连续属性完全贝叶斯分类器的学习与优化

王双成[1),2)] 杜瑞杰[1)] 刘 颖[1)]

[1)](上海立信会计学院数学与信息学院 上海 201620)

[2)](上海立信会计学院开放经济与贸易研究中心 上海 201620)

摘 要 针对连续属性朴素贝叶斯分类器不能有效利用属性之间的条件依赖信息,而依赖扩展又很难实现属性条件联合密度估计和结构学习协同优化的问题,文中在使用多元高斯核函数估计属性条件联合密度的基础上,建立了具有多平滑参数的连续属性完全贝叶斯分类器,并给出将分类准确性标准与区间异步长划分完全搜索相结合的平滑参数优化方法,再通过时序扩展构建了动态完全贝叶斯分类器. 我们使用 UCI 机器学习数据仓库中连续属性分类数据和宏观经济数据进行实验,结果显示,经过优化的两种分类器均具有良好的分类准确性.

关键词 连续属性;完全贝叶斯分类器;动态完全贝叶斯分类器;高斯核函数;平滑参数

中图法分类号 TP181 **DOI 号**: 10.3724/SP.J.1016.2012.02129

The Learning and Optimization of Full Bayes Classifiers with Continuous Attributes

WANG Shuang-Cheng[1),2)] DU Rui-Jie[1)] LIU Ying[1)]

[1)](*School of Mathematics and Information, Shanghai Lixin University of Commerce, Shanghai* 210620)

[2)](*Open Economic and Trade Research Center, Shanghai Lixin University of Commerce, Shanghai* 210620)

Abstract The naive Bayes classifiers with continuous attributes can not make the effective use of conditional dependency information between attributes. In dependency extension of naive Bayes classifiers, it is very difficult that the optimization of attribute conditional joint density estimation and structure learning of classifiers are integrated. In this paper, on the basis of using multivariate Gaussian kernel function to estimate the conditional joint density of attributes, a full Bayes classifier with continuous attributes and multi smoothing parameters is presented. The smoothing parameters are optimized by combining the evaluation criteria of classification accuracy and full search method based on interval division with asynchronous long. A dynamic full Bayes classifier is also developed by combining full Bayes classifier with time series. Experiment and analysis are done by using data sets with continuous attributes in UCI machine learning repository and macroeconomic field. The results show that two kinds of optimized classifiers have very good classification accuracy.

Keywords continuous attributes; full Bayes classifiers; dynamic full Bayes classifiers; Gaussian kernel function; smoothing parameters

1 引 言

贝叶斯分类器是一个基础概率分类器,它使用满条件分布进行分类. 可以在理论上证明贝叶斯分类器是最优分类器,但直接使用这种分类器进行分类比较困难,需要对所依据的满条件分布进行转化或增加一些约束条件来提高运算效率和计算的可行

收稿日期:2012-06-30;最终修改稿收到日期:2012-08-10. 本课题得到国家自然科学基金(11101284)、教育部人文社科基金(10YJA630154,12YJA630123)、上海市教委重点学科建设项目(J51702)及上海市教委科研创新项目(11YZ240)资助.**王双成**,男,1958 年生,博士,教授,主要研究领域为人工智能、机器学习、数据挖掘与应用. E-mail: wangsc@lixin.edu.cn. **杜瑞杰**,女,1980 年生,博士,讲师,主要研究方向为机器学习与数据挖掘. **刘 颖**,女,1980 年生,博士,副教授,主要研究方向为图形理论和机器学习.

性,这样便产生一系列贝叶斯分类器的衍生分类器.朴素贝叶斯分类器(naive Bayes classifiers,简记为NBC)是最简单的衍生分类器,它以高效率和良好的分类准确性而著称,被广泛用于医疗诊断、文本分类、邮件过滤和信息检索等.NBC基于一个很强的条件独立性假设,这使得属性之间的条件依赖信息无法得到有效的利用,而这部分信息往往也是分类的重要信息.鉴于此,对NBC的依赖扩展便成为贝叶斯分类器衍生分类器的重要研究内容.其中对离散属性NBC的依赖扩展研究较多,如Chow和Liu[1]的依赖树、Friedman和Geiger等人[2]的TAN(Tree augmented naive Bayes)分类器、Grossman和Domingos[3]的基于条件似然打分-搜索的贝叶斯网络分类器学习,Jing和Pavlović等人[4]对TAN分类器的属性选择和参数集成、Webb和Boughton等人[5]对NBC的k阶依赖扩展的理论分析和对比实验(如果综合考虑分类器偏差、方差和学习效率,认为二阶依赖扩展分类器具有最好的性能)等.对于连续属性NBC,可采用两种处理方法:一种是连续属性的离散化,最终将其转化为离散属性的分类器问题;另一种是不离散化连续属性,但需要估计属性条件密度.两种方法各有优势和不足,第1种方法适用于具有较少类的大数据集,以保证属性条件概率得到可靠的估计;第2种方法更适合于多类较小数据集(估计属性条件密度不需要很多数据)的情况,能够避免由离散化所导致的信息丢失、引入噪声和类对属性的变化不够敏感等问题.连续属性NBC研究的两个核心问题是属性条件密度估计和属性之间条件依赖信息的利用.在属性条件密度估计方面,John和Langley[6]研究了使用经典的高斯函数和高斯核函数估计属性条件密度而得到的Gaussian Naive Bayes Classifier和Flexible Bayes Classifier,并在UCI机器学习数据仓库中选择了一些具有连续属性的数据集,对这两个分类器与C4.5进行了分类准确性比较,结果基于高斯核函数的分类器优于高斯函数分类器,但不如C4.5,其主要原因是:使用高斯函数来估计属性条件密度可能与实际密度有较大的差距,而采用高斯核函数的估计又没有对拟合数据的程度进行控制,再有连续属性之间的依赖信息也得不到有效的利用,这些都会影响分类器的分类准确性.Pérez和Larranga等人[7-8]在John和Langley研究的基础上,为用于估计属性条件密度的高斯核函数引入单平滑参数(smoothing parameter),并使用MISE(Mean Integrated Square Error)统计标准优化平滑参数,将经过平滑参数优化的分类器称为Flexible Naive Bayes Classifier,实验结果显示,这种分类器的分类准确性优于Flexible Bayes Classifier和C4.5,其原因是平滑参数的优化使所估计的属性条件密度更接近于真实的密度,从而提高了分类器的分类准确性.Huang[9]对基于高斯核函数估计属性条件密度的NBC与支持向量机进行了比较,发现经过优化的NBC的分类准确性优于支持向量机,并将NBC用于信用风险预测.在连续属性之间依赖信息利用方面,李旭升和郭春香等人[10]基于似然打分对连续属性NBC进行了树结构的依赖扩展,使扩展后分类器的分类准确性有所改进.Pérez和Larranga等人通过连续属性的互信息计算对NBC进行依赖扩展,分别给出了对NBC进行树、k依赖和完全依赖扩展而得到的分类器,但在依赖扩展过程中,属性之间是否增加边取决于属性之间条件互信息的大小,这与对分类的贡献大小可能不一致,从而会影响分类准确性.

