随机森林针对小样本数据类权重设置

李建更,高志坤
LI Jian-geng,GAO Zhi-kun

北京工业大学 人工智能与机器人研究所,北京 100124
Institute of Artificial Intelligence and Robotics,Beijing University of Technology,Beijing 100124,China
E-mail:gaozhikun1314@emails.bjut.edu.cn

LI Jian-geng,GAO Zhi-kun.Setting of class weights in random forest for small-sample data.Computer Engineering and Applications,2009,45(26):131-134.

Abstract: Random forest has been proved to be an efficient algorithm for classification and feature selection in bioinformatics.Although the effect of parameter setting on results is very limited,a group of appropriate parameters can generate excellent performance.This paper focuses on the setting of class weights in random forest to deal with classification and feature selection problems of unbalanced small-sample data and determines the optimal class weight.In order to compare the performance of feature selection with different weights,SVM is applied in the paper.The results show that optimal class weight is variable and cannot form a standard.However,people can find some weights with which not only classification but also feature selection can get better performance.

Key words: random forest;class weight;small-sample;Support Vector Machine(SVM);feature selection

摘 要: 随机森林已经被证明是一种高效的分类与特征选择方法。尽管参数的设置对结果影响较小,但合适的参数可以使分类器得到理想的效果。主要针对癌症研究中小样本不均衡数据的分类和特征选择问题,研究了随机森林中类权重的设置。为了比较在不同的类权重下特征选择的效果,同时使用支持向量机(Support Vector Machine,SVM)方法。最终结果显示最优的类权重是不确定的。最后总结出几条规律指导研究者选择合适的权重使分类和特征选择效果得到改善。

关键词: 随机森林;类权重;小样本;支持向量机;特征选择
DOI: 10.3778/j.issn.1002-8331.2009.26.038 **文章编号:** 1002-8331(2009)26-0131-04 **文献标识码:** A **中图分类号:** TP391

1 前言

随机森林[1]是由 Breiman 首先提出的一种基于树分类器的分类算法,该方法具有很多优点,如不需要预处理,不会过拟合(Overfitting),同时进行特征基因的选择等。对于普通数据,使用时通常只需设置 3 个参数:终节点规模(nodesize),每个分离点选择的变量个数(mtry)及树分类器的个数(ntree)便可以得到较理想的结果。但实际问题中数据大多是不均衡的[2-3],即数据中类与类之间所含样本数相差很大。许多学者已经对不均衡问题进行了大量的研究并提出了许多方法,如过采样与欠采样技术[4-5]、代价敏感学习方法[6]、类均衡法[7]以及集成算法[8]等。过采样通过复制小类样本增加该类的样本数,这容易带来过拟合的问题。文[9]提出了一种新方法 SMOTE,该方法人工生成一些小样本代替了简单地复制,从而避免了过拟合。与过采样相反,欠采样减少大样本数目以达到均衡目的。去掉多少样本及去掉哪些样本是欠采样要考虑的问题,文[5]使用 one-sided selection 法将大类样本分为“安全样本”、“边界样本”和“噪音样本”,然后利用 Tomek-link 法去掉后两种。代价敏感学习法考虑各类的错分代价而非错分率,这种方法的意义在医学上有明显的体现。文[6]将各类样本数目之比设置为类错分代价的反比,使算法具有代价敏感性。类别均衡是把所有小类合并为一个或几个较大的类,使新类具有和原数据中的大类相同的数量级[7]。集成算法通过集合多个分量分类器降低错分风险,文[8]提出了 EasyEnsemble 和 BalanceCascade 两种方法解决类别不均衡问题。其中,前者是从大类中随机选择与小类相同数量级的样本,与小类组成训练数据;后者每次去掉大类中被正确分类的样本,直至训练样本达到均衡。随机森林(Random Forest,RF)在处理不均衡数据时有天然的优势,该方法本身属于集成算法,通过拔靴法(bootstrap)选择数据构建分量分类器(树分类器)。由于最终结果由各分量分类器投票(voting)得出,能在一定程度上减少因不均衡数据产生的影响。Chen C 等提出了两种基于随机森林的方法:一种是加权随机森林(Weighted Random Forest,WRF);另一种为均衡随机森林(Balanced Random Forest,BRF)[10]。WRF 将类权重与 RF 结合起来;BRF 注重改进 bootstrap 技术,使选择的训练数据达到均衡。

