交叉验证容噪分类算法有效性分析及其在数据流上的应用

张健沛,杨显飞,杨 静
(哈尔滨工程大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 交叉验证容噪分类算法是处理含噪音数据集分类问题的重要手段之一.从样本复杂度理论出发,对其有效性进行了详细的理论证明,并给出适用条件.提出一种容噪数据流集合分类算法,理论分析和实验验证表明,该算法与传统交叉验证容噪算法相比,具有更高的分类准确率.

关键词: 交叉验证;容噪;分类;集合分类器
中图分类号: TP301 **文献标识码:** A **文章编号:** 0372-2112 (2011) 02-0378-05

Effectiveness Analysis and Application in Data Streams of Cross Validation Noise-Tolerance Classification Algorithm

ZHANG Jian-pei, YANG Xian-fei, YANG Jing
(*College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*)

Abstract: Cross validation noise-tolerance classification algorithm is an important method which deals with noisy data set classification problem. According to sample complexity theory, cross validation noise-tolerance classification algorithm validity was proved and applied conditions was given. And noise-tolerance data stream ensemble classifiers was proposed in this paper. Theory and experiment indicated, in contrast with tradition cross validation noise-tolerance classification algorithm, this method had more prediction accuracy.

Key words: cross validation; noise-tolerance; classification; ensemble classifiers

1 引言

传统机器学习研究的分类算法假定训练数据是清洁的,但在现实应用中,大部分数据含有噪音.在有噪音的数据集上训练分类器,其分类准确率受两方面制约:训练数据的质量和分类算法的归纳偏置[1].因此,如何有效训练含有噪音的数据集是分类研究的重要内容之一[2~9].噪音包括属性噪音和类别噪音,研究表明在训练数据集上去除属性噪音建立的分类器,其分类准确率并不一定有所提高,而去除类别噪音分类准确率必会有所提高[10],因此本文仅考虑类别噪音.

容噪分类算法可分为两类:封装式和过滤式.封装式容噪分类算法是指修改学习算法本身使其具有容噪性.过滤式容噪分类算法则是在学习分类模型之前对数据集进行预处理,识别并删除噪音数据.由于过滤式容噪分类算法具有设计简单、噪音数据不能影响最终分类模型等优点,现有的大部分容噪分类算法设计都是基于此方法.

交叉验证容噪分类算法是一类典型的过滤式容噪分类算法.文献[5]首次提出了基于验证思想的容噪决策树分类算法,该算法反复利用决策树删除训练数据集中被错误分类的数据,并重新建立分类模型,最终获得没有噪音的数据集.为了提高噪音过滤的准确性,文献[6]利用集合分类器获得分类结果,当确定某部分数据是否为噪音时,采用剩余数据训练集合分类器对其进行判断,从而实现交叉验证.文献[7,8]在此框架下进一步进行了研究,提出了面向大数据集容噪分类算法及容噪集合分类算法.文献[9]对容噪决策树算法进行扩展,解决了归纳逻辑分类问题中的噪音处理问题.上述方法虽然都用到了验证或交叉验证思想剔除训练数据集中的噪音,但却没有对其有效性进行系统的理论分析.本文针对这一问题,从含噪音的样本复杂度理论出发,对交

收稿日期:2010-03-29;修回日期:2010-10-26
基金项目:国家自然科学基金(No.60873037,No.61073041,No.61073043)

叉验证容噪分类算法的有效性进行了详细的理论证明,并给出适用条件,最后在该理论指导下,提出一种容噪数据流集合分类算法 NTDSEC(Noise-Tolerance Data Stream Ensemble Classifiers).

2　交叉验证容噪分类算法及其有效性分析

设实例空间为 $\boldsymbol{R}$,训练数据集 $L=\{(x_1,y_1),(x_2,y_2)\cdots,(x_n,y_n)\}$,交叉验证容噪分类算法主要包括以下 4 个步骤:

步骤 1　通过 bootstrap 方法或非重复抽样方法对数据集 L 进行抽样,取得多个训练子集并加以训练,获得集合分类器.

步骤 2　对于每个数据(x_i,y_i),判断其类标是否与集合分类器的分类结果不一致.

步骤 3　对于不一致的训练数据,从训练数据集中删除.

步骤 4　判断其是否到达停止条件,如没有转步骤 1.

