文章编号: 1001-0920(2008)07-0828-05

采用二重扰动机制的支持向量机的集成训练算法

贾华丁[1,2], 游志胜[1], 王 磊[2]
(1. 四川大学 计算机科学与工程学院, 成都 610064; 2. 西南财经大学 经济信息工程学院, 成都 610074)

摘 要: 为了有效提升支持向量机的泛化性能,提出两种集成算法对其进行训练. 首先分析了扰动输入特征空间和扰动模型参数两种方式对于增大成员分类器之间差异性的作用;然后提出两种基于二重扰动机制的集成训练算法. 其共同特点是,同时扰动输入特征空间和模型参数以产生成员分类器,并利用多数投票法对它们进行组合. 实验结果表明,因为同时缩减了误差的偏差部分和方差部分,所以两种算法均能显著提升支持向量机的泛化性能.

关键词: 支持向量机; 集成算法; 二重扰动机制; 成员分类器
中图分类号: TP18 **文献标识码**: A

Ensemble algorithms for training support vector machine based on the double disturbance mechanism

JIA Hua-ding[1,2], *YOU Zhi-sheng*[1], *WANG Lei*[2]
(1. School of Computer Science and Engineering, Sichuan University, Chengdu 610064, China; 2. School of Economics Information Engineering, Southwest University of Finance and Economics, Chengdu 610074, China. Correspondent: JIA Hua-ding, E-mail: jihad@swufe.edu.cn)

Abstract: For improving the generalization performance of support vector machine (SVM) effectively, two ensemble algorithms are proposed to train SVM. Firstly, the effectivity of two different disturbance mechanisms on augmenting the diversities among member classifiers, disturbing feature subspace and disturbing model parameters is analyzed. Then, two ensemble algorithms are proposed based on the double disturbance mechanism. The common character of them is that, member classifier is generated by disturbing feature subspace and model parameters, and the finial decision is made by the majority voting procedure. The experimental results show that both algorithms have the ability of improving the generalization performance of SVM significantly because they reduce the bias part and the variance part of the error simultaneously.

Key words: Support vector machine; Ensemble algorithm; Double disturbance mechanism; Member classifier

1 引 言

支持向量机(SVM)是在统计学习理论基础上发展起来的一种新型机器学习方法[1],具有泛化能力强、维数不敏感等优点,适于求解高维、小样本、非线性情况下的模式分类和回归分析等问题.

为了进一步提高 SVM 的泛化性能,人们将集成技术用于其训练过程中[2]. 然而,Dong 等[3]指出,简单的 Bagging SVM 和 Boosting SVM 算法不能有效提高支持向量机的泛化性能,原因是 Bagging 和 Boosting 通常对不稳定的学习机有效,而 SVM 是稳定的机器学习方法[4]. 相反,一些学者指出,特殊设计的集成算法能够显著提升 SVM 的泛化性能,例如 Robert 等[5]提出的"特征 Bagging"算法和 Valentini 等[6]提出的"低偏差 Bagging"算法等.

本文首先分析了特征 Bagging 算法和低偏差 Bagging 算法能够提升支持向量机泛化性能的原因;然后通过将扰动输入特征空间和扰动模型参数两种扰动机制相结合,提出两种基于二重扰动机制的集成训练算法. 实验结果表明,将其用于训练 SVM 时能够显著提升分类器的泛化性能,明显优于 Bagging SVM 算法. 此外,本文通过"偏差-方差"分析技术解释了两种集成训练算法能够提升泛化性能的原因.

2 集成学习和 Bagging 算法

收稿日期: 2007-05-15; **修回日期**: 2007-09-19.
基金项目: 国家自然科学基金项目(69732010, 60272095).
作者简介: 贾华丁(1966—),男,成都人,博士生,从事图像处理与模式识别、智能信息处理等研究; 游志胜(1945—),男,成都人,教授,博士生导师,从事图像处理与模式识别、智能决策系统等研究.

集成学习的数学描述为:给定输入空间 $X \subset R^n$ 的某个样本集 $S = \{(x_1, y_1), \cdots, (x_l, y_l)\}$,其中 $x_i \in X$ 且 y_i 是 x_i 的目标值;获得 S 上的 q 个假设(或分类器) $f_1, f_2, \cdots, f_q$,并采用某种集成策略(包括线性方式和非线性方式) 将它们组合成集成假设 f_E;此后,对于任何测试样本 $x \in X$,由 f_E 对它进行预测.