目前,对连续属性贝叶斯分类器的衍生分类器虽有一些研究,但所建立分类器的分类准确性还有待提高.

本文的主要贡献如下:

(1)在0-1损失下证明连续属性贝叶斯分类器是最优分类器,并给出贝叶斯分类器衍生分类器家族的构成.

(2)将多元高斯核函数用于属性条件联合密度估计,建立与贝叶斯分类器等价的连续属性完全贝叶斯分类器(Full Bayes Classifier,FBC).

(3)为连续属性完全贝叶斯分类器的每一个属性引入一个平滑参数,并给出平滑参数优化方法,从而实现属性之间依赖信息利用和属性条件联合密度估计优化的统筹兼顾.

(4)将连续属性完全贝叶斯分类器与时间序列相结合,建立连续属性动态完全贝叶斯分类器(Dynamic Full Bayes Classifier,DFBC),并将其用于宏观经济指标波动转折点预测.

我们使用UCI机器学习数据仓库①中连续属性分类数据和宏观经济数据进行实验,结果显示,经过优化的连续属性完全贝叶斯分类器和动态完全贝叶斯分类器均具有良好的分类准确性.

本文第1节对贝叶斯分类器的衍生分类器的研

① Murphy S L, Aha D W. UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.Html, 2011

究现状进行评述；第 2 节给出连续属性贝叶斯分类器的最优性证明和贝叶斯分类器的衍生分类器家族的构成；第 3 节在使用多元高斯核函数估计属性条件联合密度的基础上，建立连续属性完全贝叶斯分类器和优化方法；第 4 节结合连续属性完全贝叶斯分类器与时间序列构建动态完全贝叶斯分类器；第 5 节是实验和分析；第 6 节是结论和进一步的工作.

2　连续属性贝叶斯分类器

用 $X_1,X_2,\cdots,X_n,C$ 表示连续属性和类，$x_1,x_2,\cdots,x_n,c$ 为其值，D 是具有 N 个记录的数据集，数据随机产生于混合分布 P，$x_{im}(1\leq i\leq n,1\leq m\leq N)$和 c_m表示 X_i和 C 在数据集 D 中第 m 个记录的观测值.

用 $c^F(x_1,x_2,\cdots,x_n)$表示分类器 F 的分类结果，$c(x_1,x_2,\cdots,x_n)$为真正的结果.

定义 1.　对连续属性分类器 F，称

$$CR(F)=\int_{x_1}\cdots\int_{x_n}p(x_1,x_2,\cdots,x_n)\,p(c^F(x_1,x_2,\cdots,x_n)\neq c(x_1,x_2,\cdots,x_n))\mathrm{d}x_1\cdots\mathrm{d}x_n \tag{1}$$

为 F 的平均 0-1 损失(或风险)，使 $CR(F)$最小的分类器称为最优分类器.

定义 2.　对概率分布 $p(c,x_1,x_2,\cdots,x_n)$，称使用满条件分布 $p(c|x_1,x_2,\cdots,x_n)$进行分类的分类器

$$\arg\max_{c(x_1,x_2,\cdots,x_n)}\{p(c|x_1,x_2,\cdots,x_n)\} \tag{2}$$

为贝叶斯分类器.

定理 1.　在 0-1 损失下，连续属性贝叶斯分类器是最优分类器.

证明.　设 F 和 F^* 分别为任意的分类器和连续属性贝叶斯分类器，对给定的情况 $x_1,x_2,\cdots,x_n$，类 C 的可能取值为 $c^1,c^2,\cdots,c^{r_c}$，记 $p^i=p(c^i|x_1,x_2,\cdots,x_n)$，$p^*=\max\limits_{1\leq i\leq r_c}\{p^i\}$，则有

$$\begin{aligned}p(c^{F^*}(x_1,x_2,\cdots,x_n)=c(x_1,x_2,\cdots,x_n))&=\max_{c(x_1,x_2,\cdots,x_n)}\{p(c|x_1,x_2,\cdots,x_n)\}\\&=p^*,\\p(c^F(x_1,x_2,\cdots,x_n)=c(x_1,x_2,\cdots,x_n))&=p(\arg\max_{c(x_1,x_2,\cdots,x_n)}\{p(c^F(x_1,x_2,\cdots,x_n))\})\\&=p^{i_0}\in\{p^1,p^2,\cdots,p^{r_c}\},\end{aligned}$$

因此，$p(c^{F^*}(x_1,x_2,\cdots,x_n)=c(x_1,x_2,\cdots,x_n))\geq p(c^F(x_1,x_2,\cdots,x_n)=c(x_1,x_2,\cdots,x_n))$，从而 $CR(c^{F^*}(x_1,x_2,\cdots,x_n))\leq CR(c^F(x_1,x_2,\cdots,x_n))$，所以连续属性贝叶斯分类器是最优分类器.　证毕.

虽然连续属性贝叶斯分类器在理论上是最优分类器，但直接计算满条件概率 $p(c|x_1,x_2,\cdots,x_n)$非常困难. 根据概率公式，可得

$$\begin{aligned}p(c|x_1,x_2,\cdots,x_n)&=\frac{p(c,x_1,x_2,\cdots,x_n)}{p(x_1,x_2,\cdots,x_n)}\\&=\frac{p(c)p(x_1,x_2,\cdots,x_n|c)}{p(x_1,x_2,\cdots,x_n)}\\&=\alpha p(c)p(x_1,x_2,\cdots,x_n|c)\end{aligned} \tag{3}$$

其中 α 是与 C 无关的量. 式(3)将满条件概率 $p(c|x_1,x_2,\cdots,x_n)$计算转化为类先验概率 $p(c)$与属性条件联合密度 $p(x_1,x_2,\cdots,x_n|c)$的计算问题. 由对 $p(x_1,x_2,\cdots,x_n|c)$计算方式的不同便产生了下面图 1 中列出的各种贝叶斯分类器的衍生分类器.

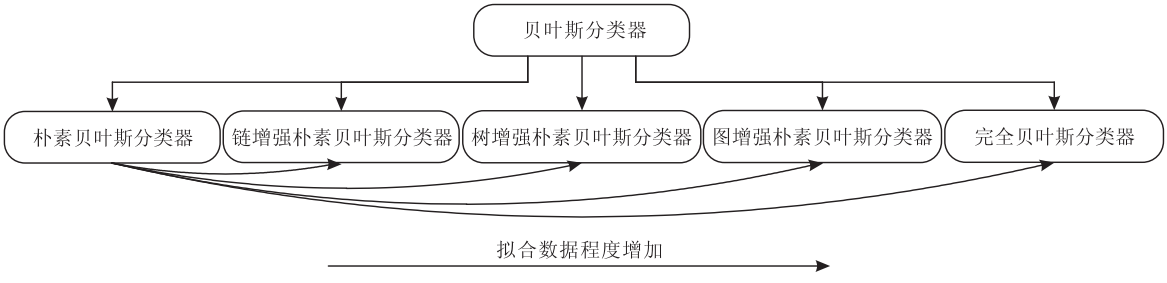


图 1　贝叶斯分类器和它的衍生分类器

在 NBC 结构中，属性结点除类父结点外没有属性父结点(给定类时，属性之间条件独立)，这样可以得到 $p(x_1,x_2,\cdots,x_n|c)=\prod\limits_{i=1}^{n}p(x_i|c)$；当属性结点最多只能有一个属性父结点和一个子结点时，$p(x_1,x_2,\cdots,x_n|c)=\prod\limits_{i=1}^{n}p(x_i|\psi(x_i),c)$，其中 $\psi(x_i)$是 X_i的属性父结点 $\Psi(X_i)$的取值，得到的是链增强 NBC；当属性结点最多只能有一个属性父结点，但可以有多个子结点时，$p(x_1,x_2,\cdots,x_n|c)=\prod\limits_{i=1}^{n}p(x_i|$

