操作风险等级预测的朴素贝叶斯方法研究

王双成[1,2],冷翠平[1],侯彩虹[1]
WANG Shuang-cheng[1,2],LENG Cui-ping[1],HOU Cai-hong[1]

1.上海立信会计学院 信息科学系,上海 201620
2.上海立信会计学院 中国立信风险管理研究院,上海 201620
1.Department of Information Science,Shanghai Lixin University of Commerce,Shanghai 201620,China
2.Risk Management Research Institute,Shanghai Lixin University of Commerce,Shanghai 201620,China
E-mail:wangsc@lixin.edu.cn

WANG Shuang-cheng,LENG Cui-ping,HOU Cai-hong.Naive Bayes method in operational risk level prediction.Computer Engineering and Applications,2008,44(12):26-28.

Abstract: It is difficult to accumulate a large number of data with high quality in operational risk.Naive Bayes classifier is the one of best classifiers used to small data set classification.It is suitable for operational risk level prediction.In this paper,firstly, the process of learning and classing is presented on naive Bayes classifier with complete data sets.Then,a method naive Bayes classifier learning with missing data is developed based on star structure and Gibbs sampling.The existing problems can be avoided in local optimization,information losing and redundancy.

Key words: operational risk;level prediction;naive Bayes classifier;missing data;Gibbs sampling

摘 要:操作风险数据积累比较困难,而且往往不完整,朴素贝叶斯分类器是目前进行小样本分类最优秀的分类器之一,适合于操作风险等级预测。在对具有完整数据朴素贝叶斯分类器学习和分类的基础上,提出了基于星形结构和 Gibbs sampling 的具有丢失数据朴素贝叶斯分类器学习方法,能够避免目前常用的处理丢失数据方法所带来的局部最优、信息丢失和冗余等方面的问题。

关键词:操作风险;等级预测;朴素贝叶斯分类器;丢失数据;Gibbs 抽样
文章编号:1002-8331(2008)12-0026-03 **文献标识码:**A **中图分类号:**TP181

1 引言

随着世界多极化和全球经济一体化进程的加快,人类不仅受到不断变化的技术发展的挑战,而且面临着全新不断增长的系统性风险,世界正进入一个不同于传统“常态社会”的“风险社会”,风险已成为当今时代的主题之一。

操作风险[1]是目前企业所面临的主要风险之一,这种风险一度不被人们所重视,不断的惨痛教训使人们逐渐意识到操作风险管理的重要性。操作风险等级预测是操作风险管理的一项重要内容,其过程就是分类问题。由于操作风险数据积累比较困难,而且往往不完整,朴素贝叶斯(Naive Bayes,简称为 NB)分类器是现有进行小样本分类最优秀的分类器之一[2,3],适合于操作风险等级预测。目前,具有丢失数据 NB 分类学习主要基于众数均值法、记录删除法、设置新值法和 EM 算法等[4-7]进行丢失数据处理,易于导致局部最优、信息丢失和冗余等问题,从而降低分类器的分类准确性。本文针对操作风险等级预测实际需求,以及具有丢失数据 NB 分类器学习存在的问题,在具有完整数据 NB 分类器学习和分类的基础上,建立基于星形结构和 Gibbs sampling[8,9]的具有丢失数据 NB 分类器学习方法。这种方法使用星形结构分解联合概率解决了标准 Gibbs sampling 的指数复杂性问题,Gibbs sampling 迭代收敛到全局平稳分布,可摆脱使用 EM 算法的局部最优问题,同时也可避免使用众数均值法、记录删除法和设置新值法所带来的信息丢失和冗余等问题。

用 $X_1,\cdots,X_n$ 和 C 分别表示属性变量和类变量,属性既可以是连续变量,也可以是离散变量,$x_1,\cdots,x_n$ 和 c 为其值,例子集(数据库)D 中具有 N 个例子(记录),数据随机产生于概率分布 P。