Krogh 等[7] 的研究表明,集成假设的泛化误差满足如下关系式:

$$E = \overline{E} - \overline{A}. \tag{1}$$

其中:$\overline{E}$ 是成员假设的泛化误差的加权平均,体现了固有误差;$\overline{A}$ 是成员假设相对于集成假设 f_E 的泛化误差的加权平均,体现了成员之间的"差异性". 因此,设计集成算法时,一方面应使成员假设具有较高的精度;另一方面应使它们之间保持足够的差异性.

Bagging 是一种直接对训练样本集进行扰动的集成算法[4]. 它采用 bootstrap 技术,使得成员分类器的训练集之间产生明显差异(约有 37.7 % 的相异样本). 对于不稳定的机器学习方法(如决策树和神经网络),训练的成员分类器之间将产生明显的差异. 因此,Bagging 能够利用多数投票方法有效地缩减方差,从而提高了集成分类器的泛化性能.

其中,多数投票过程为:设 $\omega_j (j = 1, 2, \cdots, C)$ 是第 j 类的标记. 对于某个测试样本 x,每个成员 f_i 分别为其所预测的类别投上一票,每一类得票数目为 $N_j = \#\{i \mid f_i(x) = \omega_j\}$. 于是,得票最多的类即为集成分类器 f_E 关于样本 x 的预测结果,即

$$f_E(x) = \arg\max_j (N_j). \tag{2}$$

然而,支持向量机是稳定的机器学习方法[3],对于训练样本集的扰动不能产生具有明显差异的成员分类器. 因此,简单地将 Bagging 算法应用于支持向量机的训练一般不能获得非常好的泛化性能.

3　基于二重扰动机制的集成训练算法

文献[5,6] 指出,支持向量机对于输入特征空间和模型参数的扰动比较敏感. 这里结合上述两种扰动机制的优点,提出两种非常有效的基于二重扰动机制的集成训练算法,并用于支持向量机的训练.

3.1　扰动输入特征空间

研究表明,针对输入特征空间 $X \subset R^n$ 进行扰动能更有效地增大成员分类器之间的差异. 根据文献[5] 的解释,当 X 的维数较高时,特征之间存在冗余或者相关性. 在特征子空间中训练的成员分类器不会显著地降低精度,相反,因特征子空间是随机选取的,不同的成员分类器更倾向于关注问题域的不同侧面,故它们之间的差异性是显著的.

Robert 等将采用扰动输入特征空间方式的集成算法称为"特征 Bagging (AB)"算法. 它实际是将 Bagging 算法推广到输入特征空间,使得成员分类器在不同的特征子空间中训练.

文献[8] 的研究指出,AB 算法对于提升支持向量机的泛化性能是非常有效的. 原因在于,SVM 对于输入特征空间的扰动十分敏感,在不同特征子空间中生成的成员分类器之间具有显著差异,而利用多数投票方法恰好能够对差异性带来的方差进行缩减,从而提高 SVM 的泛化性能.

3.2　扰动模型参数

支持向量机的模型参数(如高斯核参数 σ 和罚参数 C) 与它所能获得的泛化能力密切相关[1].

Valentini 通过大量实验研究了对模型参数进行扰动后支持向量机的期望误差、偏差和方差的变化趋势[6]. 其中,期望误差 $\mathrm{Err} = \mathrm{E}^x[L(y, t)]$ 在"0-1"损失函数 $L(\cdot, \cdot)$ 下可分解成 3 项,即

$$\mathrm{Err} = \mathrm{E}^x[N(x)] + \mathrm{E}^x[B(x)] + \mathrm{E}^x[V(x)]. \tag{3}$$

其中:$N(x)$,$B(x)$ 和 $V(x)$ 分别表示期望误差的固有噪声部分、偏差部分和方差部分.

Valentini 的实验结论指出,扰动模型参数能够显著影响支持向量机的期望误差、偏差和方差(如图 1 所示). 并且,模型参数在 (σ, C) 的取值空间内,存在一个"低偏差区域",使得 (σ, C) 在该区域内取值时,产生的成员分类器具有低偏差的特性.