$\pi(x_i),c)$，其中 $\pi(x_i)$ 是 X_i 的属性父结点 $\Pi(X_i)$ 的取值，得到的是树增强 NBC；当属性结点可以有多个属性父结点和子结点时（不能产生有向环），$p(x_1,x_2,\cdots,x_n|c)=\prod_{i=1}^{n}p(x_i|\pi_i,c)$，其中 π_i 是 X_i 的属性父结点集 Π_i 的配置，得到的是图（或网络）增强 NBC；当属性结点之间构成完全有向无环图（不考虑条件独立关系），$p(x_1,x_2,\cdots,x_n|c)=\prod_{i=1}^{n}p(x_i|x_1,x_2,\cdots,x_{i-1},c)$，得到的是 FBC，这种分类器能够充分利用属性之间的条件依赖信息，但易于导致对数据的过度拟合，需要对分类器与数据的拟合程度进行控制.

3 连续属性 FBC

定义 3. 对概率分布 $p(c,x_1,x_2,\cdots,x_n)$，称分类器

$$\underset{c(x_1,x_2,\cdots,x_n)}{\arg\max}\{p(c,x_1,x_2,\cdots,x_n)\}$$

或

$$\underset{c(x_1,x_2,\cdots,x_n)}{\arg\max}\{p(c)p(x_1,x_2,\cdots,x_n|c)\} \quad (4)$$

为完全贝叶斯分类器.

推论 1. 在 0-1 损失下，连续属性完全贝叶斯分类器是最优分类器.

FBC 不施加任何条件独立性假设，这使得该分类器的结构是一个完全有向无环图（所有的完全有向无环图等价），如图 2 所示.

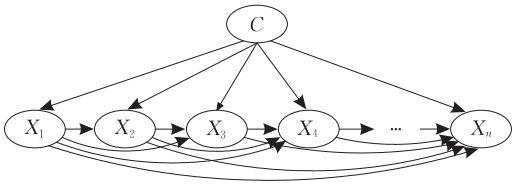


图 2 FBC 结构

与 NBC 一样，FBC 也不需要结构学习，建立连续属性 FBC 的核心是属性条件联合密度估计与优化.

3.1 属性条件联合密度估计

我们采用统计多元核函数方法[11]来估计属性条件联合密度，并在多元核函数中为每个属性引入一个平滑参数，通过平滑参数的调整来控制分类器与数据的拟合程度.

用 $\hat{p}(x_1x_2\cdots x_n|c,D)$ 表示在数据集 D 基础上的属性条件联合密度估计，基于统计多元核函数的属性条件联合密度估计一般形式为

$$\hat{p}(x_1x_2\cdots x_n|c,D)=\frac{1}{N(c)h_1h_2\cdots h_n}\sum_{m=1}^{N}\left[signa(c_m)\prod_{i=1}^{n}K_i\left(\frac{x_i-x_{im}}{h_i}\right)\right] \quad (5)$$

其中 $K_i(\cdot)$ 和 $h_i(i=1,2,\cdots,n)$ 分别是 X_i 的核函数和平滑参数，$signa(c_m)=\begin{cases}1, & c_m=c\\ 0, & c_m\neq c\end{cases}$.

我们取 $K_i(\cdot)$ 为高斯核函数（也可以取其它的核函数），即 $K_i\left(\frac{x_i-x_{im}}{h_i}\right)=g(x_i;x_{im},h_i)$，$g(x_i;x_{im},h_i)=\frac{1}{\sqrt{2\pi}h_i}\exp\left[-\frac{(x_i-x_{im})^2}{2h_i^2}\right]$，那么

$$\hat{p}(x_1x_2\cdots x_n|c,D)=\frac{1}{N(c)h_1h_2\cdots h_n}\sum_{m=1}^{N}\left[signa(c_m)\prod_{i=1}^{n}g(x_i;x_{im},h_i)\right] \quad (6)$$

其中 $N(c)$ 是训练集中 $C=c$ 的情况数量.

结合 $\hat{p}(c)=\frac{N(c)}{N}$，具有多平滑参数的完全贝叶斯分类器对 $N+1$ 情况的分类结果为

$$\underset{c(x_{1(N+1)},x_{2(N+1)},\cdots,x_{n(N+1)})}{\arg\max}\left\{\sum_{m=1}^{N}\left[\prod_{i=1}^{n}signa(c_m)g(x_{i(N+1)};x_{im},h_i)\right]\right\} \quad (7)$$

除采用多元高斯核函数进行属性条件联合密度估计外，还可以使用多元高斯函数、扩展的多元高斯核函数和高斯 Copula 函数（从计算效率和可靠性方面考虑，扩展的高斯核函数中的平滑参数矩阵和高斯 Copula 函数中的协方差矩阵可采用三对角对称矩阵）等来估计属性条件联合密度，在实验部分分别给出了几种属性条件联合密度估计方法的对比实验.

3.2 属性条件联合密度优化

基于多元高斯核函数估计属性条件联合密度的优化有许多统计方法（但需要某种分布的假设，如联合高斯分布等），我们通过对平滑参数的打分-搜索来实现属性条件联合密度的优化. 以分类器的分类准确性为打分标准，搜索策略采用区间异步长划分完全搜索.

分类准确性估计采用 10 折交叉有效性（10-fold cross-validation）验证方法，Kohavi[12]曾对各种常用的分类准确性估计方法进行过综合实验与统计分析，认为基于 10 折交叉有效性验证方法所进行的分类准确性估计更加可靠. 平滑参数决定着做叠加的高斯函数曲线形状，因此能够控制所估计的属性条件联合密度与数据的拟合程度. 随着平滑参数接近于零，所估计的属性条件联合密度会更加拟合数

据(也称为噪声估计)，并会产生过度拟合现象(欠平滑)；而随着平滑参数的增大，将使所估计的属性条件联合密度逐渐趋近于真实密度，直到取得最优平滑参数(具有最佳分类效果的平滑参数). 如果平滑参数持续增大，所估计的属性条件联合密度对数据的拟合程度将继续下降，以至于会出现欠拟合(过平滑)现象.