对于步骤 2,有两种方式判断其类标是否与集合分类器的分类结果不一致:一种是数据类标与所有个体分类器的分类结果都不一致,即完全不一致方法;另一种是与多数个体分类器的分类结果不一致,即多数不一致方法.同时,根据数据删除方式不同,也可将其分为有放回删除和无放回删除,有放回删除即删除的数据仅不参与下一轮训练,而无放回删除是指数据一经删除,将不再参与后续的训练.本文根据删除数据是否放回分别讨论其有效性及适用条件.

2.1　无放回交叉验证容噪分类算法有效性分析

Angluin 和 Laird 证明,在有噪音的数据集 L 上训练分类器,当分类准确率满足 PAC(Probably Approximately Correct)框架时,训练数据集大小需要满足式(1)[11]:

$$m\geqslant\frac{2}{\varepsilon^2(1-2\eta)^2}\ln(\frac{2N}{\sigma}) \tag{1}$$

其中 ε 表示有限假设空间中学习器的最坏情况分类错误率,σ 是一个较小的概率数值,η 为数据集 L 的最大噪音率,m 是 L 中数据的个数,N 表示假设空间包含的学习器个数.

将式(1)右侧乘以系数 μ,$\mu\geqslant1$,使得不等式转化为等式,则式(1)可改写为[12]:

$$m=\frac{2\mu}{\varepsilon^2(1-2\eta)^2}\ln(\frac{2N}{\sigma}) \tag{2}$$

因此,可以获得如下假设:

$$v=m(1-2\eta)^2=\frac{1}{\varepsilon^2}[2\mu\ln(\frac{2N}{\sigma})] \tag{3}$$

因为 $2\mu\ln(2N/\sigma)$是常数,所以学习器的最坏情况分类错误率平方与 v 成反比.

设 L_t 是第 $t-1$ 轮消除噪音后剩余的训练数据集,η_t 为数据集 L_t 含有的噪音率,p 是训练数据集 L_t 获得的集合分类器的分类准确率,S 为数据集 L_t 内类标与集合分类器分类不一致的数据集合.则在第 t 轮消除噪音后,数据集仍然含有的噪音率 η_{t+1}等于:

$$\eta_{t+1}=\frac{|L_t|\times\eta_t-|S|\times p}{|L_t|-|S|}$$

定理 1　如果数据集 L_t 所含的噪音率 $\eta_t<1/2$,且集合分类器的分类准确率 $p>1/2$,则 $\eta_{t+1}<1/2$

证明:$\eta_{t+1}-\frac{1}{2}=\frac{|L_t|\times\eta_t-|S|\times p}{|L_t|-|S|}-\frac{1}{2}$

$$=\frac{2|L_t|\times\eta_t-2|S|\times p-|L_t|+|S|}{2(|L_t|-|S|)}$$

$$=\frac{|L_t|\times(2\eta_t-1)+|S|\times(1-2p)}{2(|L_t|-|S|)}$$

因为 $\eta_t<1/2,p>1/2$.

所以　$\eta_{t+1}<1/2$

将 η_{t+1}带入式(3)得:

$$v_{t+1}=(|L_t|-|S|)\times(1-2\frac{|L_t|\times\eta_t-|S|\times p}{|L_t|-|S|})^2 \tag{4}$$

$$=\frac{[|L_t|(1-2\eta_t)+|S|(2p-1)]^2}{(|L_t|-|S|)}$$

设 $u=|L_t|(1-2\eta_t)+|S|(2p-1)$,对 v_{t+1}中的 $|S|$求导得:

$$v'_{t+1}=\frac{2u(2p-1)(|L_t|-|S|)+u^2}{(|L_t|-|S|)^2}$$

若 $\eta_t<1/2,p>1/2$,则 $v'_{t+1}>0$,因此 v_{t+1}随变量 $|S|$的增加而增加,当$|S|=0$时,v_{t+1}取得最小值,即:

$$v_{t+1}\geqslant|L_t|\times(1-2\eta_t)^2=v_t$$

因此当$|S|\neq0$,数据集 L_t 的噪音率小于 1/2 且集合分类器的分类准确率大于 1/2 时,$v_{t+1}>v_t$,则 $\varepsilon_{t+1}<\varepsilon_t$,即删除不一致训练数据集可以有效降低学习器的最坏情况分类错误率.通过定理 1 可知,当训练数据集噪声率小于 1/2 时,利用交叉验证去除噪音剩余的数据集噪音率仍小于 1/2,从而保证了交叉验证可以循环使用,不断提高学习器的分类准确率.