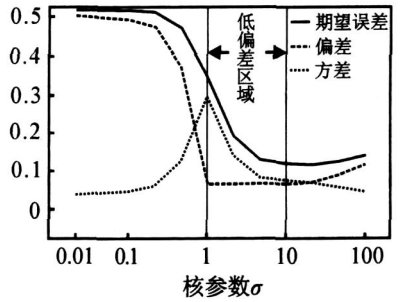

图 1　Letter 数据集上 SVM 分类器的期望误差、偏差和方差随核参数 σ 的变化曲线($C = 10$)

据此,Valentini 设计了一种新型集成算法用于实现支持向量机的训练,称其为"低偏差 Bagging"(或 LoBag). 该算法相当于 Bagging 算法在模型参数空间中的推广,使得成员分类器在"低偏差区域"随机选取模型参数对 (σ, C) 进行训练.

由分析可知,通过随机扰动模型参数,一方面可使成员分类器具有低偏差的特点;另一方面,成员之间由于模型参数的不同而具有一定的差异,可利用多数投票方法缩减方差. 因此,产生的集成分类器同时具有低偏差和低方差的特点,根据式(3) 的解释,它容易获得更好的泛化性能.

3.3 随机特征 Bagging 集成训练算法

根据前文分析,AB 算法和 LoBag 算法的关键是利用不同的样本扰动机制,使得产生的成员分类器之间具有显著的差异性,然后利用多数投票方法对方差进行缩减. 显然,两种扰动机制起作用的机理不同,因此可以对它们进行组合.

本文设计如下二重扰动机制:1) 对 n 维输入特征空间 X 进行扰动,使得随机生成的特征子空间 X_i 的维数 $m \approx 0.5n$;2) 在每个特征子空间 X_i 内,从"低偏差区域"$\mathrm{Reg}_{\mathrm{low}}$ 中随机选取一个模型参数对 (σ_j, C_j),并训练一个成员分类器.

显然,二重扰动机制是 3.1 和 3.2 节的两种样本扰动机制的并联组合. 采用该机制,可给出 SVM 的一种新型集成训练算法,称为"随机特征 Bagging (或 RAB)"算法. 随机的含义是指,特征子空间和模型参数分别随机地从输入特征空间 X(采用无放回 m 次重抽样的方式) 和低偏差区域中选取.

算法 1(RAB 算法)

输入:训练样本集 $S = \{x_i, y_i\}_{i=1}^{l}$;

1) for $i = 1,2,\cdots,q$;

2) 从输入特征空间 X 的 n 个特征中随机选取 $m \approx 0.5n$ 个构成特征子空间 X_i,并将 S 在 X_i 上投影,得到维数为 m 的新训练集 S_i;

3) 从低偏差区域 $\mathrm{Reg}_{\mathrm{low}}$ 中随机选取模型参数对 (σ_j, C_j);

4) 以 (σ_j, C_j) 作为模型参数在 S_i 上训练成员分类器 f_i;

5) 将成员分类器 $f_i (i = 1,2,\cdots,q)$ 按多数投票方法获得集成分类器 f_E.

经观察发现,RAB 算法与 AB 算法很相像,差别仅在于特征子空间中的模型参数是经过随机扰动过的. 因此,与 AB 算法相比,RAB 算法产生的成员分类器具有低偏差的特点,并且采用的二重扰动机制容易使得成员之间产生更大的差异性. 因而,RAB 算法更有利于提高分类器的泛化性能.

最后,判断某个随机扰动后的模型参数对 (σ_j, C_j) 是否位于"低偏差区域"内. 可按如下方法进行:假设 (σ^*, C^*) 是最优模型参数对,由它产生的分类器的偏差为 $(\mathrm{E}^x[B(x)])^*$,若由 (σ_j, C_j) 训练的分类器的偏差小于该值,则说明它位于"低偏差区域".

3.4 二维随机 Bagging 集成训练算法

在 RAB 算法中,两种样本扰动机制同时产生作用. 如果让两种扰动机制依次产生作用,则可得到另一种二重扰动机制:1) 对 n 维输入特征空间 X 进行扰动,使得随机生成的特征子空间 X_i 的维数 $m \approx 0.5n$; 2) 在每个 X_i 内采用扰动模型参数的方式训练 p 个成员分类器 $f_{i,j}, j = 1,2,\cdots,p$.

显然,该二重扰动机制是 3.1 和 3.2 节的两种样本扰动机制的串联组合. 同样,可给出基于该扰动机制的集成训练算法. 因最终的集成分类器 f_E 是经过两次多数投票过程获得的,故将该算法称为支持向量机的"二维随机 Bagging"集成训练算法 (2D-RBagging 算法).