2 具有完整数据的 NB 分类器学习

NB 分类器由结构和参数两部分构成,结构始终保持不变,不需要学习,因此具有完整数据的 NB 分类器学习的核心是分类器参数估计。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60675036);上海市重点学科(No.P1601);上海市教委重点项目(No.05zz66)。

作者简介:王双成(1958-),男,博士,教授,主要研究领域为人工智能、机器学习和数据采掘,以及在风险管理中的应用;冷翠平(1979-),女,博士,讲师,研究方向为智能控制与数据采掘;侯彩虹(1978-),女,博士,讲师,研究方向为智能评价与数据采掘。

收稿日期:2007-12-05 **修回日期:**2008-01-24

2.1 NB 分类器结构

NB 分类器假设属性变量在给定类变量时条件独立，因此分类器具有星形结构,如图 1 所示。

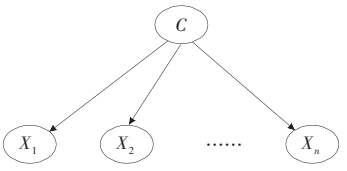

图 1　朴素贝叶斯分类器结构

2.2 NB 分类器参数估计

NB 分类器参数估计包括类参数估计(先验概率估计)和条件概率或密度估计(似然函数估计)。

(1)类概率估计

$$p(c_k)=\hat{p}(c_k)=\frac{N_{c_k}}{N}$$

其中 N_{c_k} 为第 c_k 类中例子数量。

(2)条件概率或密度估计

①条件概率估计

当 X_i 是离散的属性变量时,$p(x_i|c_k)=\hat{p}(x_i|c_k)=\frac{N_{c_k}^{(x_i)}}{N_{c_k}}$,其中 $N_{c_k}^{(x_i)}$ 为第 c_k 类中 $X_i=x_i$ 的例子数量。

如果某一个例子的数量是 0,那么 $\hat{p}(x_i|c_k)=\frac{\frac{1}{N}}{N_{c_k}+\frac{N_{x_i}}{N}}$,$N_{x_i}$ 是属性变量 X_i 取值数量。

②条件密度估计

当 X_i 是连续的属性变量时,取条件密度为正态密度(也可采用核密度或混合密度),$p(x_i|c)=g(x_i,\mu_{c_k},\sigma_{c_k})$,其中

$$g(x_i;\mu_{c_k},\sigma_{c_k})=\frac{1}{\sqrt{2\pi}\sigma_{c_k}}e^{-\frac{(x_i-\mu_{c_k})^2}{2(\sigma_{c_k})^2}}$$

$$\mu_{c_k}=\frac{x_{i1}(c_k)+\cdots+x_{iN}(c_k)}{N_{c_k}}$$

$$x_{ij}(c_k)=\begin{cases}x_{ij} & x_{ij}\in c_k\text{类}\\ 0 & x_{ij}\notin c_k\text{类}\end{cases}\qquad N_{c_k}=\begin{cases}1 & x_{ij}\in c_k\text{类}\\ 0 & x_{ij}\notin c_k\text{类}\end{cases}$$

即第 c_k 类的样本平均值。

$$\sigma_{c_k}^2=\frac{(x_{i1}(c_k)-\mu_{c_k}(c_k))^2+\cdots+(x_{iN}(c_k)-\mu_{c_k}(c_k))^2}{N_{c_k}}$$

$$\mu_{c_k}(c_k)=\begin{cases}\mu_{c_k} & x_{ij}\in c_k\text{类}\\ 0 & x_{ij}\notin c_k\text{类}\end{cases}$$

即第 c_k 类的样本方差。

2.3 NB 分类器表示形式与分类原理

由贝叶斯公式可得

$$p(c|x_1,\cdots,x_n)=\frac{p(c)p(x_1,\cdots,x_n|c)}{p(x_1,\cdots,x_n)}=\alpha p(c)p(x_1,\cdots,x_n|c)=\beta p(c)\prod_{i=1}^{n}p(x_i|c)\qquad(1)$$