2.2　有放回交叉验证容噪分类算法有效性分析

设 L 为原始数据集,η 是数据集 L 所含有的噪音率,S_t 为数据集 L 内类标与第 t 轮训练的集合分类器分类不一致的数据集合,p_t 是第 t 轮训练的集合分类器分类准确率.S_{t-1}为数据集 L 内类标与第 $t-1$ 轮训练的集合分类器分类不一致的数据集合,p_{t-1}是第 $t-1$ 轮训练的集合分类器分类准确率,则 $t+1$ 轮训练数据集含有的噪音率为:

$$\eta_{t+1}=\frac{|L|\times\eta-|S_t|\times p_t}{|L|-|S_t|}$$

将其带入式(3):

$$v_{t+1}=(|L|-|S_t|)\times(1-2\frac{|L|\times\eta-|S_t|\times p_t}{|L|-|S_t|})^2$$

同理:

$$v_t=(|L|-|S_{t-1}|)\times(1-2\frac{|L|\times\eta-|S_{t-1}|\times p_{t-1}}{|L|-|S_{t-1}|})^2$$

如果 $v_{t+1}>v_t$,则 $\varepsilon_{t+1}<\varepsilon_t$.当$|S_{t-1}|>|S_t|$时,$v_t$ 的右侧第一项小于 v_{t+1}第一项.同时当$|S_t|\times p_t>|S_{t-1}|\times p_{t-1}$时,$v_t$ 的右侧第二项小于 v_{t+1}第二项.因此当满足下面公式时,有放回交叉验证容噪分类算法可以有效降低学习器的最坏情况分类错误率.

$$\frac{p_t}{p_{t-1}}>\frac{|S_{t-1}|}{|S_t|}>1 \tag{5}$$

对于条件$|S_{t-1}|>|S_t|$,可以通过抽样方法保证,当$|S_{t-1}|<|S_t|$时,对 S_t 进行抽样,利用抽样子集代替 S_t.对于条件式(5),则需要在数据集 L 中选出部分不含噪声的数据组成纯净数据集,利用算法在纯净数据集中的准确率加以验证.显然,与无放回交叉验证容噪分类算法相比,该算法需要更加复杂的验证条件.

3 容噪数据流集合分类算法

通过式(4)可知,在无放回交叉验证容噪分类算法中,分类准确率 p 值越高,不一致数据集 S 越大,学习器的最坏情况分类错误率就越小.确定不一致数据集时有绝对不一致方法和相对不一致方法可供选择,研究表明,绝对不一致方法具有较高的 p 值,但其获得的不一致数据集较小;相对不一致方法获得的不一致数据集虽然较大,但其 p 值却很低[7].因此两种方法均无法同时获得较高的 p 值和较大的不一致数据集 S.

数据流是一种大量连续达到的、潜在无限输入的数据有序序列.由于现实应用环境的复杂性,使得数据流中的噪音问题更加突出.因此,本文结合交叉验证容噪分类理论,提出一种容噪数据流集合分类算法,即NTDSEC算法,在剔除噪音的过程中,有效解决 p 值与不一致数据集 S 之间的问题.

设含有类别噪声的数据流 $DS=\{(x_1,y_1),(x_2,y_2)\cdots,(x_n,y_n),\cdots\}$.当 DS 每流入 n 个数据,即利用该数据段训练一个个体分类器.集合分类器 E 由最近建立的 k 个个体分类器组成.NTDSEC 的算法描述见算法 1.

交叉验证容噪分类算法确定不一致数据集时仅利用了分类结果,并没有考虑数据集的分布特性.当某个数据与其最近的数据具有相同类标时,其是噪音的可能性相对较小,反之是噪音的可能性相对较大,因此通过判断与其最近的数据是否具有相同的类标作为条件,分别采用相对不一致方法或绝对不一致方法确定不一致数据集可以有效解决 p 值与 S 集之间的问题.