算法 2(2D-RBagging 算法)

输入:训练样本集 $S = \{x_i, y_i\}_{i=1}^{l}$;

1) for $i = 1,2,\cdots,q$;

2) 从输入特征空间 X 的 n 个特征中随机选取 $m \approx 0.5n$ 个构成特征子空间 X_i,并将 S 在 X_i 上投影,得到维数为 m 的新训练集 S_i;

3) for $j = 1,2,\cdots,p$;

4) 从低偏差区域 $\mathrm{Reg}_{\mathrm{low}}$ 中随机选取模型参数对 (σ_j, C_j);

5) 以 (σ_j, C_j) 作为模型参数在 S_i 上训练成员分类器 $f_{i,j}$;

6) 将成员 $f_{i,j} (j = 1,2,\cdots,p)$ 按多数投票方法获得特征子空间 X_i 中的集成分类器 $f_E^{X_i}$;

7) 将 $f_E^{X_i} (i = 1,2,\cdots,q)$ 按多数投票方法获得最终的集成分类器 f_E.

可见,2D-RBagging 算法由嵌套的内外两个循环组成. 其中,内循环是 LoBag 算法,它通过扰动模型参数能够同时缩减 $f_E^{X_i}$ 的偏差和方差;而外循环则是以 $f_E^{X_i}$ 作为成员分类器的 AB 算法,由于每个 $f_E^{X_i}$ 分别属于不同的特征子空间,它们之间仍存在较大差异,使得算法的步骤 7) 对 $f_E^{X_i}$ 进行集成时仍能进一步缩减方差.

由分析可知,与 AB 算法相比,2D-RBagging 算法相当于采用具有低偏差特点的特征子空间 X_i 中的集成分类器 $f_E^{X_i}$ 作为成员,因而能够获得更优的泛化性能. 与 LoBag 算法相比,2D-RBagging 算法相当于在前者的基础上,再一次利用 $f_E^{X_i}$ 之间的差异性缩减方差,因而在泛化性能上更具优势.

4 数值实验和结果分析

在 9 个来自 UCI 数据库的小样本数据上进行实验,以验证 RAB 算法和 2D-RBagging 算法的性能,并将它们与 Bagging,AB 以及 LoBag 算法的性能进行比较.

实验中,所有成员分类器均由传统的 $\mathrm{SVM}^{\mathrm{light}}$ 算法[9] 训练获得. 并且选用高斯核函数,相应的最优模型参数 (C 和 σ) 由 5-fold 交叉验证方法得到. 重复抽样次数 p 和 q 均按照文献[5,6] 的建议设置为 25.

对于每个数据集,随机选取 50 % 的样本用于训

练,剩余部分则用于测试,并且实验重复 10 次. 最后,每种待比较的集成算法的泛化性能通过测试集上的平均测试精度进行评价.

表 1 统计了上述 5 种集成训练算法的主要实验结果. 从实验结果可以观察到以下现象:

1) Bagging 算法不能有效提升 SVM 的泛化性能,在 ionosphere 和 chess 上甚至降低了测试精度.

2) AB 算法的测试精度在多数情况下均明显优于 Bagging,特别是当样本维数较高时,例如 sonar, ionosphere, DNA 数据集.

3) LoBag 算法的测试精度在各种情况下同样优于 Bagging,并且在 tic 上取得最优的测试精度.

4) 本文提出的 RAB 算法和 2D-RBagging 算法所获得的测试精度,绝大多数情况下均比 Bagging, AB, LoBag 以及 SVM^{light} 算法更优,仅在 tic 数据集上比 LoBag 算法稍差.

表 1　5 种集成训练算法的测试精度比较
(包括传统的 SVM^{light} 算法)

数据集	Bagging	AB	LoBag	RAB	2D-RBagging	SVM^{light}
sonar	89.95	90.81	90.50	91.24	91.47	89.81
heart	83.80	83.85	84.21	84.32	84.78	83.47
ionosphere	94.00	94.48	94.30	94.55	94.55	94.06
diabetes	77.83	78.16	78.08	78.64	78.91	77.68
tic	86.02	85.85	86.40	86.16	86.31	85.90
german	75.29	76.07	75.76	76.48	76.64	74.52
DNA	95.61	96.60	95.82	96.90	96.85	95.28
segment	97.36	97.64	97.90	98.05	98.23	97.21
chess	98.29	98.76	98.64	99.11	99.17	98.40

上述现象表明,采用扰动输入特征空间机制和扰动模型参数机制的集成训练算法,对于提升支持向量机的泛化性能均是有效的;而采用二重扰动机制的 RAB 算法和 2D-RBagging 算法则更能显著地提升分类器的泛化性能.