对于单平滑参数($h=h_1=h_2=\cdots=h_n$)的完全贝叶斯分类器，采用区间异步长划分完全搜索方法发现最优平滑参数. 一般根据实验确定平滑参数的界值 $h_{\min}$ 和 $h_{\max}$，十进位点将 $[h_{\min},h_{\max}]$ 分成一些子区间，在这些子区间中的步长依次记为 $\Delta_1,\Delta_2,\cdots$(为提高效率，步长可逐渐增加)，这样便可得到平滑参数的取值集合，通过以分类准确性为标准的遍历打分-搜索来发现最优平滑参数. 对于具有多平滑参数的情况，首先根据 Quinlan[13] 的信息增益率为属性排序，其中的条件密度计算采用高斯核函数，平滑参数使用 John 和 Langley 的方法进行设置. 采用单平滑参数的优化方法依次进行多平滑参数的优化(将单参数最优值作为多参数的初始配置)，最终获得所有平滑参数的局部最优配置.

3.3　时间复杂性分析

建立 FBC 的主要运算是计算高斯函数，因此，对确定的平滑参数 h_i，属性条件联合密度估计需要进行 Nn 次高斯函数计算. 用 M 表示单平滑参数所有可能的取值数量，那么，在多平滑参数优化的过程中需要进行 MNn^2 次高斯函数计算，而 M 是一个与 N 和 n 都无关的量，可以看做是一个常量，因此，建立最优多参数 FBC 的时间复杂度是 $O(Nn^2)$.

4　连续属性 DFBC

DFBC 是 FBC 与时间序列的结合，能够有效利用类的动态时序信息和时间片内属性之间的依赖信息，是多变量时间序列预测的有力工具. 分别用 $X_i[1],X_i[2],\cdots,X_i[T]$ $(1\leq i\leq n)$ 和 $C[1],C[2],\cdots,C[T]$ 表示属性和类序列，$x_i[1],x_i[2],\cdots,x_i[T]$ 和 $c[1],c[2],\cdots,c[T]$ 是具体的取值；$D[1],D[2],\cdots,D[T]$ 是累计时间片数据集序列，$D[1]\subset D[2]\subset\cdots\subset D[T]$，$N[1],N[2],\cdots,N[T]$ 是对应时序数据集中的例子数量.

4.1　分类器结构和表示形式

在 DFBC 结构中，类时间序列构成马尔科夫链，给定一个时间片内的类时，所属的时间片属性与其它时间片内的属性和类条件独立，图 3 给出的是 DFBC 结构.

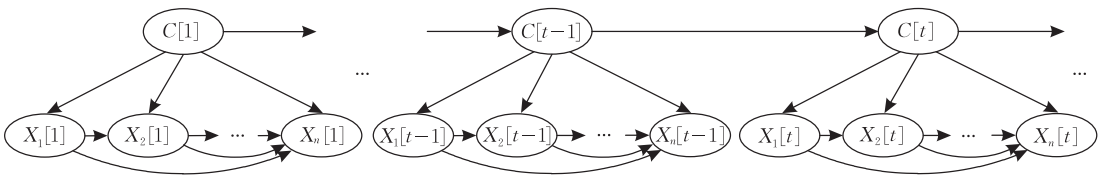


图 3　DFBC 结构

依据贝叶斯网络理论、概率公式和图 3 中所蕴含的条件独立性关系可得

$$\begin{aligned}&p(c[t]|c[1],c[2],\cdots,c[t-1],x_1[1],x_2[1],\cdots,\\&\qquad x_n[1],\cdots,x_1[t],x_2[t],\cdots,x_n[t])\\&=p(c[t]|c[t-1],x_1[t],x_2[t],\cdots,x_n[t])\\&=\frac{p(c[t],c[t-1],x_1[t],x_2[t],\cdots,x_n[t])}{p(c[t-1],x_1[t],x_2[t],\cdots,x_n[t])}\\&=\beta p(c[t]|c[t-1])\,p(x_1[t],x_2[t],\cdots,x_n[t]|c[t])\end{aligned}\tag{8}$$

其中 β 是与 $c[t]$ 无关的量.

DFBC 可表示为

$$\begin{aligned}&\underset{c[t](c[t-1],x_1[t],x_2[t],\cdots,x_n[t])}{\arg\max}\\&\{p(c[t]|c[t-1])\,p(x_1[t],x_2[t],\cdots,x_n[t]|c[t])\}\end{aligned}\tag{9}$$

其中，$p(c[t]|c[t-1])$ 是类转移概率，$p(x_1[t],x_2[t],\cdots,x_n[t]|c[t])$ 为时间片属性条件联合密度.

4.2　分类准确性评价标准

设有时序数据 $x_1[1],x_2[1],\cdots,x_n[1],c[1],\cdots,x_1[T],x_2[T],\cdots,x_n[T],c[T]$，选择一个界值 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 $accuracy(dfbc,\boldsymbol{\rho},D[T],T_0)$ 表示 DFBC 的分类准确率，$\boldsymbol{\rho}=(\rho_1,\rho_2,\cdots,\rho_n)$，$c_{prediction}[t]$ 是使用 $D[t-1]$ 进行训练，并依据 $x_1[t],x_2[t],\cdots,x_n[t]$ 的配置对 $c[t]$ 的预测结果，$c_{true}[t]$ 是真正的结果，那么

$$\begin{aligned}&accuracy(dfbc,\boldsymbol{\rho},D[T],T_0)=\\&\qquad\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}signb(c_{prediction}[t],c_{true}[t])\end{aligned}\tag{10}$$

其中

$$signb(c_{prediction}[t],c_{true}[t])=\begin{cases}1, c_{prediction}[t]=c_{true}[t]\\ 0, c_{prediction}[t]\neq c_{true}[t]\end{cases}.$$

5 实验与分析

首先,在 UCI 机器学习数据仓库中选择 28 个连续属性的分类数据集,删除具有丢失数据的记录,对属性数据进行规范化处理,数据集中记录的位置也进行随机初始化,从不同分类器之间的分类准确性比较和平滑参数对分类准确性的影响两方面进行 FBC 的实验与分析;然后,再使用 3 个宏观经济指标时序数据集进行 DFBC 的实验与分析.

5.1 UCI 数据集描述

在所选择的 UCI 数据集中,对几个较大的数据集顺序截取其中的一部分数据,打 * 号的数据集为经过截取的数据集,数据集的基本情况如表 1 所示.

