基于间隔分布集成优化的齿轮箱故障诊断

胡清华[1] 朱鹏飞[1] 左 明[2]
(天津大学计算机科学与技术学院 天津 300072)[1]
(Department of Mechanical Engineering, University of Alberta, Edmonton, Alberta T6G 2G8)[2]

摘 要 齿轮裂纹等级的识别对于齿轮箱故障诊断具有重要意义。通过随机化邻域约简,生成一系列邻域可分子空间,从而形成不同的子分类器。通过最小化间隔损失或者求解L1正则最小平方损失问题来改变间隔分布,从而得到子分类器权值,对分类器按权值排序后,选择使得训练集分类精度最高的子分类器集合。实验结果表明,对于齿轮裂纹等级的识别,该方法的性能大大优于现有的其它方法。

关键词 邻域粗糙集,随机约简,集成学习,间隔分布
中图法分类号 TP181 **文献标识码** A

Gear Fault Diagnosis Based on Margin Distribution Ensemble Optimization

HU Qing-hua[1] ZHU Peng-fei[1] ZUO Ming[2]
(School of Computer Science and Technology, Tianjin University, Tianjin 300072, China)[1]
(Department of Mechanical Engineering, University of Alberta, Edmonton, Alberta T6G 2G8, Canada)[2]

Abstract Gear crack level identification is of great significance for gear box fault diagnosis. Regarding instability and performance limitation in current identification, we generated a set of neighborhood separable subspaces based on randomized attribute reduction, on which a set of base classifiers were obtained. The weight vector of the base classifiers was learned by optimizing loss of ensemble margin and regularization learning to change margin distribution. Base classifiers were ranked according to the weight value and a set of classifiers that make the ensemble classification accuracy highest on the training set was got. The experiment analysis shows that the proposed method is much better than other state of the art methods in crack level identification.

Keywords Neighborhood rough set, Randomized reducts, Ensemble learning, Margin distribution

1 引言

齿轮箱作为机械设备中最常用的动力传递部件,被广泛应用于飞机、汽车、电厂汽轮机以及钢铁厂等[12]。在长期运转的过程中,由于制造误差、冲击载荷和工作环境等因素的影响以及疲劳、老化等效应的存在,齿轮都将不可避免地出现一些故障[10]。对于大型复杂、自动化以及连续化程度很高的设备,其一旦出现故障,就会对整个生产造成很大的损失。因此,研究齿轮箱故障诊断对于降低设备的维修费用,提高产品的竞争力,防止突发性事故,具有很大的经济效益和社会效益[10]。

裂纹故障是齿轮最常出现的损坏形式之一。识别不同的裂纹等级,对于裂纹故障诊断来说是一个重要任务。雷等[12]提出一种基于加权K近邻的裂纹等级识别方法,其能够有效地识别齿轮裂纹等级。对于目前的裂纹等级识别方法存在的鲁棒性差以及识别率受限等问题,可通过集成学习予以解决。

在集成学习中,如何获得有差异的子分类器是影响集成学习效果的关键[3]。Bagging[4]和Boosting等[5]方法通过对训练样本进行处理来获得不同的基分类器。同时选择不同的特征子空间也可获得不同的子分类器[8]。Hu[6]应用集成粗糙子空间的方法,通过粗糙集生成有效的特征子空间,从而建立不同的分类器,取得了良好的效果。

集成学习的融合方式对于分类性能也有很大的影响[1]。简单投票[2]、线性加权投票[13]是两种最常见的融合方式。如何选择更优的子分类器进行融合以及分类器的权值学习成为一个重要的研究方向。Schapire等[15]从1范数间隔最大化角度解释了Boosting方法。Garg等[14]提出了基于间隔分布的复杂性度量,同时得到了基于间隔分布的泛化界,通过优化泛化界取得了良好的实验效果。

本文提出了一种基于随机化邻域属性约简和间隔分布的集成学习方法,并将其应用于齿轮裂纹等级识别中。首先通过随机化邻域约简,得到一系列分类性能较强的邻域可分子空间,在每个子空间中学习一个分类器,由此得到一系列的子分类器。通过对间隔的平方损失的直接优化或者1-范数正则优化学习分类器的权,进而优化融合后的分类间隔分布。然后根据权值对子分类器进行排序,选择使得训练集融合分类