期望误差的"偏差-方差"分析技术是研究集成算法工作机理的重要工具[6]. 下面的实验将分析和比较上述集成算法对于缩减 SVM 分类器的偏差和方差所起的作用. 其中,对期望误差、偏差和方差的估计由如下 bias_variance_analysis 过程获得.

过程 1 (bias_variance_analysis)

输入: 训练集 $S = \{x_i, y_i\}_{i=1}^{l}$ 满足 $x_i \in X \subset R^n$, $y_i \in \{\tilde{\omega}_1, \tilde{\omega}_2, \cdots, \tilde{\omega}_c\}$; 算法 L; 抽样次数 $s = 100$; 样本的抽样概率 $p = 0.5$.

输出: 期望误差 Err, 偏差 $E^x[B(x)]$ 和方差 $E^x[V(x)]$ 的估计值.

1) 重复如下步骤 s 次:

依概率 p 将 S 中的样本随机抽样出训练集 D_i, 剩余作为测试集 T_i; 由算法 L 在 D_i 上训练分类器 f_i.

2) for each $x \in S$,

$$t_m = \arg\max_{\tilde{\omega}_j} \frac{\#\{i \mid f_i(x) = \tilde{\omega}_j,\ x \in T_i\}}{\#\{i \mid x \in T_i,\ i = 1, \cdots, s\}}, \quad (4)$$

$$B(x) = \begin{cases} 1, & t_m \neq y, \\ 0, & t_m = y, \end{cases} \quad (5)$$

$$V(x) = \frac{\#\{i \mid t_m \neq f_i(x),\ x \in T_i\}}{\#\{i \mid x \in T_i,\ i = 1, \cdots, s\}}, \quad (6)$$

$$\mathrm{Err}(x) = B(x) + V(x). \quad (7)$$

3) 计算

$$\mathrm{Err} = \sum \mathrm{Err}(x)/l,$$

$$E^x[B(x)] = \sum B(x)/l,$$

$$E^x[V(x)] = \sum V(x)/l.$$

分别计算 RAB, 2D-RBagging, Bagging, AB 和 LoBag 算法在上述数据集上训练的集成分类器的偏差和方差. 图 2 和图 3 分别给出了它们与 SVM^{light} 算法所取得结果的相对值.

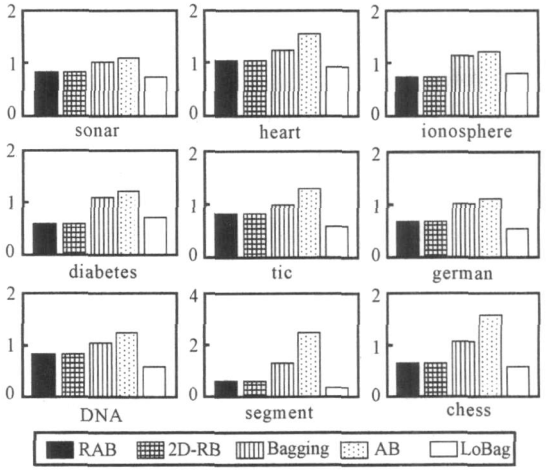

图 2　5 种集成算法的相对偏差的比较

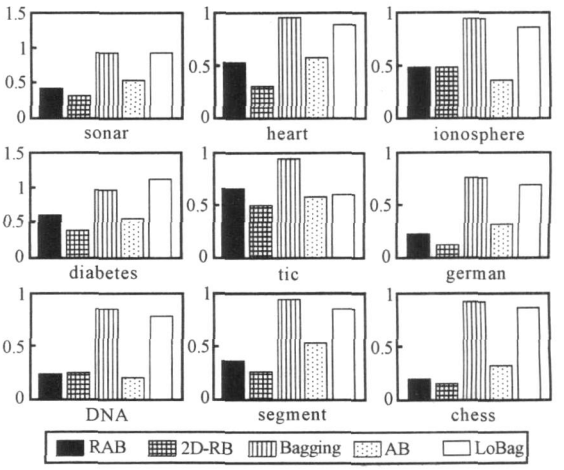

图 3　5 种集成算法的相对方差的比较

显然,在绝大多数情况下,RAB 算法和 2D-RBagging 算法均能有效地同时缩减分类器的偏差和方差,这是它们取得优异测试精度的根本原因.比较而言,Bagging 算法不能明显地缩减方差,相反在某些情况下却增大了偏差;AB 算法能明显地缩减方差,但多数情况偏差较大;LoBag 算法能明显地缩减偏差,但对方差的缩减有限.