表 1 UCI 数据集描述

编号	数据集	例子数量	属性数量	类数	编号	数据集	例子数量	属性数量	类数
1	Ae_train*	774	12	9	15	Iris	150	4	3
2	Arabic_digit*	736	13	2	16	Liver_disease	345	6	2
3	Breast_cancer	699	10	2	17	Magic_Gamma_telescope*	718	10	2
4	Breast_tissue	106	9	6	18	New_thyroid	215	5	3
5	Cardiotocography*	726	27	10	19	Parkinsons	195	22	2
6	Cmc	1376	9	2	20	Pima	768	8	2
7	Column_3c	310	6	3	21	Sensor_reading*	456	24	3
8	Connectionist_Bench*	528	10	11	22	Spambase	601	30	2
9	Ecoli	292	5	4	23	Statlog*	1310	16	7
10	Glass	214	9	6	24	Transfusion	748	4	2
11	Heart_disease	270	13	2	25	Wdpc	569	31	2
12	Horse_colic	300	22	2	26	Wine	178	13	3
13	Image_segmentation*	209	16	7	27	Wpbc	198	34	2
14	Ionosphere	349	33	2	28	Yeast	1484	6	4

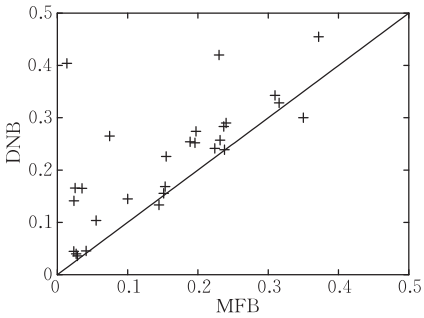
5.2 分类准确性比较

经过实验发现,平滑参数的峰值一般在 0.001～0.1 之间. 取 $h_{\min}=0.001$, $h_{\max}=0.1$, $\Delta_1=0.001$, $\Delta_2=0.005$,分别选择对连续属性离散化的 NBC 和 TAN 分类器(DNB,DTAN)、基于高斯函数估计属性条件密度的分类器(GNB)、John 等人给出的分类器(Flexible Bayes Classifier,FLBC)、Pérez 等人建立的使用 MISE 标准优化平滑参数的分类器(Flexible Naive Bayes Classifier,FNBC)、单参数和多参数优化的朴素贝叶斯分类器(SNB,MNB)、使用高斯核函数进行属性之间条件信息计算的朴素贝叶斯分类器树结构依赖扩展分类器(CTAN)、C4.5、支持向量机(SVM)、基于具有三对角协方差矩阵多元高斯函数的完全贝叶斯分类器(GFBC)、具有三对角平滑参数矩阵的多元高斯核函数完全贝叶斯分类器(KFBC)、具有三对角协方差矩阵 Copula 函数的完全贝叶斯分类器(CFBC)、基于多元高斯核函数的具有单参数和多参数完全贝叶斯分类器(SFB,MFB),采用 10 折交叉有效性验证方法进行分类器的分类错误率估计,分类器的分类错误率实验结果如表 2 所示.

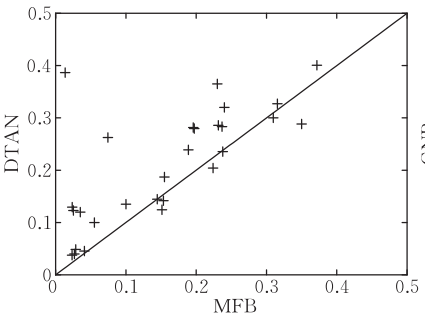
从表 2 的总体平均值来看,经过多参数优化的 FBC 相对于其它 14 个分类器具有优势的程度依次是 8.23%、6.57%、18.97%、17.89%、14.36%、5.98%、2.43%、10.58%、4.87%、22.69%、9.84%、6.89%、4.87%和 3.87%,这显示了经过多参数优化的 FBC 具有良好的分类准确性,在多类值的数据集中尤其如此. 总的来看,FBC 优于 NBC,这说明 FBC 能够有效地利用属性之间的依赖信息,提高了分类器的分类准确性.

使用表 2 中的数据所绘制的 FBC 与其它分类器,关于 28 个数据集的分类错误率比较散点图如图 4 所示. 图中每一个点的坐标是用于比较的两个分类器的分类错误率,在 45°线上方、下方和线上的点分别表示经过多平滑参数优化的完全贝叶斯分类器分类错误率小于、大于和等于用于比较的分类器.

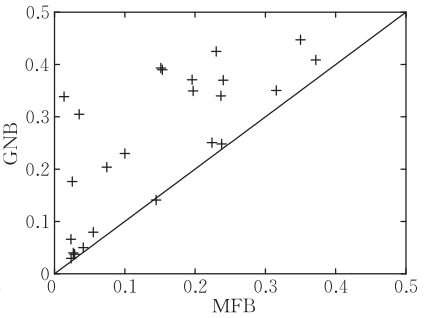
从图 4 中的 14 个比较图来看,多参数 FBC 的分类准确率优于其它分类器,在 28 个数据集中分类准确性占优的百分比依次是 92.85%、75.00%、96.42%、96.42%、100.00%、82.14%、53.57%、85.71%、82.14%、100.00%、85.71%、92.85%、78.57%和 89.28%. 其中具有最小百分比的是 MNB,可见,经过优化的 MNB 也具有良好的分类准确性.