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.60234020);北京市自然科学基金(the Natural Science Foundation of Beijing City of China under Grant No.4092021);北京市教育委员会科技计划项目(No.JC002011200903)。
作者简介: 李建更(1965-),男,教授,硕士生导师,主要研究方向:生物信息学、自动控制等;高志坤(1982-),男,硕士生,主要研究方向:生物信息学。
收稿日期: 2008-05-15 **修回日期:** 2008-09-01

该文继续了 Chen C 等的研究，讨论了利用 WRF 在对小样本数据进行分类和特征提取时类权重的设置，使用支持向量机(Support Vector Machine，SVM)验证特征选择的结果。此外，也涉及到多类问题的类权重设置。因为主要研究 WRF 在癌症中的应用，故实验中所选数据绝大多数与癌症有关。实验结果显示最优的类权重是可变的，但可以总结出几个规律以寻找合适的权重，从而在一定程度上改进分类器性能。

2 方法

在 6 组数据上使用 WBF，计算袋外预测错误(Out-of-Bag error，OOB error)，同时得到特征属性集。在得到的特征集上运用 SVM 检验特征选择的效果。

2.1 加权随机森林

加权随机森林(Weighted Random Forest，WBF)是基于 RF 的一种改进的预测分类器，其本质是赋给小类较大的权重而给大类相对小的权重。这一措施为小类的错分设置了更大的惩罚。类权重在两个地方影响 RF 的结果，一个是树分类器的生长过程。这个过程中每个子节点(非终节点)最佳分离值的寻找涉及类权重的计算；另一个是在终节点。类权重直接参与每个终节点的类标签的确定。类权重在 WRF 中扮演关键作用，通常根据 OOB error 和类之间样本数之比来确定。

在树的生长过程中，加权 Gini 不纯度被选择用来寻找分离点。具体的计算如公式(1)(2)所示：

$$i(N)=\frac{\sum_{i=1}^{C}(n_i W_i)^2}{\sum_{i=1}^{C} n_i W_i} \tag{1}$$

$$\Delta i=i(N)-i(N_L)-i(N_R) \tag{2}$$

公式中 N 代表未分离的节点；N_L 和 N_R 分别表示分离后的左侧节点和右侧节点(*RF* 中采用“二叉”树)；W_i 为 C 类样本的类权重；n_i 表示节点内各类样本的数量；Δi 是不纯度减少量，该值越大代表分离点分离效果越好。

在终节点，类权重用来决定其类标签。具体公式如(3)所示：

$$nodeclass=\arg\max_i (n_i W_i) \quad (i=1,2,\cdots,C) \tag{3}$$

2.2 支持向量机

SVM[11]是一种基于统计学习理论的机器学习方法，主要用于小样本分类的研究。该方法首先将数据由低维空间转换到高维，然后计算超平面将两类数据分离。由于癌症数据通常包含成千上万的属性(一般为不同的基因表达值)，分类之前需要利用 WRF 进行特征基因的选择以适应 SVM 的需要。同时，SVM 最初被设计用来分离两类问题，为了解决多类问题的分类，采用了扩展的 SVM 算法[12]。该算法利用 OVR(one-versus-rest)选择某一类而将其余类合为一类，然后运行 SVM。重复以上过程直到所有类都被分离。目前已经有关于 SVM 的 MATLAB 工具箱可供使用，如 SVM_steveGunn、LS_SVMlab 等。

3 实验及结果分析

3.1 实验数据

为了比较不同类权重下分类及特征选择的效果，使用了 6 组不同的数据，数据属性如表 1 所示，其中包括数据名称(Name)、样本数(n)、类别数(c)、样本向量维数(a)、数据类型(T)和缺失率(Miss)。肺癌(Lung)、白血病(Leukemia)和前列腺癌(Prostate)数据来自 BROAD 数据库[13]；结肠癌(Colon)和白酒(Wine)数据来自 UCI Machine Learning Repository[14]；胃癌(Gastric)数据由北京肿瘤防治所提供。所有样本均为小样本数据，实验过程将一个数据集分为包含不同样本数的几组，针对每组数据赋予一系列不同的权重。

表 1 实验数据

数据名称	样本数 n	类别数 c	样本向量维数 a	数据类型 T	缺失率/(%)
Colon	62	2	1 712	Num	0
Leukemia	72	2	7 129	Num	0
Prostate	102	2	2 766	Num	0
Gastric	38	3	21 378	Num	6.84
Lung	197	4	1 000	Num	0
Wine	178	3	13	Num	0

