# 基于贝叶斯分类的增强学习协商策略

孙天昊[1,2] 陈 飞[1] 朱庆生[1] 曹 峰[2]
(重庆大学计算机学院 重庆 400030)[1] (中国嘉陵工业股份有限公司(集团)信息技术部 重庆 400032)[2]

**摘 要** 为了帮助协商Agent选择最优行动实现其最终目标,提出基于贝叶斯分类的增强学习协商策略。在协商过程中,协商Agent根据对手历史信息,利用贝叶斯分类确定对手类型,并及时动态地调整协商Agent对对手的信念。协商Agent通过不断修正对对手的信念,来加快协商解的收敛并获得更优的协商解。最后通过实验验证了策略的有效性和可用性。

**关键词** 贝叶斯分类,增强学习,协商策略,协商历史
**中图法分类号** TP301 **文献标识码** A

## Reinforcement Learning Negotiation Strategy Based on Bayesian Classification

SUN Tian-hao[1,2] CHEN Fei[1] ZHU Qing-sheng[1] CAO Feng[2]
(College of Computer Science, Chongqing University, Chongqing 400030, China)[1]
(Department of Information Technology, China Jialing Industrial Co., Ltd(Group), Chongqing 400032, China)[2]

**Abstract** To help negotiation Agent to select its best actions and reach its final goal, a reinforcement learning negotiation strategy based on Bayesian classification was proposed. In the middle of negotiation process, negotiation Agent makes the best use of the opponent's negotiation history to make a decision of the opponent's type based on Bayesian classification, dynamically adjust the negotiation Agent's belief of opponent in time, quicken the negotiation result convergence and reach the better negotiation result. Finally, the algorithm was proved to be effective and practical by experiment.

**Keywords** Bayesian classification, Reinforcement learning, Negotiation strategy, Negotiation history

## 1 引言

随着网络技术的不断发展,人们越来越依赖于网络进行商务活动。传统的商务活动需要交易双方对商品的各种议题进行协商。协商是双方或多方就某些共同感兴趣的议题进行交流,以获得一致的过程[1]。Agent技术是一种能够有效代替人进行协商的工具或手段,可以实现电子协商的自动化和智能化[2]。

将学习机制引入基于Agent的电子商务协商阶段,也就是在协商过程中学习对手的信念、偏好以及协商环境知识,而在协商结束后学习和评价协商结果,使得Agent在动态变化的环境中能够根据对方信念进行推理,更新自身信念,自主地提高协商能力,更有效地与对手进行协商,使利益最大化[3]。

机器学习的主流算法在协商中都已有相应的研究,包括增强学习[4-6]、贝叶斯学习[7-10]、遗传算法[10-12]、支持向量机[13,14]和神经网络[15]等。

增强学习(Reinforcement learning)[4]通过感知环境状态和从环境中获得不确定回报来学习动态系统的最优行为策略,它以获得极大化期望回报为学习目的。文献[5,6]在协商过程中引入增强学习,加快协商进度和提高协商效率。

文献[7]依据对手提议序列和让步假设,使用贝叶斯学习来决定对手提议所属类别。文献[8,9]通过贝叶斯学习来更新每个Agent关于环境和其它Agent的知识和信念。文献[10]利用贝叶斯学习预测对手的保留价和时间期限,并使用遗传算法产生每一轮协商的提议。

文献[11]将混合遗传算法(HGA)应用于协商模型,以提高Agent协商的效率。文献[12]利用遗传算法学习协商中对手们的偏好和约束。

文献[13,14]通过支持向量机的方法来学习协商轨迹,得到协商对手在每个协商项的态度。然后利用学习得到的对手协商态度,构造了一个协商的决策模型。

文献[15]使用人工神经网络在时间序列上建模协商过程。训练神经网络时使用Levenberg-Marquardt算法和贝叶斯规则。

本文综合应用贝叶斯分类和增强学习方法提高协商效率。贝叶斯学习主要是通过学习更新信念,以便更加准确地获得对手的私有信息。$Q$学习算法是最重要的增强学习算法。本文依据对手协商历史,使用贝叶斯学习确定对手的协商类型,然