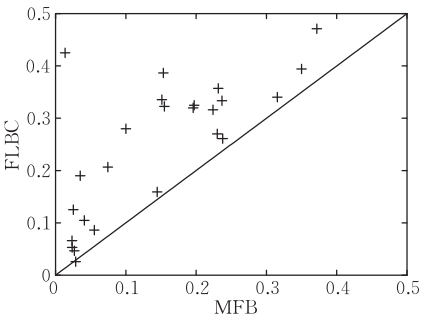
(a) MFB对DNB



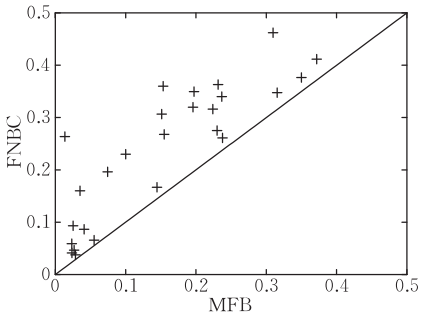
(b) MFB对DTAN



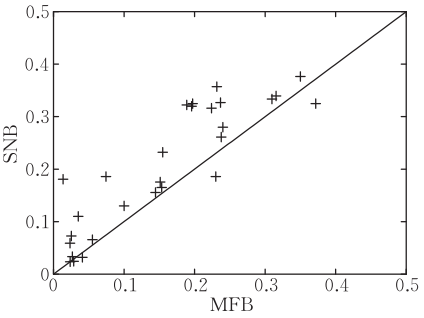
(c) MFB对GNB



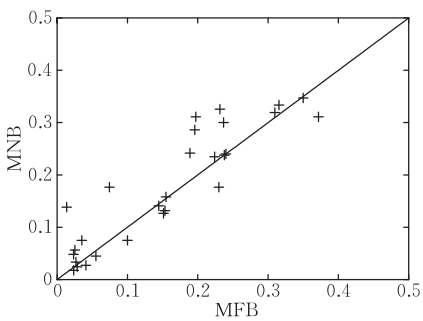
(d) MFB对FLBC



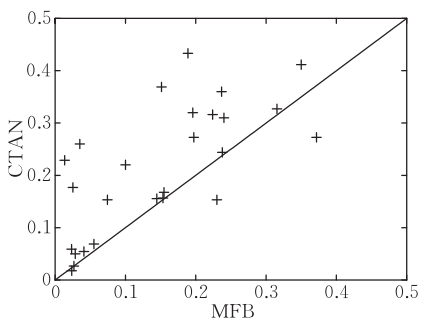
(e) MFB对FNBC



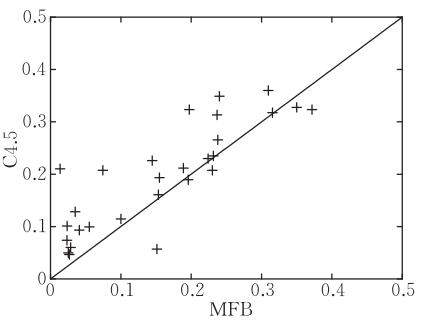
(f) MFB对SNB



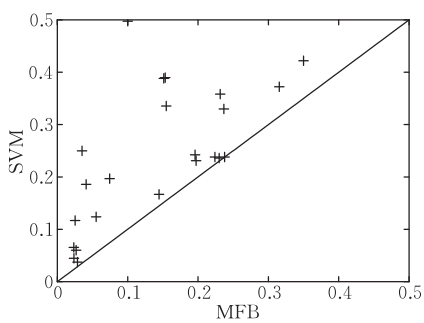
(g) MFB对MNB



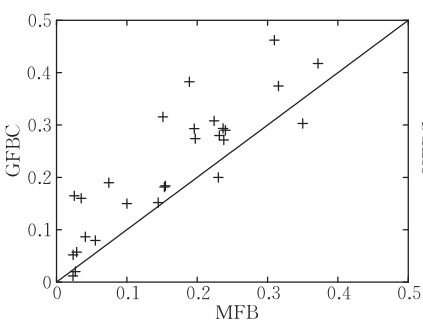
(h) MFB对CTAN



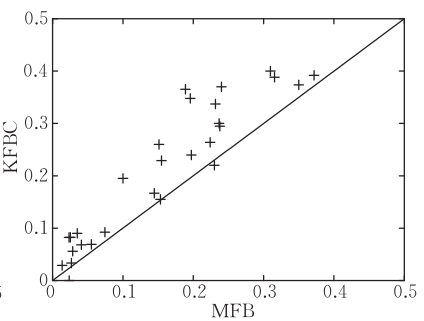
(i) MFB对C4.5



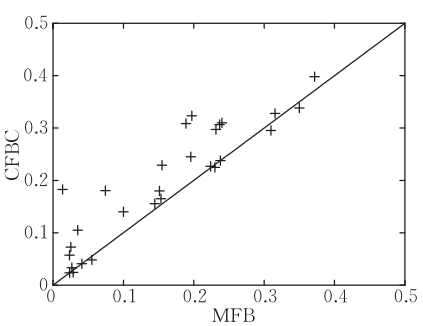
(j) MFB对SVM



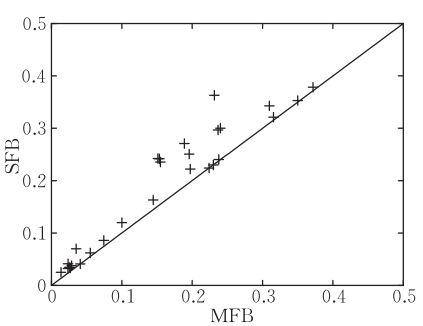
(k) MFB对GFBC



(l) MFB对KFBC



(m) MFB对CFBC



(n) MFB对SFB

图 4　分类错误率比较散点图

表 2 分类器之间的错分率实验结果

数据集	DNB	DTAN	GNB	FLBC	FNBC	SNB	MNB	CTAN	C4.5	SVM	GFBC	KFBC	CFBC	SFB	MFB
Ae_train*	0.2650	0.2624	0.2039	0.2065	0.1962	0.1858	0.1767	0.1532	0.2076	0.1969	0.1897	0.0923	0.1806	0.0858	0.0743
Arabic_Digit*	0.2740	0.2795	0.3494	0.3247	0.3494	0.3247	0.3110	0.2726	0.3233	0.2311	0.2740	0.2398	0.3233	0.2220	0.1973
Breast_cancer	0.0358	0.0486	0.0372	0.0258	0.0377	0.0243	0.0243	0.0500	0.0601	0.0372	0.0572	0.0558	0.0243	0.0377	0.0286
Breast_tissue	0.2900	0.3200	0.3700	0.6100	0.5600	0.2800	0.2400	0.3100	0.3491	0.5936	0.2900	0.3700	0.3100	0.3000	0.2400
Cardiotocography*	0.2542	0.2389	0.5420	0.6292	0.5820	0.3223	0.2417	0.4333	0.2118	0.7158	0.3827	0.3653	0.3084	0.2709	0.1889
Cmc	0.3285	0.3271	0.3504	0.3402	0.3475	0.3395	0.3336	0.3270	0.3176	0.3722	0.3745	0.3884	0.3278	0.3212	0.3154
Column_3C	0.2259	0.1871	0.5162	0.3226	0.2678	0.2323	0.1581	0.1677	0.1935	0.3355	0.1839	0.2291	0.2291	0.2355	0.1549
Connectionist_Bench*	0.4039	0.3866	0.3385	0.4251	0.2635	0.1808	0.1385	0.2288	0.2102	0.6535	0.5424	0.0289	0.1827	0.0250	0.0135
Ecoli	0.1035	0.1000	0.0794	0.0863	0.0656	0.0656	0.0449	0.0689	0.0994	0.1237	0.0794	0.0690	0.0483	0.0621	0.0552
Glass	0.3429	0.3000	0.5096	0.5239	0.4620	0.3334	0.3197	0.5333	0.3599	0.6427	0.4620	0.4000	0.2953	0.3429	0.3096
Heart_disease	0.1334	0.1445	0.1408	0.1593	0.1667	0.1555	0.1408	0.1555	0.2260	0.1667	0.1519	0.1667	0.1556	0.1630	0.1445
Horse_colic	0.2834	0.2834	0.3400	0.3334	0.3400	0.3267	0.3000	0.3600	0.3133	0.3300	0.2934	0.3000	0.3067	0.2967	0.2367
Image_Segmentation*	0.1450	0.1350	0.2300	0.2800	0.2300	0.1300	0.0750	0.2200	0.1145	0.4974	0.1500	0.1950	0.1400	0.1200	0.1000
Ionosphere	0.2572	0.2858	0.5372	0.3572	0.3629	0.3572	0.3258	0.5200	0.2349	0.3581	0.2800	0.3372	0.2972	0.3629	0.2315
Iris	0.0400	0.0400	0.0400	0.0467	0.0467	0.0334	0.0334	0.0267	0.0467	0.0601	0.0200	0.0334	0.0334	0.0334	0.0267
Liver_disease	0.3000	0.2883	0.4471	0.3942	0.3765	0.3765	0.3471	0.4117	0.3276	0.4220	0.3030	0.3736	0.3383	0.3530	0.3500
MAGIC_Gamma	0.2522	0.2817	0.3709	0.3198	0.3198	0.3198	0.2862	0.3198	0.1894	0.2422	0.2930	0.3479	0.2451	0.2508	0.1958
New_thyroid	0.0454	0.0454	0.0500	0.1046	0.0864	0.0319	0.0273	0.0545	0.0931	0.1859	0.0864	0.0682	0.0410	0.0410	0.0410
Parkinsons	0.1650	0.1200	0.3050	0.1900	0.1600	0.1100	0.0750	0.2600	0.1283	0.2500	0.1600	0.0900	0.1050	0.0700	0.0350
Pima	0.2390	0.2356	0.2481	0.2611	0.2611	0.2611	0.2377	0.2441	0.2657	0.2382	0.2715	0.2949	0.2377	0.2403	0.2377
Sensor_readings*	0.1556	0.1245	0.3934	0.3356	0.3067	0.1756	0.1267	0.3688	0.0571	0.3883	0.3156	0.2600	0.1800	0.2423	0.1512
Spambase	0.1684	0.1417	0.3900	0.3867	0.3600	0.1650	0.1317	0.1566	0.1610	0.3898	0.1817	0.1551	0.1650	0.2417	0.1534
Statlog*	0.1657	0.1230	0.1764	0.1252	0.0932	0.0726	0.0565	0.1770	0.0496	0.1168	0.1649	0.0825	0.0726	0.0344	0.0252
Transfusion	0.2414	0.2040	0.2507	0.3160	0.3160	0.3160	0.2347	0.3160	0.2300	0.2382	0.3081	0.2640	0.2267	0.2240	0.2240
Wine	0.1412	0.1295	0.0295	0.0530	0.0412	0.0236	0.0177	0.01774	0.1012	0.0445	0.0518	0.0000	0.0236	0.0412	0.0236
Wdbc	0.0447	0.0375	0.0661	0.0661	0.0590	0.0590	0.0483	0.0589	0.0739	0.0654	0.0118	0.0822	0.0572	0.0322	0.0233
Wpbc	0.4200	0.3650	0.4250	0.2700	0.2750	0.1858	0.1767	0.1532	0.2076	0.2362	0.2000	0.2200	0.2250	0.2300	0.2300
Yeast	0.4548	0.4007	0.4088	0.4710	0.4115	0.3247	0.3110	0.2726	0.3233	0.6155	0.4176	0.3919	0.3980	0.3784	0.3717
平均值	0.2206	0.2084	0.2909	0.2844	0.2623	0.2040	0.1764	0.2371	0.1956	0.3124	0.2320	0.2108	0.1956	0.1878	0.1564