3.2 实验结果

3.2.1 WRF 分类结果

为了比较类权重对分类器的影响，实验中设置 *nodesize*=1，*mtry*=150(在 Wine 中为 13)及 *ntree*=1 200，并保持这些参数值不变。每次实验 WRF 均会产生 OOB 错误率和变量重要性分析，前者用于评价分类器分类性能；后者是特征选择的依据。所有 OOB 错误率见表 2~表 7。表中每一行代表同一数据集在不同的类权重下分类结果；每一列代表同一权重下不同数据集的分类结果。OOB 错误率后小括号内数字分别代表各类中的错分数，黑色粗体数字代表最小错误率。

表 2 Colon

OOB errs (100%)		Class weights(w_1,w_2) (1,1)	(2,1)	(3,1)	(4,1)	(5,1)
Sample ratios	1:2	12.9(3,5)	**12.9**(3,5)	14.5(2,7)	14.5(1,8)	20.97(0,13)
	1:3	13.2(4,3)	13.2(3,4)	13.2(2,5)	13.2(1,6)	**11.32**(1,5)
	1:4	**12**(3,3)	14(2,5)	14(2,5)	14(2,5)	14(2,5)
	1:5	14.58(4,3)	12.5(2,4)	14.58(2,5)	14.58(2,5)	**10.42**(1,4)

表 3 Leukemia

OOB errs (100%)		Class weights(w_1,w_2) (1,1)	(2,1)	(3,1)	(4,1)	(5,1)
Sample ratios	1:2	2.77(1,1)	2.77(0,2)	2.77(0,2)	4.166(0,3)	5.55(0,4)
	1:3	**1.578**(1,0)	3.17(0,2)	3.17(0,2)	4.762(0,3)	4.762(0,3)
	1:4	3.39(2,0)	**1.695**(0,1)	3.39(0,2)	4.762(0,3)	6.779(0,4)
	1:5	3.57(2,0)	**0**(0,0)	0(0,0)	3.571(0,2)	5.357(0,3)

表 4 Prostate

OOB errs (100%)		Class weights(w_1,w_2) (1,1)	(2,1)	(3,1)	(4,1)	(5,1)
Sample ratios	1:2	**8.974**(6,1)	14.10(3,8)	16.66(1,12)	19.23(1,14)	20.51(1,15)
	1:3	7.246 3(4,1)	**7.246 3**(2,3)	13.043(2,7)	17.39(1,11)	17.39(1,11)
	1:4	12.31(7,1)	**6.153 8**(2,2)	7.692 3(2,3)	16.92(2,9)	16.92(2,9)
	1:5	8.064(5,0)	**4.838 7**(2,1)	9.677 4(2,4)	12.903(1,7)	11.290(1,6)

表 5 Gastric

OOB errs (100%)		Class weights(w_1,w_2,w_3) (1,1,1)	(1,2,4)	(1,1,2)	(1,1,3)	(1,1,4)
Sample ratios	20:13:5	5.263 1 (0,0,2)	**0** (0,0,0)	0 (0,0,0)	0 (0,0,0)	0 (0,0,0)

表 6 Lung

OOB errs (100%)		Class weights(w_1,w_2,w_3,w_4) (1,1,1,1)	(1,8,7,7)	(1,2,2,1)	(1,3,3,1)	(1,1,2,1)
Sample ratios	139:17:21:20	4.060 9 (3,2,3,0)	6.598 9 (12,0,1,0)	**3.045 7** (3,1,2,0)	3.553 3 (5,1,1,0)	3.553 3 (3,2,2,0)

表 7　Wine

OOB errs (100%)		Class weights(w_1,w_2) (1,1,1)	(6,5,7)	(2,2,1)	(1,2,1)	(2,3,1)
Sample ratios	59:71:48	2.247 2 (1,3,0)	1.685 4 (0,3,0)	**1.685 4** (1,2,0)	3.370 7 (3,2,1)	2.247 2 (1,1,2)

3.2.2　SVM 分类结果

WRF 首先根据变量重要性分析将属性值(因数据主要为癌症,故文中属性值以基因表达值代替)排序,然后选择序列中前 30 个(Wine 数据选前 8 个)基因作为特征基因用于 SVM 分类。SVM 对各数据集的分类准确率见表 8~表 13。表中仍用黑色粗体数字代表最小错分率。除核函数外,实验过程中 SVM 的其他参数通过"留一"检验(Leave-One-Out Cross Validation,LOOCV)在一定范围内确定最佳值。各表中的结果均通过 LOOCV 得到。