其中 $p(x_i|c)$ 为条件概率或条件密度,α 和 β 是与 c 无关的量。

(1)NB 分类器分类原理

通过训练集 D 获得 $p(c),p(x_1|c),\cdots,p(x_n|c)$ 的估计值,对给定的属性值 $x_1^0,\cdots,x_n^0$,使 $p(c)\prod_{i=1}^{n}p(x_i|c)$ 最大(也使 $p(c|x_1,\cdots,x_n)$ 最大)的 c 的值便是 $x_1^0,\cdots,x_n^0$ 所属的类。

(2)NB 分类器表示形式

分类器表示形式为:$\arg\max_{c(x_1,\cdots,x_n)}\left\{p(c)\prod_{i=1}^{n}p(x_i|c)\right\}$。

3　具有丢失数据的 NB 分类器学习

具有丢失数据的 NB 分类器学习是一个迭代过程,首先对离散变量的丢失数据进行随机初始化,对连续变量的丢失数据使用均值进行初始化,按照数据库中记录的顺序,基于星形结构和 Gibbs sampling 依次对每一个记录中的丢失数据进行修正,修正所有记录中的丢失数据实现一次迭代,直到满足终止条件结束迭代,完成对数据集的修正。迭代产生两个序列,分别是数据集序列 $D^{(k)}$ 和参数向量序列 $\theta^{(k)}$(局部概率分布序列),用 $D^{(0)}$ 表示初始化后的数据集。

3.1 修正丢失数据

设已经进行了 k 次迭代，由星形结构 S 所决定的联合分布为:

$$p^{(k)}(x_1,\cdots,x_n,c|S)=p^{(k)}(c)p^{(k)}(x_1,\cdots,x_n|c,S)=p^{(k)}(c)\prod_{i=1}^{n}p^{(k)}(x_i|c)$$

按照变量的顺序和数据库中记录的顺序依次对具有丢失数据的变量进行抽样,并用抽样值修正待修正的数据。

设 X_i 和 C 在第 m 个记录的待修正值为 x_{im} 和 c_m，修正后的值为 $\hat{x}_{im}$ 和 $\hat{c}_m$,属性变量 X_i 和类变量 C 的可能取值为 $x_i^1,\cdots,x_i^{r_i}$ 和 $c^1,\cdots,c^{r_c}$,用 $D^{(k)}=D_{(1,1)}^{(k)}$ 表示第 k 次迭代前的数据集,$D_{(i,m)}^{(k)}$ 表示在第 k 次迭代中修正数据 x_{im} 之前的最新数据集,$D^{(k+1)}=D_{(1,N+1)}^{(k)}$ 表示第 k 次迭代后的数据集。当 $p^{(k)}(x_i^u|\pi_{x_{im}},c_m,D_{(i,m)}^{(k)},S)=0,u\in\{1,\cdots,r_i\}$ 时,对 $p^{(k)}(x_{im}^u|c_m,D_{(i,m)}^{(k)},S)$ 进行拉普拉斯修正(Laplace-corrected)[10],$p^{(k)}(x_{im}^u|c_m,D_{(i,m)}^{(k)})=(1/N)/(N(c_m)+N(x_{im}^u)(1/N))$,其中 $N(c_m)$ 和 $N(x_{im}^u)$ 分别为 $C=c_m$ 和 $X_i=x_{im}^u$ 的例子数量。

(1)类变量值的修正

对 $p^{(k)}(c_m|D_{(i,m)}^{(k)},S)\prod_{i=1}^{n}p(x_{im}|c_m,D_{(i,m)}^{(k)},S)$ 进行归一化处理,记

$$w_c(h)=\frac{p(c^h|D_{(i,m)}^{(k)},S)\prod_{i=1}^{n}p(x_{im}|c^h,D_{(i,m)}^{(k)},S)}{\sum_{j=1}^{r_c}p(c^j|D_{(i,m)}^{(k)},S)\prod_{i=1}^{n}p(x_{im}|c^j,D_{(i,m)}^{(k)},S)},h\in\{1,\cdots,r_c\}$$