5.3 平滑参数变化对分类准确性的影响

选择 Column_3c、Connectionist_Bench、Glass、Sensor_reading 和 Spambase 这 5 个数据集，分别从单参和多参变化两方面进行平滑参数对分类准确性的影响程度计算与分析，如图 5 和图 6 所示，其中 $a1=0.001$，…，$a9=0.009$，$a10=0.01$，$a11=0.015$，…，$a28=0.1$.

(1) 单平滑参数变化的影响

单平滑参数($h=h_1=h_2=\cdots=h_n$)变化对 FBC 分类准确性的影响情况如图 5 所示.

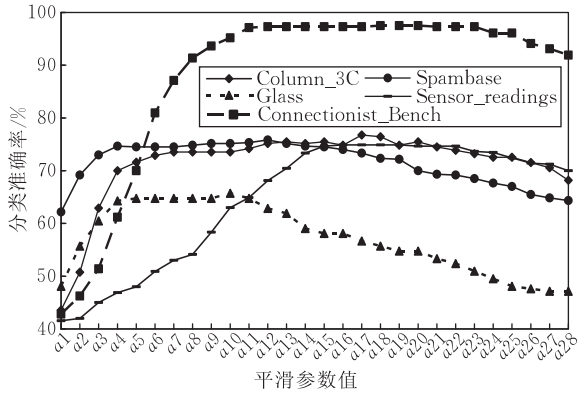


图 5 单平滑参数对分类准确性的影响

从图 5 能够看到，随着平滑参数的变化，分类器的分类准确率一般也在发生变化，曲线都有明显的峰值或高原区间. 在 0.001～0.1 范围内，对 5 个数据集，分类准确率的最大变化跨度依次是 32.87%、54.75%、18.58%、33.33% 和 13.67%，可见单平滑参数变化对分类器的分类准确性有较大的影响，因为所描述的是所有属性对分类的影响.

(2) 多平滑参数中单参数变化的影响

在 5 个数据集中，依次选择平滑参数 h_2、h_2、h_4、h_{24} 和 h_{20}. 对每一个数据集，除选择的平滑参数外，其它参数取单参数优化后的最优值，所选择的参数变化对分类器分类准确性的影响情况如图 6 所示.

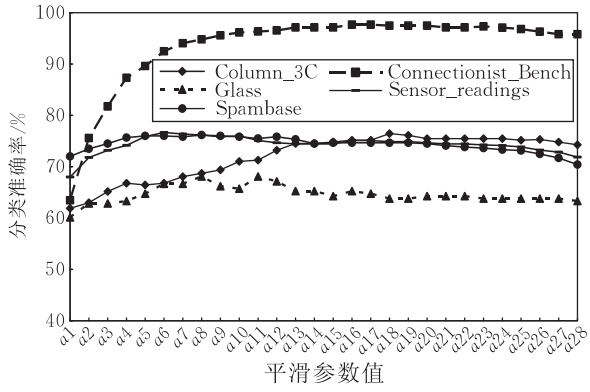


图 6 多平滑参数中的单参数对分类准确性的影响

从图 6 的总体来看，多参数中的单参变化相对更加平缓(个别情况变化较大)，但它们的累积影响往往

大于单参数，也就是通过局部调整会使所估计的属性条件联合密度更接近于真实密度，从而使经过优化的具有多平滑参数 FBC 具有更好的分类准确性.

5.4　DFBC 的分类准确性比较

选择与 GDP(Gross Domestic Product)、ERF(Exchange Rate Fluctuations) 和 EC(Energy Consumption)相关的 3 个宏观经济指标集，从国家统计局和相关数据源获取时序数据，时序数据记录数量依次是 20、31 和 25. 将 3 个数据集中的国内生产总值、实际有效汇率指数和能源消费总量，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)进行二值离散化作为类变量，进行动态分类准确性实验.

分别采用条件随机场(CRF)[14]，对基于高斯函数、单平滑参数高斯核函数和多平滑参数高斯核函数估计属性条件密度的动态朴素贝叶斯分类器(GDNB、SKDNB 和 MKDNB)，具有多平滑参数的动态完全贝叶斯分类器(MKDFB)进行比较，T_0 依次选取后 11 个时间点，情况如表 3 到表 5 所示.