此外,RAB 算法和 2D-RBagging 算法相比,后者明显能更大幅度地缩减方差,原因是它采用了两次主投票过程对成员分类器进行集成.

5 结 论

特殊设计的集成算法能够有效提升支持向量机的泛化性能.本文首先分析了扰动输入特征空间和扰动模型参数两种机制对于增大成员分类器间差异性的作用;然后将它们进行组合,得到两种基于二重扰动机制的集成训练算法.实验结果表明,两种新算法均能显著提升 SVM 分类器的泛化性能.通过"偏差-方差"分析解释了其中的原因:两种新算法能够同时缩减误差的偏差部分和方差部分.

参考文献(References)

[1] Vapnik V N. 统计学习理论的本质[M]. 北京: 清华大学出版社, 2000.
(Vapnik V N. The nature of statistical learning theory [M]. Beijing: Tsinghua University Press, 2000.)

[2] Kim H, Pang S, Je H, et al. Constructing support vector machine ensemble [J]. Pattern Recognition, 2003, 36(12): 2757-2767.

[3] Dong Y S, Han K S. A comparison of several ensemble methods for text categorization[C]. IEEE Int Conf on Services Computing. Shanghai, 2004: 419-422.

[4] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[5] Robert B, Ricardo G O, Francis Q. Attribute Bagging: Improving accuracy of classifier ensembles by using random feature subsets[J]. Pattern Recognition, 2003, 36(6): 1291-1302.

[6] Valentini G, Dietterich T. Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods[J]. J of Machine Learning Research, 2004, 5(6): 725-775.

[7] Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning[C]. Advances in Neural Information Processing Systems. Denver, 1995: 231-238.

[8] Tao D C, Tang X O, Wu X. Asymmetric Bagging and random subspace for SVM-based relevance feedback in image retrieval[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2006, 28(7): 1088-1099.

[9] Joachims T. Making large-scale SVM learning practical [C]. Advances in Kernel Methods: Support Vector Learning. Cambridge: MIT Press, 1999.

(上接第 827 页)

[3] Yang G H, Yee J S, Wang J L. An iterative LMI method to discrete-time state-feedback controller design with mixed H_2/H_∞ performance [J]. European J of Control, 2002, 8(2): 126-135.

[4] Yang G H, Lam J, Wang J L. Reliable H_∞ control for affine nonlinear systems[J]. IEEE Trans on Automatic Control, 1998, 43(8): 1112-1117.

[5] Takatgi T, Sugeno M. Fuzzy identification of systems and its application to model and control[J]. IEEE Trans on Systems, Man and Cybernetics, 1985, 15(1): 116-132.

[6] Cao S G, Rees N W, Feng G. Stability analysis and design for a class of continuous time fuzzy control systems[J]. Int J of Control, 1996, 64(3): 1069-1087.

[7] Hui-Ning Wu. Reliable mixed fuzzy static output feedback control for nonlinear systems with sensors faults[J]. Automatica, 2005, 41(6): 1925-1932.

[8] Feng G, Cao S G, Rees N W, et al. Design of fuzzy systems with guaranteed cost control stability[J]. Fuzzy Sets and Systems, 1997, 85(1): 1-10.

[9] Kim S W, Seo C J, Kim B K. Robust and reliable H_∞ controllers for discrete-time systems with parameter uncertainty and actuator failure[J]. Int J of Systems Science, 1999, 30(12): 1249-1258.

[10] Liu Xiaodong, Zhang Qingling. New approaches to H_∞ controller designs based on fuzzy observers for T-S fuzzy systems via LMI[J]. Automatica, 2003, 39(4): 1571-1582.

[11] Liu Guo-yi, Zhang Qing-ling, Zhai Ding. LMI-based H_2/H_∞ mixed controller design for T-S fuzzy systems [J]. Control and Decision, 2007, 22(9): 1032-1034.

[12] Zhai Ding, Zhang Qing-ling, Liu Guo-yi. Robust non-fragile controller for a class of linear time-delay systems [J]. Control and Decision, 2006, 21(5): 559-562.

[13] Zhai Ding, Zhang Qing-ling. Investigation on management information system of transport and sales for enterprises[J]. Control and Decision, 2002, 17(S): 837-839.