表 8　Colon

Accuracy rate		Class weights(w_1,w_2) (1,1)	(2,1)	(3,1)	(4,1)	(5,1)
Sample ratios	1:2	0.838 7	0.822 6	**0.887 1**	0.838 7	0.854 8
	1:3	0.849 1	0.849 1	0.867 9	0.867 9	**0.886 8**
	1:4	**0.900 0**	0.800 0	0.860 0	0.840 0	0.880 0
	1:5	0.833 3	0.895 8	0.854 2	0.812 5	**0.916 7**

表 9　Leukemia

Accuracy rate		Class weights(w_1,w_2) (1,1)	(2,1)	(3,1)	(4,1)	(5,1)
Sample ratios	1:2	0.972 2	0.972 2	0.958 3	**0.986 1**	0.972 2
	1:3	0.984 1	0.984 1	0.984 1	0.984 1	**1.000 0**
	1:4	0.966 1	**1.000 0**	1.000 0	1.000 0	1.000 0
	1:5	0.964 3	**1.000 0**	1.000 0	0.982 1	1.000 0

表 10　Prostate

Accuracy rate		Class weights(w_1,w_2) (1,1)	(2,1)	(3,1)	(4,1)	(5,1)
Sample ratios	1:2	0.897 4	0.897 4	0.897 4	**0.910 3**	0.897 4
	1:3	0.942 0	**0.942 0**	0.942 0	0.927 5	0.942 0
	1:4	0.876 9	0.907 7	0.923 1	0.923 1	**0.953 8**
	1:5	0.903 2	**0.935 5**	0.919 4	0.935 5	0.935 5

表 11　Gastric

Accuracy rate		Class weights(w_1,w_2,w_3) (1,1,1)	(1,2,4)	(1,1,2)	(1,1,3)	(1,1,4)
Sample ratios	20:13:5	0.789 5	**0.921 1**	0.815 8	0.789 5	0.921 1

表 12　Lung

Accuracy rate		Class weights(w_1,w_2,w_3) (1,1,1,1)	(1,8,7,7)	(1,2,2,1)	(1,3,3,1)	(1,1,2,1)
Sample ratios	139:17:21:20	0.949 2	**0.959 4**	0.944 2	0.944 2	0.939 1

表 13　Wine

Accuracy rate		Class weights(w_1,w_2,w_3) (1,1,1)	(6,5,7)	(2,2,1)	(1,2,1)	(2,3,1)
Sample ratios	20:13:5	0.988 8	0.988 8	**0.988 8**	0.988 8	0.988 8

3.2.3　多类问题权重设置

多类问题权重设置比二类问题要复杂的多,单就计算量讲,遍历所有的权重便是相当大的工作,尤其是当类别数较大的时候。文中仅选择几组有代表性的类权重进行实验,实验结果见表 5~表 7 及表 11~表 13。表中第一组权重未改变原始数据;第二组权重设置为各类样本数的反比;其余各组权重设置与类错分率相结合,类错分率高者设置较大权重。

3.3　实验分析

从实验结果可以得到两个显然的事实:其一是对于绝大多数数据集,WRF 均能取得优于普通 RF 的结果。在每个表中,第一列的结果在绝大多数情况下不如其他几列,说明类权重的设置是十分必要的;其二是随着类权重的增加,该类的错分数是递减的。所有的结果均说明了这个事实,没有例外。此外,恰好与类样本数成反比的类权重一般都不能取得最佳的结果,而一类的样本数在总样本数中所占比例越高,该类错分数越少。同时,对比 WRF 与 SVM 的结果发现,若 WRF 结果较理想,则 SVM 的分类效果同样较好,即便两者不能同时达到最佳。说明合理设置权重能提高特征选择的效果。鉴于以上事实,文章总结了几条规律,以帮助研究者方便地选择权重。

规律 1　除非有特殊要求,否则不需设置太大权重。同时考虑 WRF 和 SVM 的结果,为小类设置 2 或 3 的权重便足够。较大权重值可能令特征选择的效果有所提高,但其代价是 WRF 的分类错误率会急剧增加,如表 4、表 10 所示。即便 2 或 3 的权重不是最优选择,实验显示至少它们为次优的。