算法 1 NTDSEC 算法

输入:最近流入的 n 个数据组成的集合 D_n,现有的集合分类器 E,现有个体分类器个数 k,允许个体分类器最大个数 $k_{\max}$.

输出:新集合分类器 $\hat{E}$,新的个体分类器个数 $\hat{k}$

(1)利用集合分类器 E 对数据集合 D_n 分类.

(2)$S\leftarrow\phi$. //存放不一致数据集

(3)对于任意 $j,j\in\{1,2,\cdots,k\}$,设 $n_j=0$.

(4)数据集 D_n 中的任意一个数据(x_i,y_i),如果 $f_j(x_i)\neq y_i$,则 $n_j=n_j+1$.

//用 n_j 存储个体分类器$f_j(x)$的分类结果与数据类标不一致的数据个数.

(5)如果 $n_j>\frac{1}{2}|D_n|$,则从集合分类器 E 中删除个体分类器 f_j. //出现概念漂移时,去除旧分类器

(6)对于数据集 D_n 中的任意一个数据(x_i,y_i),与其距离最近的数据(x_j,y_j),如果 $y_i\neq y_j$,则利用相对不一致方法判断其是否与集合分类器分类结果不一致,否则用绝对不一致方法判断.

(7)当判断其为不一致时,$S\leftarrow(x_i,y_i)$

(8)利用数据集$\tilde{D}_n=|D_n-S|$训练个体分类器 f_n.

(9)$k=size(E)$,如果 $k=k_{\max}$,则删除第一个个体分类器,且 $k=k-1$. //判断现有分类器个数是否达到上限

(10)集合分类器 $\tilde{E}=E\cup f_n,\hat{k}=k+1$.

4 实验验证及分析

为了验证 NTDSEC 算法的有效性,在 hyper-plane 数据集上进行对比实验,该数据集被广泛应用在数据流分类算法验证领域[13,14].实验环境为 windows XP 操作系统,主频 3.0Ghz,主存 512M,算法实现工具 Matlab7.噪音采用成对注入方式(X,Y),即对于类别为 X 的任意实例,按概率 p_e 将其类标转换为 Y,其中 p_e 为注入的噪音比例.该方法能够有效的刻画真实数据集含有噪音的特点.

为了更好的分析 NTDSEC 算法的去噪性能,本文引入了去噪准确率、噪音剩余率指标.设 N 是数据集 D 内含有的噪音集合,S_n 为不一致数据集,去噪准确率 $p1$、噪音剩余率 $p2$ 的计算公式如下:

$$p1=\frac{|S_n\bigcap N|}{|S_n|},p2=\frac{|N|-|S_n\bigcap N|}{|D-S_n|}$$

$p1$ 值越高表明算法去除噪音的准确率越高,其误删除的非噪音数据就越少.$p2$ 值越低表示剩余的数据集噪音率越低.

hyper-plane 数据集实验参数设置如下:数据段大小 $D_n=300$,数据维数 $n=5$,个体分类器个数阈值 $k_{\max}=10$,为了模拟概念漂移情况,在第 30、60 和 90 个数据段更改决定数据类别的超平面斜率,个体分类器采用 SVM

分类器.

图 1 为未容噪处理 (No noise-tolerance)、绝对不一致方法(Consensus vote)、相对不一致方法(Majority vote)和 NTDSEC 方法四种算法在数据流各个数据段上的分类准确率对比图,其中数据流的噪音率为 0.2.从图 1 可以看出,NTDSEC 算法的分类准确率最高,No noise-tolerance 算法的分类准确率最低,其他两种方法居中.四种算法刚开始运行时分类准确率都相对较低,随着流入的数据段不断增加,四种算法的分类准确率不断提高,原因在于四种算法在训练之初所包含的个体分类器个数与流入的数据段个数成正比,当到达个体分类器个数阈值时,才淘汰最初建立的个体分类器,因此其集合分类器的分类准确率随个体分类器个数的增加而增加.当出现概念漂移时,四种算法的分类准确率都有明显的下降,因为旧概念下训练的集合分类器必然无法对新概念产生的数据进行有效分类,但分类准确率能快速恢复表明 4 种算法能够有效删除旧概念下训练的个体分类器,从而适应新的概念.