表 3　GDP 波动转折点预测

分类器	$T_0=9$	$T_0=10$	$T_0=11$	$T_0=12$	$T_0=13$	$T_0=14$	$T_0=15$	$T_0=16$	$T_0=17$	$T_0=18$	$T_0=19$
CRF	36.36	40.00	44.44	50.00	57.14	66.66	60.00	50.00	33.33	50.00	0.00
GDNB	45.45	50.00	55.55	62.50	71.42	66.66	60.00	50.00	33.33	50.00	0.00
SKDNB	63.63	70.00	77.77	87.50	85.71	83.33	80.00	75.00	66.66	100.00	100.00
MKDNB	81.81	90.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
MKDFB	81.81	90.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

表 4　ERF 波动转折点预测

分类器	$T_0=20$	$T_0=21$	$T_0=22$	$T_0=23$	$T_0=24$	$T_0=25$	$T_0=26$	$T_0=27$	$T_0=28$	$T_0=29$	$T_0=30$
CRF	54.54	60.00	66.66	75.00	71.42	83.33	80.00	75.00	66.66	50.00	0.00
GDNB	36.36	40.00	33.33	37.50	42.85	50.00	40.00	50.00	33.33	0.00	0.00
SKDNB	54.54	60.00	55.55	62.50	71.42	66.66	60.00	75.00	66.66	50.00	0.00
MKDNB	63.63	60.00	66.66	75.00	85.71	100.00	100.00	100.00	100.00	100.00	100.00
MKDFB	63.63	70.00	77.77	87.50	100.00	100.00	100.00	100.00	100.00	100.00	100.00

表 5　EC 波动转折点预测

分类器	$T_0=14$	$T_0=15$	$T_0=16$	$T_0=17$	$T_0=18$	$T_0=19$	$T_0=20$	$T_0=21$	$T_0=22$	$T_0=23$	$T_0=24$
CRF	81.81	80.00	77.77	75.00	71.42	66.66	60.00	50.00	66.66	50.00	0.00
GDNB	54.54	60.00	55.55	50.00	57.14	66.66	60.00	50.00	66.66	50.00	0.00
SKDNB	72.72	70.00	77.77	75.00	71.42	66.66	60.00	50.00	66.66	50.00	0.00
MKDNB	81.81	80.00	88.88	87.50	100.00	100.00	100.00	100.00	100.00	100.00	100.00
MKDFB	81.81	80.00	88.88	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

从表 3 到表 5 综合来看，经过多参优化的动态朴素贝叶斯分类器，在分类准确率方面优于前面的分类器，而经过多参优化的动态完全贝叶斯分类器还要好于动态朴素贝叶斯分类器. 可见，动态完全贝叶斯分类器同样具有良好的分类准确性.

6　结论和进一步的工作

本文在使用具有多平滑参数的多元高斯核函数来估计属性条件联合密度的基础上，建立了能够有效利用属性之间条件依赖信息的完全贝叶斯分类器和动态完全贝叶斯分类器，并给出了将分类准确性标准与平滑参数区间异步长划分完全搜索相结合的分类器优化方法，使属性条件依赖信息利用和属性条件密度估计优化能够统筹兼顾. 使用 UCI 机器学习数据仓库中连续属性分类数据和宏观经济数据的实验结果显示，经过优化的完全贝叶斯分类器和动态完全贝叶斯分类器具有良好的分类准确性. 但以分类准确性为标准的分类器优化对大数据集会存在效率问题，而且多参数贪婪搜索也可能导致局部最优的问题，我们进一步的研究工作是如何提高学习效率和实现多平滑参数的全局优化.

参　考　文　献

[1] Chow C K, Liu C N. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory, 1968, 14(3): 462-467

[2] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning, 1997, 29(2-3): 131-161

[3] Grossman D, Domingos P. Learning Bayesian network classiers by maximizing conditional likelihood//Proceedings of the 21th International Conference on Machine Learning, Alberta, Canada, 2004: 361-368

[4] Jing Y S, Pavlović V, Rehg J M. Boosted Bayesian network classifiers. Machine Learning, 2008, 73(2): 155-184

[5] Webb G I, Boughton J R, Zheng F et al. Learning by extrapolation from marginal to full-multivariate probability distributions: Decreasingly naive Bayesian classification. Machine Learning, 2012, 86(2): 233-272

[6] John G H, Langley P. Estimating continuous distributions in Bayesian classifiers//Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI-1995). San Mateo, USA, 1995: 338-345

[7] Pérez A, Larrañaga P, Inza I. Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naïve Bayes. International Journal of Approximate Reasoning, 2006, 43(1): 1-25

[8] Pérez A, Larranga P, Inza I. Bayesian classifiers based on kernel density estimation: Flexible classifiers. International Journal of Approximate Reasoning, 2009, 50(2): 341-362

[9] Huang S C. Using Gaussian process based kernel classifiers for credit rating forecasting. Expert Systems with Applications, 2011, 38(7): 8607-8611

[10] Li Xu-Sheng, Guo Chun-Xiang, Guo Yao-Huang. The credit scoring model on extended tree augment naive Bayesian network. Systems Engineering Theory & Practice, 2008, 28(6): 129-136(in Chinese)
(李旭升, 郭春香, 郭耀煌. 扩展的树增强朴素贝叶斯网络信用评估模型. 系统工程理论与实践, 2008, 28(6): 129-136)

[11] Silverman B W. Using kernel density estimates to investigate multimodality. Journal of the Royal Statistical Society, 1981, 43(1): 97-99

[12] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection//Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). Montréal, Canada, 1995: 1137-1143

[13] Quinlan J R. Induction of decision trees. Machine Learning, 1986, 1(1): 81-106

[14] Lafferty J D, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data//Proceedings of the 18th International Conference on Machine Learning (ICML). Williams Town, MA, USA, 2001: 282-289

WANG Shuang-Cheng, born in 1958, Ph. D., professor. His main research interests include artificial intelligence, machine learning, data mining and their application.

DU Rui-Jie, born in 1980, Ph. D., lecturer. Her main research interests include machine learning and data mining.

LIU Ying, born in 1980, Ph. D., associate professor. Her main research interests include graph theory and machine learning.

Background

The study of Bayes classifier family with continuous attributes is of an important part of machine learning and data mining. At present, the research of Bayes derivative classifiers is respectively concentrated in the optimization of attribute conditional density and dependency extension to naive Bayes classifiers. Integrated optimization in two aspects is needed to improve the classification accuracy of classifiers. In this paper, the full Bayes classifiers and dynamic full Bayes classifiers with continuous attributes and multi smoothing parameters are presented on the basis of estimating the conditional joint density of attributes using multivariate Gaussian kernel function. They can effectively use conditional dependency information between attributes. The conditional joint density estimation of attributes can also be optimized by adjusting smoothing parameters. Experiment results show that optimized full Bayes classifiers and dynamic full Bayes classifiers have very good classification accuracy. Through this paper, the derivative classifier family of Bayes classifier can be deeply understood. But full Bayes classifiers and dynamic full Bayes classifiers have broad application prospects in many areas. The contents of this article is of an important part of National Natural Science Foundation (No. 11101284), Humanities and Social Science Foundation of the Chinese Education Commission (No. 10YJA630154, No. 12YJA630123), Leading Academic Discipline Project of Shanghai Municipal Education Commission (No. J51702), and Innovation Program of Shanghai Municipal Education Commission (No. 11YZ240). We have made deep studies to derivative classifier family of Bayes classifier with discrete attributes and naive Bayes classifiers optimized by attribute subset selection, Bayesian network and Markov network classifiers, Markov blanket classifiers, restricted Bayesian classification networks and so on have been respectively developed.