到稿日期:2012-06-27 返修日期:2012-10-25
胡清华(1976—),男,教授,主要研究方向为机器学习、粗糙集理论;朱鹏飞(1985—),男,硕士生,主要研究方向为故障诊断和计算机视觉;
左 明 教授,主要研究方向为设备故障诊断。

精度最高的若干子分类器测试样本的集合。实验结果表明，对于齿轮裂纹等级的识别，提出的方法在识别率上大大优于其它方法。

2 背景知识

2.1 邻域粗糙集

鉴于 Pawlak 提出的粗糙集理论不能处理数值型数据，Hu 提出了基于邻域粒化的粗糙集模型[7]。

给定一个由 N 个属性描述的分类问题，可以将其形式化为一个决策系统$\langle U,A,D\rangle$。$U=\{x_1,\cdots,x_n\}$是全部样本构成的集合，$A=\{a_1,\cdots,a_N\}$是描述样本的属性集合，D 是分类决策属性。

定义 1 设$\langle U,\Delta\rangle$是非空度量空间，$x\in U,\delta\geqslant 0$，称点集 $N(x)=\{y|\Delta(x,y)\leqslant\delta,y\in U\}$为 x 的 δ 邻域。

定义 2 给定$\langle U,A,D\rangle$，如果 A 生成一族论域上的邻域关系，则 $NDT=\langle U,A,D\rangle$称为邻域决策系统。

定义 3 给定 $NDT=\langle U,A,D\rangle$，D 将 U 划分为 N 个等价类：$X_1,X_2,\cdots,X_N$，$B\subseteq A$ 生成 U 上的邻域关系 N_B，那么决策 D 关于 B 的邻域下近似和上近似分别为：

$$\underline{N_B}D=\{\underline{N_B}X_N,\underline{N_B}X_2,\cdots,\underline{N_B}X_N\}$$

$$\overline{N_B}D=\{\overline{N_B}X_N,\overline{N_B}X_2,\cdots,\overline{N_B}X_N\}$$

定义 4 给定 $NDT=\langle U,A,D\rangle$，决策属性 D 对条件属性 $B\subseteq A$ 的依赖度为：

$$\gamma_B(D)=\mathrm{Card}(\underline{N_B}D)/\mathrm{Card}(U)$$

给定邻域决策系统 $NDT=\langle U,A,D\rangle$，$B\subseteq A$，$a\subseteq B$，如果(1)$\gamma_B(D)=\gamma_A(D)$；(2)$\forall a\in B:\gamma_{(B-a)}(D)<\gamma_B(D)$，则称 B 是一个属性约简。在邻域粗糙集的框架下，B 又可称为邻域可分子空间(NSS)。

定义 5 给定 $NDT=\langle U,A,D\rangle$，$\{B_j\,|\,j\leqslant r\rangle$是一组约简的集合，称 $Core=\bigcap\limits_{j\leqslant r}B_j$ 为核。

本质上，约简是一组保持原始数据近似能力的特征子集，粗糙集理论认为存在多个可以保持原始数据近似能力的属性子集。在不同的子空间中信息不同，而不同的子空间之间互相补充，因此通过集成多个不同子空间中的信息可以提高泛化能力[6]。

2.2 集成间隔

给定一组样本 $X=\{(x_i,y_i)\}$，$i=1,2,\cdots,n$，$y_i\in\{+1,-1\}$。在 m 个不同的邻域约简上的分类输出是 $H\in\mathcal{R}^{n\times m}$，$W=\langle w_1,w_2,\cdots,w_m\rangle$是 m 个分类器对应的权值。