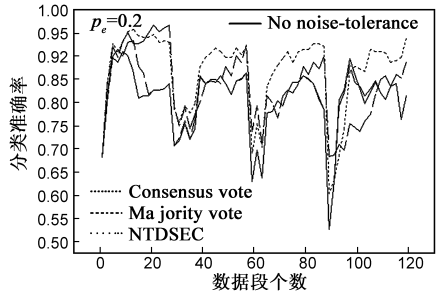

图1　四种算法在各个数据段上的分类准确率对比图

图 2 为四种算法在不同噪音率的数据流中其分类准确率的对比图,从图 2 可以看出,随着噪音率的不断增加,四种算法的分类准确率会明显下降.在噪音率为 0.1 的情况下,Consensus vote 算法比 Majority vote 算法分类准确率高,其他情况反之.从图 3、4 中分析其原因,在噪音率为 0.1 时,Consensus vote 算法的去噪准确率高出 Majority vote 算法许多,但噪音剩余率相差较小,分别是 0.0835 和 0.0274,因此 Consensus vote 算法的分类准确率相对较高,随着噪音率的不断提高,Consensus vote 算法虽然仍具有较高的去噪准确率,但其噪音剩余率也较大,在噪音率为 0.15、0.2、0.25 和 0.3 的数据流中,其噪音剩余率分别是 0.1346、0.1851、0.2386 和 0.2927,表明该算法去掉的噪音数据过少,因此其分类准确率更接近 No noise-tolerance 算法,Consensus vote 算法虽然去噪准确率较低,但噪音剩余率也非常低,表明该方法去掉了绝大多数的噪音数据,从而提高了分类准确率.

KDD-CUP99 数据集来源于标准数据库 UCI 中的网络入侵检测数据集,共 494021 条记录,,每条记录包括 7 个离散属性,34 个连续属性和 1 个类别属性,对应正常连接模式或某种入侵模式.对数据进行预处理,去掉离散属性,由于模式为 Normal 和 Dos 的记录占数据集的 98.93%,因此本文中仅保留上述两种模式的记录,将剩余的数据集进行混淆后(shuffling)再随机抽取[15,16],生成 50 个数据段,噪音注入方式同上实验.

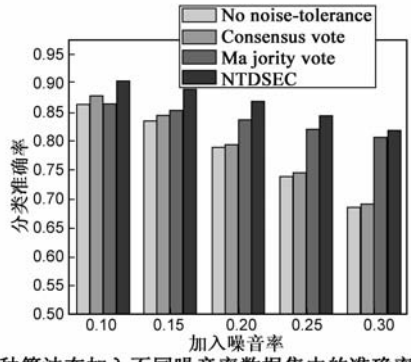

图2 四种算法在加入不同噪音率数据集中的准确率对比图

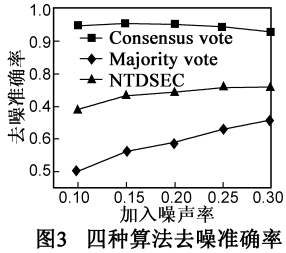

图3　四种算法去噪准确率对比图

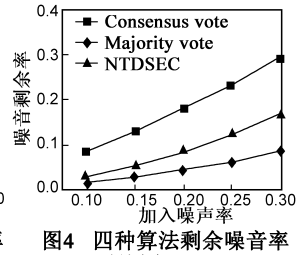

图4　四种算法剩余噪音率对比图

表 1 为数据段大小分别为 0.5k、1k 和 1.5k 三种情况下,四种算法在 KDD-CUP99 数据集上的准确率.从表 1 可以看出,在不同规模数据段上的实验,Consensus vote 算法和 Majority vote 算法的准确率都比 No noise-tolerance 算法有所提升,从而验证了交叉验证思想对噪音数据集容噪的有效性.四种算法中 NTDSEC 算法的准确率最高,因此在判断不一致数据集时,利用数据集的分布特性可以有效提高其判断的合理性.同时 NTDSEC 算法的准确率有进一步提升的空间,因为本文中仅考虑了最近数据的类标情况,当增大考虑范围,增加为最近几个数据的类标时,必然会使不一致数据集的判断更加合理.

