include applied statistics and data mining.

Fei ZHENG received his B.S. degree in Computer Science from Tsinghua University, Peking, China, in 1989, the M.S. degree and Ph.D degree in Computer Software from Shanghai Jiao Tong University, Shanghai, China, in 1992 and 1998, respectively. He has been with the School of Information Management, Shanghai Lixin University of Accounting and Finance, Shanghai, China, since 2011, where he is currently an Associate Professor of computer applications. His current research interests include data mining, machine learning, and information security.