定义 6 对于任意样本 $x_i\in X$，在 m 个邻域可分子空间上的预测值为$\{h_{ij}\}$，$j=1,2,\cdots,m$。根据加权投票的融合方式，得到新的分类器 $H=\mathrm{sgn}(\sum\limits_{j=1}^{m}w_jh_{ij})$，样本 x_i 在 H 上的分类间隔定义为：

$$\rho(x_i)=y_iH(x_i) \tag{1}$$

显然，如果 $\rho(x_i)>0$，则样本被正确分类，否则 x_i 被错分。

定义 7 对于多类分类问题，给定样本 $x_i\in X$ 在 m 个不同的邻域可分子空间上的分类输出为$\{h_{ij}\}$，$j=1,2,\cdots,m$，定义如下矩阵 $D=\{d_{ij}\}_{n\times m}$：

$$d_{ij}=f(y_i,h_{ij})=\begin{cases}+1, & \text{if } y_i=h_{ij}\\ -1, & \text{if } y_i\neq h_{ij}\end{cases}$$

$d_{ij}=+1$ 表示样本 x_i 在第 j 个分类器上分类正确，否则就被错误分类。这种定义也适用于两类分类问题。

定义 8 对于任意样本 $x_i\in X$ 在 m 个不同的邻域可分子空间上的分类输出为$\{h_{ij}\}$，$j=1,2,\cdots,m$，则对于样本 x_i，融合间隔 EM(Ensemble Margin)可定义如下：

$$\rho(x_i)=\sum_{j=1}^{m}w_jd_{ij},\ \sum_{j=1}^{m}w_j=1 \tag{2}$$

融合间隔的定义同时适用于两类和多类问题，反映了样本在加权融合过程中被正确分类的程度。如果样本 x_i 被所有基本分类器正确分类，则间隔为 1；如果样本 x_i 被所有样本基本分类器错分，则间隔为 -1。

如果 $\rho(x_i)\in(0,1]$，则 x_i 被正确分类；

如果 $\rho(x_i)\in[-1,0)$，则 x_i 被错误分类；

如果 $\rho(x_i)=0$，则 x_i 可能分对也可能分错。

定义 9 给定一组样本 X，在 m 个不同的邻域可分子空间上的分类输出是 $H\in\mathcal{R}^{n\times m}$，则样本集 X 平均间隔 AEM 可定义如下：

$$E=\frac{1}{n}\sum_{i=1}^{n}\rho(x_i)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}w_jd_{ij}$$

定义 10 对于任意样本 $x_i\in X$ 在 m 个不同的邻域可分子空间上的分类输出为$\{h_{ij}\}$，$j=1,2,\cdots,m$，对于样本 x_i，融合间隔 EM 为 $\rho(x_i)$，则样本 x_i 在 H 中的损失定义为：

$$l(x_i)=[1-\rho(x_i)]^2 \tag{3}$$

对于样本集 X，融合损失为：

$$\begin{aligned}l(X)&=\sum_{i=1}^{n}l(x_i)=\sum_{i=1}^{n}[1-\rho(x_i)]^2\\&=\sum_{i=1}^{n}[1-\sum_{j=1}^{m}w_jd_{ij}]^2=\|U-DW\|_2^2\end{aligned} \tag{4}$$

式中，U 为长度为 m、元素为 1 的向量。

3 基于间隔分布的集成学习

3.1 基本思想

本文提出的集成学习方法的结构如图 1 所示。给定一个决策表$\langle U,A,D\rangle$，条件属性集为$\{a_1,a_2,\cdots,a_n\}$，根据随机化邻域属性约简算法可以得到一组属性约简集合$\{AR_1,AR_2,\cdots,AR_m\}$。给定一个新的样本和分类器，在各个邻域可分子空间上可得到相应的输出。通过最小化融合损失可以得到分类器的权值，从而对分类器进行排序，最终选择部分分类器的输出，以简单投票或者加权投票的方式进行融合，以得到这个样本的类标号。