表 1　四种算法在 KDD-CUP99 数据集上的准确率

	No noise-tolerance	Consensus vote	Majority vote	NTDSEC
0.5k	0.8136	0.8162	0.8303	0.8307
1k	0.8196	0.8283	0.8286	0.8319
1.5k	0.8169	0.8313	0.8332	0.8369

5　结论

利用含有噪音数据集训练分类器,分类器的分类准确率会受较大影响.本文针对交叉验证容噪分类算法没有完善的理论基础,对其进行了详细的理论证明

并给出了其适用条件.在此基础上,提出一种容噪数据流集合分类算法,使得不一致数据集的判断更加合理,在不同噪音比例及不同规模数据段上实验表明该算法明显优于传统的交叉验证容噪分类算法.

参考文献:

[1] X Q Zhu, X D Wu. Class noise vs attribute noise: a quantitative study of their impacts[J]. Artificial Intelligence Review, 2004, 11(3): 177 – 210.

[2] D Shen, Q Yang, Z Chen. Noise reduction through summarization for web-page classification[J]. Information Processing and Management, 2007, 43(6): 1735 – 174.

[3] V Eglin, S Bres, C Rivero. Hermite and Gabor transforms for noise reduction and handwriting classification in ancient manuscripts[J]. International Journal on Document Analysis and Recognition, 2007, 9(2): 101 – 122.

[4] X D Wu, X Q Zhu. Mining with noise knowledge: error aware data mining[J]. IEEE Transactions on Systems Man and Cybernetics, 2008, 38(4): 917 – 932.

[5] G H John. Robust decision trees: Removing outliers from databases[A]. Proceedings of the First International Conference on Knowledge Discovery and Data Mining[C]. Menlo Park, CA: AAAI Press, 1995. 174 – 179.

[6] D Gamgerber, N Lavrac, C Groselj. Experiments with noise filtering in a medical domain[A]. Proceedings of the Sixteenth International Conference on Machine Learning[C]. San Francisco, USA: Morgan Kaufmann, 1999. 143 – 151.

[7] X Q Zhu, X D Wu, Q J Chen. Eliminating class noise in large datasets[A]. Proceedings of the Twentieth International Conference on Machine Learning[C]. Washington DC, USA: AAAI Press, 2003. 920 – 927.

[8] V Sofie, V A Anneleen. Ensemble methods for noise elimination in classification problems[A]. Fourth Workshop on Multiple Classifier Systems[C]. United Kingdom: Springer-Verlag Press, 2003. 317 – 325.

[9] S Verbaeten. Identifying mislabeled training examples in ILP classification problems[A]. Proceedings of the Twelfth Dutch-Belgian Conference on Machine Learning[C]. Netherlands: Utrecht of University, 2002. 71 – 78.

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

[11] D Angluin, P Laird. Learning from noisy examples[J]. Machine Learning, 1988, 2(4): 343 – 370.

[12] Z H Zhou, M Li. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529 – 1541.

[13] W N Street, Y S Kim. A streaming ensemble algorithm for large-scale classification[A]. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York, USA: ACM Press, 2001. 377 – 382.

[14] F F Troyano, J S Aguilar Ruiz, J C Riquelme. Data streams classification by incremental rule learning with parameterized generalization[A]. Proceedings of the 2006 ACM Symposium on Applied Computing[C]. New York, USA: ACM Press, 2006. 657 – 661.

[15] P Zhang, X Q Zhu, Y Shi. Categorizing and mining concept drifting data streams[A]. Proceeding of the fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York, USA: ACM Press, 2008. 812 – 820.

[16] 欧阳震诤,罗建书,胡东敏等.一种不平衡数据流集成分类模型[J].电子学报,2010,38(1):184 – 189.
OuYang Z Z, Luo J S, HU D M, et al. An ensemble classifier framework for mining imbalanced data streams[J]. Acta Electronica Sinica, 2010, 38(1): 184 – 189. (in china)

作者简介:

张健沛 男,1956 年 11 月出生于黑龙江,哈尔滨工程大学教授、博士生导师,主要研究方向数据库与知识库.
E-mail: zhangjianpei@hrb.edu.cn

杨显飞 男,1979 年 7 月出生于黑龙江伊春,哈尔滨工程大学博士研究生,主要研究方向为机器学习和数据挖掘.
E-mail: yangxianfei@eyou.com

杨 静 女,1962 年 2 月出生于黑龙江,哈尔滨工程大学教授、博士生导师,主要研究方向数据库、软件工程.
E-mail: yangjing@hrb.edu.cn