到稿日期:2010-10-14 返修日期:2010-12-13 本文受中央高校基本科研业务费科研专项项目(CDJRC10180012,CDJZR10180014)资助。
孙天昊(1979-),男,博士后,讲师,主要研究方向为电子商务、智能信息处理,E-mail:sthing@cqu.edu.cn;陈 飞(1986-),男,硕士生,主要研究方向为电子商务;朱庆生(1956-),男,博士,教授,主要研究方向为虚拟植物生长可视化、面向服务的软件技术、电子商务与现代物流;曹 峰(1973-),男,硕士,高工,主要研究方向为电子商务、企业信息化。

后更正自己的私有信念，使用 $Q$ 学习生成反提议，目的是在贝叶斯分类之后使用 $Q$ 学习算法产生新的更符合现实情况的报价。

## 2 基于贝叶斯分类的增强学习协商策略

### 2.1 基于增强学习协商策略

协商双方交替报价，只有在最终协商成功后，才能获得相应的回报 $r$。协商双方对于协商对象各自有不同的价格评估区间，对方报价只有在该区间内才是可以接受的价格。记买方和卖方的报价区间分别为$[p_b^{\min}, p_b^{\max}]$和$[p_s^{\min}, p_s^{\max}]$。若最终协商成功，记成交价为 $p^T$，则双方所得回报分别为：

$$\text{卖方：} r_s = p^T - p_s^{\min} \tag{1}$$

$$\text{买方：} r_b = p_b^{\max} - p^T \tag{2}$$

买卖双方各自的协商时限分别为 $T_b$、$T_s$，即买卖双方各自允许的最大报价次数。

**定义 1** 时间信念是指协商 Agent 认为对方接受其报价的概率。

买卖双方的时间信念分别记为${}^b b^s(t)$、${}^s b^b(t)$。一般分为增函数（如${}^b b^s(t)=t/T_b$）、减函数（如${}^b b^s(t)=1-t/T_b$）或者常函数（如${}^b b^s(t)=0.5$）。

**定义 2** 价格信念是指协商 Agent 对成交价格在其报价区间内概率分布的认识。

买卖双方的价格信念分别记为${}^b p^s(t)$、${}^s p^b(t)$。如${}^b p^s(t)=1/(p_b{}^{\max}-p_b^{\min})$，${}^s p^b(t)=1/(p_s^{\max}-p_s^{\min})$。

基于 Q-learning 学习的 $Q$ 函数定义为：

$$Q(s(t),p(t))=r(s(t),p(t))+\gamma \max_{p(t+1)} Q(\delta(s(t),p(t)),p(t+1)) \tag{3}$$

式中，$\gamma$ 为时间贴现率，$\delta()$为状态转移函数。

若协商在第 $t$ 次报价时成功，则

卖方 Agent 回报：

$$Q_t^s = \int_{p_s^{\min}}^{p_s^{\max}} (p^T - p_s^{\min})\, {}^s p^b \, \mathrm{d}_{p^T} \tag{4}$$

买方 Agent 回报：

$$Q_t^b = \int_{p_b^{\min}}^{p_b^{\max}} (p_b^{\max} - p^T)\, {}^b p^s \, \mathrm{d}_{p^T} \tag{5}$$

卖方 Agent 第 $t$ 阶段 $Q$ 值的平均期望为：

$$\bar{Q}_s(s(t),p(t)) = \left(\sum_{i=t}^{T_s} {}^s b^b(i)\gamma^{i-t} Q_e^s\right)/(T_s - t + 1) \tag{6}$$

买方 Agent 第 $t$ 阶段 $Q$ 值的平均期望为：

$$\bar{Q}_b(s(t),p(t)) = \left(\sum_{i=t}^{T_b} {}^b b^s(i)\gamma^{i-t} Q_e^b\right)/(T_b - t + 1) \tag{7}$$

最后，得到卖方 Agent 的报价策略为：

$$p_s(t) = p_s^{\min} + \bar{Q}_s(s(t),p(t)) \tag{8}$$

买方 Agent 的报价策略为：

$$p_b(t) = p_b^{\max} - \bar{Q}_b(s(t),p(t)) \tag{9}$$

### 2.2 对手类型

在协商过程中，协商者可以充分利用对手的历史信息并加以学习，以促进协商过程[7,13,14]。为了更好、更快地实现买卖双方报价的收敛，在 $Q$ 学习中加入对对手协商历史的学习。通过对手协商历史信息，判断对手类型，动态调整对对手的信念。