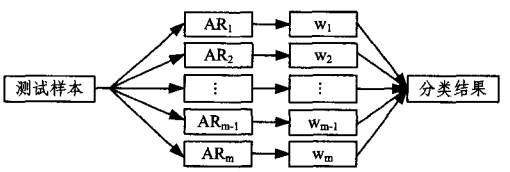

图 1 基于属性约简的集成学习

3.2 邻域随机约简

可采用贪心搜索策略实现基于邻域粗糙集的属性约简。从空集开始，每一步增加一个使得区分能力增长最大的特征，

直到增加任何属性，且区分能力都不再增长为止。通过这一步计算，将生成一个嵌套的特征子集序列：$B_1 \subset B_2 \subset \cdots \subset B_k$。这一步采用属性重要度来计算特征的区分能力。属性的重要度计算为：

$$SIG(a,B,D)=\gamma_{B\cup a}(D)-\gamma_B(D)$$

由于每次都是选择区分能力最大的特征，这样就只能得到一个约简，称其为在这种搜索方法下的最优约简。为了得到多个约简，可以放宽每一步中对于选中属性的要求，可随机选择区分能力最大的前 K 个特征中的一个作为选中属性，这样就可以得到多个具有区分能力的约简。

算法 1 基于邻域粗糙集的随机属性约简

输入：⟨U,A,D⟩，参数 delta 和随机数 N

输出：约简 red

(1) $\emptyset \rightarrow$ red;

(2) For each $a_i \in A-$red

(3) 计算 $\gamma_{red\cup a_i}(D)=\dfrac{|POS_{B\cup a_i}(D)|}{|U|}$

(4) 计算 $SIG(a_i,red,D)=\gamma_{red\cup a}(D)-\gamma_{red}(D)$

(5) end

(6) 选择 a_k，a_k 为属性集{A−red}中属性重要度 $SIG(a_i,red,D)$ 前 N 个最大中的一个

(7) If $SIG(a_k,red,D)>0$

(8) $red\cup a_k \rightarrow$ red

(9) go to (2)

(10) else

(11) 返回 red

(12) end if

选定随机数 N 后，每运行一次程序即可得到一个随机化的属性约简。这个算法的复杂度为 $(2n-k)(k+1)\times(k+1)\times m\log m/2$，其中 n 和 m 分别为样本和特征的数目，k 为约简中属性的个数。

在邻域粗糙集模型中，属性约简为保持原始空间近似能力的特征子集。最优约简具有更强的近似能力，但并不一定具有更好的泛化性能，即属性约简的选取并不依赖于分类性能，因此在相应的特征子空间上，最优约简对应的分类性能不一定最优。

3.3 分类器融合

在分类器融合过程中，间隔应越大越好，融合损失应越小越好。可以改变分类器的权值使得融合损失最小化，同时得到分类器的权值后，可对分类器进行排序，选择部分更优的分类器进行融合。

首先以融合损失为优化目标，如式(5)所示。

$$\begin{aligned} &J_W=\arg\min_W \|U-DW\|_2^2 \\ &\text{s.t.} \sum_{j=1}^{m} w_j=1 \end{aligned} \tag{5}$$

式(5)为一个具有线性约束的最小二乘问题，为一个凸优化问题。在求解的时候根据对 w_j 的非负要求可以分为两种情况：最小化损失问题和带有非负约束的最小化损失问题。

同时我们可以引入 1-范数正则项，在最小化融合损失的同时可以保证分类器权值的稀疏性，如式(6)所示。

$$\begin{aligned} &J_W=\arg\min_W\{\|U-DW\|^2+\lambda\|W\|_1\} \\ &\text{s.t.} \sum_{j=1}^{m} w_j=1 \end{aligned} \tag{6}$$

式中，λ 为平衡融合损失项和正则项的参数，λ 越大，稀疏性越强。

对于 $\sum_{j=1}^{m} w_j=1$，可将其写成 $1=LW$，其中 $L=[1;1;\cdots,1]$ 为长度为 m 的列向量，代入到 $\|U-DW\|_2^2$，则：

$$\begin{aligned} \|U-DW\|_2^2 &= \|U-DW+1-LW\|_2^2 \\ &= \|[U,1]-[D,L]W\|_2^2 \end{aligned}$$