规律 2　考虑特异性(Specificity)与敏感性(Sensibility)。敏感性为病人中得出阳性检测的样本占病人总数的百分比;特异性为健康人中得到阴性检测的样本占健康人总数的百分比。通常情况下二者是对立的,即一方增加必然导致另一方减少。若将小类看作病人样本,为了得到较高的敏感性,需要给该类设置较大的权重。经常被用作评价这两个指标的工具是"受者操作特性曲线"(Receiver Operating Characteristic Curve,ROC Curve)。在分类器分类准确率相等的情况下,ROC 曲线及曲线下面积(Area under the Curve,AUC)为判断分类器对一类样本的分类效果提供依据。图 1 显示的是设置不同权重的 WRF 分类器对结肠癌(Colon)数据中小类的分类 ROC 曲线,其中小类与大类样本比为 1:3(表 2)。随小类权重的增加,相应 AUC 的值分别为 0.808 7、0.834 6、0.869 6 及 0.886 5。AUC 值越大,分类器性能越好。

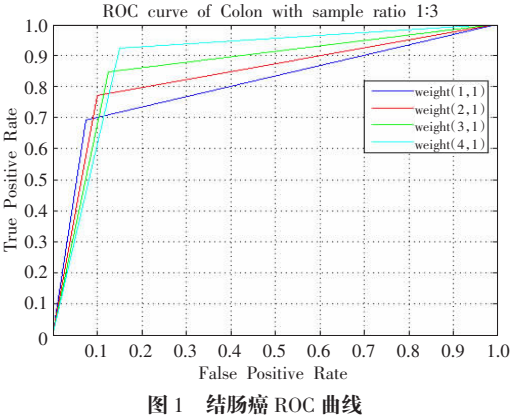

图 1　结肠癌 ROC 曲线

规律 3　考虑错分代价。当设置一个较大的类权重时,需要考虑哪类会产生大的错分代价,即如果类别 A 的错分代价是 B 的 n 倍,则权重可设为 n:1,因为实验结果表明,随类权重的增加该类错分率是减少的。这与代价敏感学习算法有相似之处,所不同的是后者改变的是样本数而非权重。此外,实验表明高分类准确率需要较小的权重而理想的特征选择效果需要相对

较大的权重。从 RF 的内部机制分析,尽管权重设置不能提高小类样本被选中的概率,但一旦有小类样本被选中,权重会直接参与分类与特征选择过程,使结果避免过多地向大类倾斜。

规律 4 多类问题的类权重设置要结合样本错分率及类之间样本数关系。对于错分数多或样本数少的类,可以设置较大权重。通常选择 2 或 3 作为这些类的权重,不需设置太大。

4 结论

进一步研究了 WRF 针对小样本数据类权重的设置问题。实验结果证明 WRF 比普通 RF 方法无论在分类还是特征选择上都能取得更优的效果。详细对比了在不同样本数之比条件下各种类权重对分类和特征选择的影响,发现设置很大的权重是没有必要的。尽管没有一个严格的标准可供使用,该文总结了几条规律来帮助研究者方便地选择权重。今后的工作需要寻找一种更好的方法来处理多类问题的权重设置。此外,将 WRF 与样本选择技术结合起来也是将来研究的重点。

参考文献:

[1] Breiman L.Random forest[J].Machine Learning,2001,45:5-32.

[2] Stolfo S J,Fan D W S,Lee W,et al.Credit card fraud detection using meta-learning:Issuesand initial results[C]//AAAI-97 Wrokshop on AI Methods in Fraud and Risk Mangement,1997.

[3] Pednault E P D,Rosen B K,Apte C.Handling imbalanced data sets in insurance risk modeling,Technical Report RC-21731[R].IBM Research Report,2000-03.

[4] Batista G E A P A,Bazzan A L C.Balancing training data for automated annotation of keywords:A case study[C]//Proc of the Second Brazilian Workshop on Bioinformatics,SBC,2003.

[5] Kubar M,Matwin S.Addressing the course of imbalanced training sets:One-sided selection[C]//Proceedings of 14th International Conference in Machine Learning,San Francisco,CA,1997:179-186.

[6] Breiman L,Freidman J.Classification and regression trees [M].[S.l.]:Wadsworth,1984.

[7] 张启蕊,张凌,董守斌,等.训练集类别分布对文本分类的影响[J].清华大学学报:自然科学版,2005,45(9):1802-1805.