对生成的随机数 λ,得到 $\hat{c}_m=\begin{cases}c^1 & 0<\lambda\leqslant w_c(1)\\ \cdots & \\ c^h & \sum_{j=1}^{h-1}w_c(j)<\lambda\leqslant\sum_{j=1}^{h}w_c(j)\\ \cdots & \\ c^{r_c} & \lambda>\sum_{j=1}^{r_c-1}w_c(j)\end{cases}$。

(2)属性变量值的修正

属性变量值的修正包括离散变量和连续变量值的修正。

①离散变量 X_j 值的修正

对 $p(x_{im}|c_m,D^{(k)}_{(i,m)},S)$ 进行归一化处理，记

$$w_i(h)=\frac{p^{(k)}(x_i^h|c_m,D^{(k)}_{(i,m)},S)}{\sum_{j=1}^{r_i-1}p^{(k)}(x_i^j|c_m,D^{(k)}_{(i,m)},S)},h\in\{1,\cdots,r_i\}$$

对生成的随机数 λ，得到 $\hat{x}_{im}=\begin{cases}x_i^1 & 0<\lambda\leqslant w_i(1)\\ \cdots & \\ x_i^h & \sum_{j=1}^{h-1}w_i(j)<\lambda\leqslant\sum_{j=1}^{h}w_i(j)\\ \cdots & \\ x_i^{r_i} & \lambda>\sum_{j=1}^{r_i-1}w_i(j)\end{cases}$。

②连续变量 X_j 值的修正

首先生成两个随机数 ξ_1 和 ξ_2，则 $\hat{x}'_{jm}=(-2\ln\xi_1)^{1/2}\sin(2\pi\xi_2)$，服从 $N(0,1)$ 分布[8]，X_j 的取值为：$\hat{x}_{jm}=\sigma_{x_{(n+1)m}}\hat{x}'_{jm}+\mu_{x_{(n+1)m}}$，可知 $\hat{x}_{jm}$ 服从 $N(\mu_{x_{(n+1)m}},\sigma_{x_{(n+1)m}})$ 分布。

(2)局部概率(参数)调整

如果 $c_m\neq\hat{c}_m$，则需要调整对应的类变量参数；如果 $x_{im}\neq\hat{x}_{im}$，则需要调整对应的属性变量参数。

①类参数调整

$$p^{(k)}(c_m|D^{(k)}_{(i+1,m)},S)=p^{(k)}(c_m|D^{(k)}_{(i,m)},S)-1/N$$

$$p^{(k)}(\hat{c}_m|D^{(k)}_{(i+1,m)},S)=p^{(k)}(\hat{c}_m|D^{(k)}_{(i,m)},S)+1/N$$

②离散属性参数调整

$$p^{(k)}(x_{im}|c_m,D^{(k)}_{(i+1,m)},S)=\frac{p^{(k)}(x_{im},c_m|D^{(k)}_{(i,m)},S)-1/N}{p^{(k)}(c_m|D^{(k)}_{(i,m)},S)-1/N}$$

$$p^{(k)}(\hat{x}_{im}|c_m,D^{(k)}_{(i+1,m)},S)=\frac{p^{(k)}(\hat{x}_{im},c_m|D^{(k)}_{(i,m)},S)+1/N}{p^{(k)}(c_m|D^{(k)}_{(i,m)},S)+1/N}$$

③连续属性参数调整

经过简单的数学推导可得如下调整公式：

$$\hat{\mu}^{(k)}_{jm}=\mu^{(k)}_{jm}-\frac{x_{jm}(c_k)-\hat{x}_{jm}(c_k)}{N_{c_i}}$$

$$(\hat{\sigma}^{(k)}_{jm})^2=(\sigma^{(k)}_{jm})^2-\frac{x^2_{jm}(c_k)-\hat{x}^2_{jm}(c_k)}{N_{c_i}}+N_{c_i}[(\hat{\mu}^{(k)}_{jm})^2-(\mu^{(k)}_{jm})^2]$$