**定义 3** 买方 Agent 的协商历史 $H_b(t)$是在协商过程中买方 Agent 给出的报价的序列：

$$H_b(t) = p_b(1), p_b(2), \cdots, p_b(t)$$

式中，$p_b(1) = p_b^{\min}$。对卖方 Agent，$p_s(1) = p_s^{\max}$。

根据协商对手在协商历史信息中所表现出来的特点，一般把参加协商的 Agent 分为 3 种类型[1]：易妥协型($C$)、固执型($O$)、均匀线型($L$)。

根据对手的协商行为可以把对手让步幅度分为绝对平均让步幅度、绝对最小让步幅度、绝对最大让步幅度等几种[1]，本文采用最常用的绝对平均让步幅度，以买方 Agent 为例，记 $\Delta = (p_b(t) - p_b(1))/(t-1)$，其中 $t>1$。记 $\alpha = \Delta / p_b^{\min}$ 表示平均让步幅度与协商初始值的比值，则对手类型：

$$C = \begin{cases} C_C, & \alpha > \theta_1 \\ C_L, & \theta_2 \leqslant \alpha \leqslant \theta_1 \\ C_O, & \alpha < \theta_2 \end{cases}$$

式中，$\theta_1$ 和 $\theta_2$ 分别为分类的上限和下限，例如 $\theta_1 = 0.05$，$\theta_2 = 0.1$。

### 2.3 贝叶斯分类

贝叶斯公式为：

$$P(C_i|x) = \frac{P(C_i)P(x|C_i)}{\sum_{j=1}^{k} P(x|C_j)P(C_j)} = \frac{P(C_i)P(x|C_i)}{P(x)} \tag{10}$$

式中，$P(C_i)$为先验概率，$P(x|C_i)$为联合概率，$P(C_i|x)$为后验概率。

贝叶斯分类就是候选分类集合 $C$ 中寻找给定数据 $x$ 时可能性最大的假设 $C_{MAP}$，$C_{MAP}$ 被称为最大后验概率。

$$C_{MAP} = \arg\max_{C_i \in C} P(C_i|x) = \arg\max_{C_i \in C} \frac{P(C_i)P(x|C_i)}{P(x)} = \arg\max_{C_i \in C} P(C_i)P(x|C_i) \tag{11}$$

式中，$i \in [1, |C|]$。于是 $x$ 就属于 $C_{MAP}$ 类。

贝叶斯分类已被广泛应用。文献[7]使用贝叶斯分类确定对手偏好的类型。

### 2.4 信念调整

在经过贝叶斯分类得到对手类型后，调整对对手的时间信念函数。调整的原则为：

若对手是保守类型，则信念函数调整为减函数；

若对手是激进类型，则信念函数调整为增函数；

若对手是一般类型，则信念函数调整为常函数。

以买家为例，

$${}^b b^s(t) = \begin{cases} t/T_b, & C^s = C_C \\ 1 - t/T_b, & C^s = C_O \\ 0.5, & C^s = C_L \end{cases} \tag{12}$$

### 2.5 协商报价策略

在协商过程的每次报价中，先计算让步比例 $\alpha$；然后根据 $\alpha$ 判断本次报价时对手的类型，再用贝叶斯分类调整对手的类型信念。在得到对手的类型后调整对对手的信念函数，$Q$ 学习将根据新的时间信念函数来计算下次报价值。

报价策略算法：

(1)计算让步比例 $\alpha$。

(2)根据 $\alpha$ 判断本次报价时对手的类型。

(3)贝叶斯分类调整对手的类型信念。

(4)调整对对手的信念函数。

(5)$Q$ 学习计算下次报价值。

## 3 实验

表 1 共有 3 组实验数据，分别设置买卖双方各自允许的

最大报价次数、报价区间、时间信念和价格信念。贝叶斯学习所用先验概率 $P(C_i)$ 随机生成，其中，${}^bp^s(t)=1/(p_b^{max}-p_b^{min})$，${}^sp^b(t)=1/(p_s^{max}-p_s^{min})$，$C_i\in(C_C,C_O,C_L)$，$\sum C_i=1$。$\theta_1=0.05$，$\theta_2=0.1$，$\gamma=0.9$。

表 1　实验数据

| | $T_s$ | $T_b$ | $p_s^{min}$ | $p_s^{max}$ | $p_b^{min}$ | $p_b^{max}$ | ${}^bb^s(t)$ | ${}^sb^b(t)$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 22 | 20 | 20 | 40 | 15 | 21 | $1-\frac{t}{T_b}$ | $1-\frac{t}{T_s}$ |
| 2 | 18 | 18 | 20 | 35 | 15 | 21 | $\frac{t}{T_s}$ | $1-\frac{t}{T_s}$ |
| 3 | 22 | 22 | 15 | 50 | 12 | 20 | $1-\frac{t}{T_b}$ | $1-\frac{t}{T_s}$ |