[8] Liu X Y,Wu J.Exploratory under-sampling for class-imbalance learning[C]//Proceedings of the 6th IEEE International Conference on Data Mining(ICDM'06),Hong Kong,China,2006.

[9] Chawla N V,Bowyer K W.SMOTE:Synthetic minority over-sampling technique[J].Journal of Artificial Intelligence Research,2002,16:321-357.

[10] Chen C,Liaw A,Breiman L.Using random forest to learn imbalanced data,Technical Report 666[R].Statistics Department,University of California at Berkeley,2003.

[11] Vapnik V N.Statistical learning theory[M].[S.l.]:John Wiley and Sons,1998.

[12] Weston J,Warkins C.Multi-class support vector machines,CSD-TR-98-04[R].Royal Holloway,University of London,1998.

[13] Broad institute[EB/OL].http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.

[14] UCI.Machine learning repository[EB/OL].http://archive.ics.uci.edu/ml/datasets.html.

(上接 29 页)

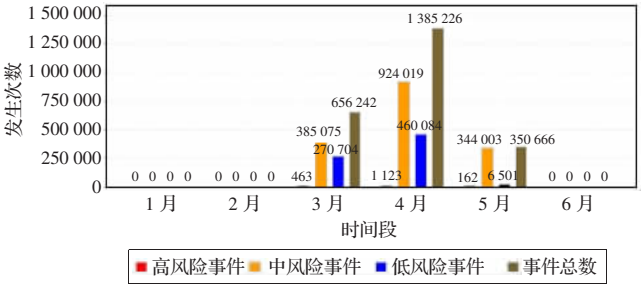

图 3 系统风险事件报告统计情况

从系统实现效果分析,“软件人”群网络安全风险评估模型具有以下优势:

(1)利用“软件人”的灵活机动的特点,通过“软件人”间的协调协作,能及时、动态地反应网络安全状况。

(2)利用“软件人”的自治性和学习性,不断完善自身的功能,减少人为的参与,并隐藏了复杂性,提高系统的运行效率。

(3)利用“软件人”根据风险分析结果,可针对动态的网络安全威胁及时做出响应,在提供分析结果的同时,可对安全设备的设置自动做出调整。

(4)利用“软件人”进行数据检测和风险评估,不需要像传统风险评估系统那样分散收集数据再集中处理,可将“软件人”派到主机上直接分析处理数据, 降低了对网络带宽的依赖程度,提高系统的服务能力和工作效率。

(5)根据网络系统的不同复杂程度,在不影响“软件人”正常工作的情况下,可以复制、繁衍相应数量的子“软件人”进行分布式部署,提高了系统的动态可扩展性和适应性。

6 结束语

该模型充分利用各“软件人”之间相互协调而又相互独立自治的特性,对网络系统安全对象进行风险评估,能动态地反应网络系统的安全性,并使系统结构具有良好的自治性、灵活性、扩展性、适应性、分布式控制和应急响应能力。有效地解决了传统风险评估系统中检测和评估模块不能在网络中动态移动、随需分布,限制了评估速度、效率和范围等问题。随着该模型系统的进一步完善,将使网络安全风险评估更为准确快捷。

参考文献:

[1] 曾广平,涂序彦.软件人[C]//中国人工智能学会第 10 届全国学术年会论文集.北京:北京邮电大学出版社,2003:677-682.

[2] Butler S A,Fischbeck P.Multi-attribute risk assessment,Technical Report CMD-CS-01-169[R].2001.

[3] Wooldridge M.An introduction to multi-agent systems[M].[S.l.]:John Wiley & Sons Inc,2002.

[4] 冯登国,张阳,张玉清.信息安全风险评估综述[J].通信学报,2004,25:10-18.

[5] Tu Xu-yan,Zeng Guang-ping,Tang Tao.Humanized autonomous decentralized system[C]//Proc of the Int Symposium on Autonomous Decentralized Systems,Sichuan,2005:593-598.

[6] Ma Zhong-gui,Ye Bin,Ban Xiao-juan,et al.The study on model and architecture of SoftMan group based on intelligent autonomous decentralized systems[C]//Proc of the Int Conf on Autonomous Decenralized Systems,Chengdu,2005:641-646.

[7] Ma Zhan-fei,Zheng Xue-feng,Zeng Guang-ping,et al.Multi-SoftMan coordination control in detection system[J].Conrol and Decision,2008,8:944-948.