3.2 数据集迭代终止检验

把丢失数据按照某一顺序排成一个序列，称为丢失数据序列。采用相邻两次迭代丢失数据序列的一致性检验进行终止迭代判断。设相邻两次迭代所得到的丢失数据序列分别为 $x^{(k)}_{i1},x^{(k)}_{i2},\cdots,x^{(k)}_{iM}$ 和 $x^{(k+1)}_{i1},x^{(k+1)}_{i2},\cdots,x^{(k+1)}_{iM}$，$sig(x^{(k)}_{ij},x^{(k+1)}_{ij})=\begin{cases}0,x^{(k)}_{ij}=x^{(k+1)}_{ij}\\ 1,x^{(k)}_{ij}\neq x^{(k+1)}_{ij}\end{cases}$，$1\leqslant j\leqslant M$。对给定的阈值 $\eta_0>0$，如果 $\frac{1}{M}\sum_{j=1}^{M}sig(x^{(k)}_{ij},x^{(k+1)}_{ij})<\eta_0$，则结束迭代。

4 操作风险等级预测

操作风险等级预测的目的是根据影响操作风险各因素现状，通过推理得到目前操作风险等级，以便决策者掌握操作风险的威胁程度，制定相应的策略化解或转移操作风险的影响。

图 2 是一个用于操作风险管理的贝叶斯网络，其中变量的实际含义是：风险文化-RC，管理技能-MS，操作过程-OP，损失报告系统质量-RM，技术-TE，欺骗威胁-DT，恐怖威胁-TT，系统威胁-ST，薄弱环节-WT，实际损失事件-LN，损失事件的严重性-LI，损失控制过程能力-LC，操作风险等级 OR，员工技能-ES，报告的损失事件-RS。

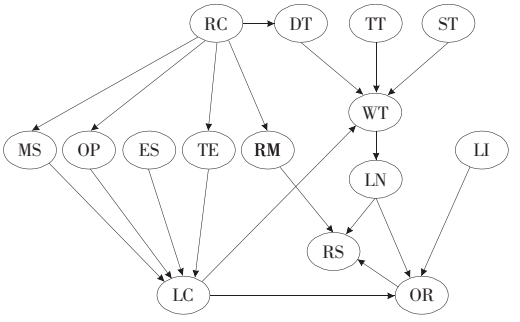

图 2 操作风险管理贝叶斯网络

根据各种因素的当前配置，使用朴素贝叶斯方法可预测出操作风险的等级 OR 的值，从而能够保障及时了解操作风险的现状，为决策者制定相应的策略提供依据。

5 实验

在 UCI 机器学习数据仓库[11]中选择 6 个分类数据集 hepatitis、iris、voting_records、new thyroid、wdbc、wine 进行实验，采用 10 折交叉有效性(10-fold cross-validation)验证[12]方法进行分类器的分类准确性估计。

(1)迭代收敛性实验

取 η_0=0.05，丢失数据 30%，使用 hepatitis、iris、voting_records 进行数据集修复迭代收敛性实验，情况如图 3 所示。

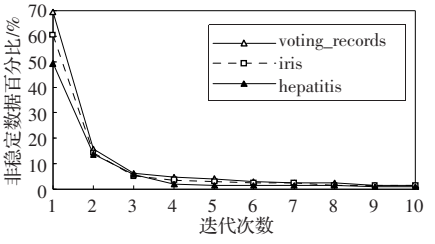

图 3 修正数据集迭代收敛情况

从图 3 可以看出，迭代 5 次后均收敛，其它的数据集也具有类似的情况，这表明了算法具有较高的效率。

使用 6 个分类数据集进行具有丢失数据的分类准确性实验，去掉标准差，并取 6 个数据集分类准确性的平均值，与普遍采用的处理丢失数据众数均值法、记录删除法、设置新值法和 EM 算法进行比较，情况如图 4 所示。

图 4 中显示，对具有不同丢失数据比例的数据集，经过修复后的数据集分类正确率明显具有优势，而且随着丢失数据比例的增大，数据迭代修复的作用逐渐增大，这说明了迭代学习方法的有效性。其主要原因是：基于星形结构和 Gibbs sampling 的数据修复迭代收敛到对应的全局平稳分布，能够避免使用

(下转 94 页)