在 VC++环境中实现了该策略，并分别对 3 组实验数据进行 2 种情形实验比较：情形 1)增强学习；情形 2)基于贝叶斯分类的增强学习。其中，情形 1)指的是只使用增强学习策略进行协商，也即在协商过程中，一直保持原有的最初的对对手的信念，不会对协商对手的信念进行调整。情形 2)在情形 1)的基础上增加基于对手历史信息的贝叶斯分类。具体指的是在协商过程中，先根据 $\alpha$ 判断对手的类型后，再进一步使用贝叶斯分类，然后才调整对手的信念函数。

图 1 是第 1 组实验数据的实验结果，情形 1)对应 s1，b1 表示基于增强学习策略的买卖双方的报价过程。双方时间信念都为减函数，卖方以递减报价，买方以递增报价，最终在 $t=18$ 时以价格 20.82 成交。

情形 2)对应 s2，b2b。其中 s2，b2b 分别表示卖方采用基于增强学习策略、买方采用基于贝叶斯分类的增强学习策略的买卖双方的报价过程。在 $t=19$ 时成交，成交价是 20.64，该成交价与情形 1)相比，对买方是有利的，即买方以更低的价格成交。由此可见，贝叶斯分类能帮助协商者获得对自己更加有利的成交价。

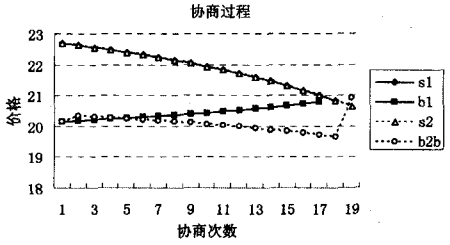


图 1　第 1 组实验数据的实验结果

图 2 是第 2 组实验数据的实验结果。情形 1)对应 s1，b1。双方在协商时限内未能达成交易。情形 2)对应 s2，b2b。双方在 $t=14$ 时以价格 20.7 成功完成交易。表明贝叶斯分类能够动态调整对手信念，最终促成交易。

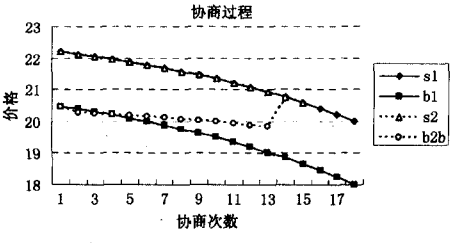


图 2　第 2 组实验数据的实验结果

图 3 是第 3 组实验数据的实验结果。情形 1)对应 s1，b1，双方在 $t=6$ 时以价格 19.05 成功完成交易。情形 2)对应 s2b，b2b，双方在 $t=3$ 时以价格 18.95 成功完成交易。表明双方在使用贝叶斯分类后更快达成了交易。

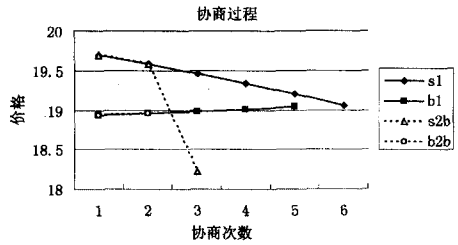


图 3　第 3 组实验数据的实验结果

以上 3 组实验表明，与基于增强学习策略相比，基于贝叶斯分类的增强学习策略能帮助协商者(1)获得对自己更加有利的成交价；(2)动态调整对手信念，使得原本失败的协商最终也能促成交易；(3)更快达成了交易。

**结束语**　本文提出了基于贝叶斯分类的增强学习协商策略。在协商过程中，充分利用已获得的对手协商历史信息，对对手协商历史进行学习，根据协商历史对对手使用贝叶斯分类进行分类，根据分类结果动态调整协商者的信念，协商者使用新的信念利用增强学习的 $Q$ 学习算法生成反提议继续协商。本方法的特点是充分利用机器学习机制，赋予协商者学习能力；充分利用协商历史，注重协商交互过程；动态调整协商者信念，满足协商过程的动态性。实验结果表明，基于贝叶斯分类的增强学习策略能够更快达到协商解并能提高协商解的质量。

## 参 考 文 献


[1] Faratin P. Automated service negotiation between autonomous computational agents [D]. University of London, 2000

[2] He M H, Jennings N R, Leung H F. On agent-mediated electronic commerce [J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 985-1003

[3] Stone P, Veloso M. Multi-agent systems: A survey from a machine learning perspective [J]. Autonomous Robots, 2000, 3(8): 345-383

[4] 孙天昊，朱庆生，李双庆，等. 一种优化的基于增强学习的协商策略[J]. 计算机工程与应用，2008，40(30)：24-25

[5] Li J. An agent bilateral multi-issue alternate bidding negotiation protocol based on reinforcement learning and its application in E-commerce[C] // Proceedings of the International Symposium on Electronic Commerce and Security (ISECS 2008). Guangzhou, China. August 2008: 217-220

[6] Cao J G. Research on electronic commerce automated negotiation in multi-agent system based on reinforcement learning [C] // Proceedings of the 2009 International Conference on Machine Learning and Cybernetics. Baoding, China, 2009(3): 1419-1423

[7] Scott B, Bruce S. A Bayesian classifier for learning opponents' preferences in multi-object automated negotiation [J]. Electronic Commerce Research and Applications, 2007, 6: 274-284

[8] Zeng D, Sycara K. Bayesian learning in negotiation [J]. Int'l J. Human-Computer Studies, 1998, 48: 125-141

[9] Shi D J, Liu Z H Q, He J. MAS learning based on Bayesian learning method [J]. Applied Mechanics and Materials, Information Technology for Manufacturing Systems, 2010 (20-23): 1292-1298


(下转第 247 页)

离。

根据模式序列的动态时间弯曲距离可以求任意两个长度的模式序列间的距离。模式序列间的动态时间弯曲距离满足正定性、对称性和三角不等式性，因而，可用它定义两个模式序列的相似性。

**定义 4** 压缩时间序列 $X=(x_1,x_2,\cdots,x_n)$ 和 $Y=(y_1,y_2,\cdots,y_m)$ 之后得到的模式序列分别为：

$S_X=\{M_1,M_2,\cdots,M_t\}$

$S_Y=\{N_1,N_2,\cdots,N_p\}$

称 $D_{drw}(S_X,S_Y)$ 为模式序列 $S_X$ 和 $S_Y$ 的相似性函数，即有：

$\mathrm{Sim}(S_X,S_Y)=D_{drw}(S_X,S_Y)$

在判断两个模式序列是否相似时，只需要计算出它们之间的动态时间弯曲距离，再给定正常数 $\varepsilon$，如果相似性函数值小于 $\varepsilon$ 时，那么它们就相似，否则它们不相似。

**定义 5** 把时间序列 $X=(x_1,x_2,\cdots,x_n)$ 转换成模式序列，结果分别为：

$S_X=\{M_1,M_2,\cdots,M_t\}$

任意序列模式 $S_{X1}$，$S_{X2}\in S_X$，$S_{X1}=\{M_j,M_{j+1},\cdots,M_{j+l_1}\}$，$S_{X2}=\{M_i,M_{i+1},\cdots,M_{i+l_2}\}$，称 $D_{drw}(S_{X1},S_{X2})$ 为序列模式 $S_{X1}$ 和 $S_{X2}$ 的相似性函数，即有：

$\mathrm{Sim}(S_X,S_Y)=D_{drw}(S_X,S_Y)$

式中，$1\leqslant i,i+l_1\leqslant t,1\leqslant j,j+l_1\leqslant t$。

## 4 仿真实验

用 2006 年沪深 300 指数的全部数据，以及 2007 年沪深 300 指数的全部数据作为实验对象，如图 2、图 3 所示。分别把它们转换成模式序列，分别含有 80 和 88 个元模式，应用模式序列动态时间弯曲距离方法计算它们的距离为 28.4。因而这两个时间序列的模式序列并不具有较好的相似性，这是因为 2006 年沪深 300 指数温和增长，而到了 2007 年几乎全民进入股市，促使 2007 年沪深 300 指数急剧增长，使它们具有很大的差异性。

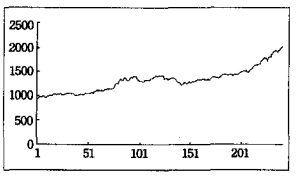


图 2 2006 年沪深 300 指数

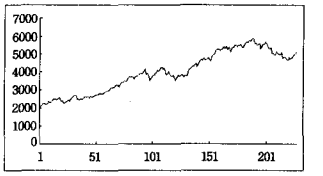


图 3 2007 年沪深 300 指数

## 参 考 文 献


[1] 翁颖钧，朱仲英. 基于分段线性动态时间弯曲的时间序列聚类算法研究. 研究与设计[J]. 微型电脑应用，2003，19(9)

[2] Berndt D J，Clifford J. Finding patterns in time series：DB，2000. A dynamic programming approach[Z]. Advances in Knowledge Discovery and Data Mining，1996

[3] Berndt J，Clifford D. Using dynamic time warping to find patterns in time series[C]// AAAI-94 Workshop on Knowledge Discovery in Database. 1994：229-248

[4] 曲文龙，张德政，杨炳儒. 基于小波和动态时间弯曲的时间序列相似匹配[J]. 北京科技大学学报，2006，28(4)

[5] Berndt D，Clifford J. Using dynamic time warping tp find patterns in time series[C]// AAAI Workshop on Knowledge Discovery in Databases. 1994：229-248

[6] Aach J，Church G. Alidning gene expression time series with time warping algorithms[J]. Bioinformatics，2001，17

[7] Kim S W，Park S，Chu W W. Efficient processing of similarity search under time warping in sequence databases：an index-based approach[J]. Inf. Syst.，2004，29(5)：405-420

[8] Das G，Lin K I，Marmila H，et al. Rule Discovery from Time Series [A]// Proc. of the 4th Int. Conf. on Knowledge Discovery and Data Mining[C]. [S. l.]：AAA I Press，1998：16-22

[9] 张保稳. 时间序列数据挖掘研究[D]. 西安：西北工业大学，2002

[10] 王亮，姜丽红. 快速挖掘最大频繁模式算法[J]. 计算机工程与应用，2006(17)

[11] 周勇. 时间序列时序关联规则挖掘研究[D]. 成都：西南财经大学，2008(6)：40-60


(上接第 229 页)


[10] Sim K M，Guo Y Y，Shi B Y. BLGAN：Bayesian learning and genetic algorithm for supporting negotiation with incomplete information [J]. IEEE Transactions on Systems，Man，and Cybernetics，Part B：Cybernetics，2009，39(1)：198-211

[11] 李剑，牛少彰. 一种基于混合遗传算法的双边多议题协商[J]. 北京邮电大学学报，2009，32(2)：1-4

[12] Ng S C，Sulaiman M N，Selamat M H. Machine learning approach in optimizing negotiation agents for E-Commerce [J]. Information Technology Journal，2009，8(6)：801-810

[13] 程昱，高济，古华茂，等. 基于对手态度学习的协商决策模型[J]. 浙江大学学报：工学版，2008，42(10)：1676-1680

[14] 程昱，高济，古华茂，等. 基于机器学习的自动协商决策模型[J]. 软件学报，2009，20(8)：2160-2169

[15] Real C，Kersten G E，Vahidov R. Predicting opponent's moves in electronic negotiations using neural networks[J]. Expert Systems with Applications，2008，34(2)：1266-1273


(上接第 244 页)


[2] 卢小甫. 切丛流行学习算法及其应用研究[D]. 苏州：苏州大学，2010

[3] Zhou Li-li，LI Fan-zhang. Research on Mapping Mechanism of Learning Expression[C]// Jian Yu，et al. Proceeding of Rough Set and Knowledge Technology. Beijing，China，October 2010：298-303

[4] 贺伟. 范畴论[M]. 北京：科学出版社，2006

[5] 李凡长，何书萍，钱旭培. 李群机器学习研究综述[J]. 计算机学报，2010，33(7)：115-1126

[6] 卢小甫. 切丛流行学习算法及其应用研究[D]. 苏州：苏州大学，2010

[7] Jolliffe I T. Principal Component Analysis [M]. New York：Springer，1989

[8] Tenenbaum J B，de Silva V，et al. A global geometric framework for nonlinear dimensionaility reduction [J]. Science，2000，290(5500)：2319-2323

[9] 赵连伟，罗四维，等. 高位数据流形的低维嵌入及嵌入维数研究[J]. 软件学报，2005，16(8)：1423-1430

[10] Roweis ST，Saul LK. Nonlinear dimensionality analysis by locally linear embedding[J]. Science，2000，290(12)：2323-2326

[11] Tenenbaum J B，de Silva V，et al. A global geometric framework for nonlinear dimensionaility reduction [J]. Science，2000，290(5500)：2319-2323