令 $\tilde{U}=[U,1]$，$\tilde{D}=[D,L]$，则式(6)变为：

$$J_W=\arg\min_W\{\|\tilde{U}-\tilde{D}W\|_2^2+\lambda\|W\|_1\} \tag{7}$$

式(7)是一个 L1 正则最小平方问题，本文中利用 l1_ls 工具箱进行求解[11]。在求解的时候根据对 w_j 的非负要求可以分为两种情况：L1 正则问题(L1)和 NL1 正则(权值大于等于零)。我们将在实验部分予以讨论。

给定样本集合 $X=X_t+X_s$，其中 X_t 为训练集，X_s 为测试集。在训练集 X_t 上，通过优化目标可以得到各个分类器的权值 W。然后根据分类器权值的大小，可以对分类器进行排序，权值越大，则分类器对于融合越重要。融合的方式有两种，包括简单投票(所有分类器的权值都相同)以及加权投票的方法。在训练集上，每次添加一个基本分类器，选择使得分类精度最高的前 P 个基本分类器。

4 故障数据获取

数据采集的实验系统由齿轮箱、驱动齿轮旋转的三相交流电机以及加装载荷的磁力制动器组成，如图 2 所示。电机的旋转速度由一个速度控制器控制，这样齿轮箱可在不同的转速下运动。齿轮箱总共有 3 个轴 4 个齿轮，并被一个同步齿型带驱动。齿轮的振动由齿轮上的两个加速度传感器测得。两个传感器分别安装在齿轮箱的水平和垂直方向上。一个数字信号处理器以及带有数据获取软件的电脑用来收集振动数据，以做进一步处理。每次实验中，保持其它 3 个齿轮不变，改变另外一个齿轮的裂纹等级。

表 1 特征提取的各项指标

F	描述	F	描述	F	描述	F	描述
1	x 方向 FM0	14	时域 x 方向均值	27	时域 y 方向波形指标	40	频域 x 方向峰值因子
2	y 方向 FM0	15	时域 x 方向标准方差	28	时域 y 方向偏斜度	41	频域 x 方向峰值脉冲指标
3	x 方向 FM4	16	时域 x 方向均方根	29	时域 y 方向峭度指标	42	频域 x 方向中心
4	y 方向 FM4	17	时域 x 方向波形指标	30	时域 y 方向峰值因子	43	频域 y 方向峰值
5	x 方向 FM4*	18	时域 x 方向偏斜度	31	时域 y 方向峰值脉冲指标	44	频域 y 方向均值
6	y 方向 FM4*	19	时域 x 方向峭度指标	32	时域 y 方向中心	45	频域 y 方向标准方差
7	x 方向 FA6	20	时域 x 方向峰值因子	33	频域 x 方向峰值	46	频域 y 方向均方根
8	y 方向 FA6	21	时域 x 方向峰值脉冲指标	34	频域 x 方向均值	47	频域 y 方向波形指标
9	x 方向 F6A*	22	时域 x 方向中心	35	频域 x 方向标准方差	48	频域 y 方向偏斜度
10	y 方向 F6A*	23	时域 y 方向峰值	36	频域 x 方向均方根	49	频域 y 方向峭度指标
11	x 方向 EOP	24	时域 y 方向均值	37	频域 x 方向波形指标	50	频域 y 方向峰值因子
12	y 方向 EOP	25	时域 y 方向标准方差	38	频域 x 方向偏斜度	51	频域 y 方向峰值脉冲指标
13	时域 x 方向峰值	26	时域 y 方向均方根	39	频域 x 方向峭度指标	52	频域 y 方向中心

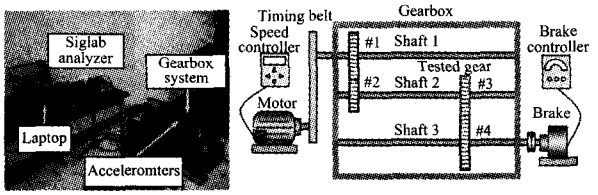

图 2 实验系统[12]

本文分别采用了 3 个负载等级、12 个不同转速和 5 个裂纹等级进行实验，并且在相同的负载、转速以及裂纹等级下重复实验 3 次，这样共获得 540 个样本。接下来对数据进行特征提取，分别对 x 方向(水平)以及 y 方向(垂直)进行时域和频域分析，之后分别计算了各个方向的均值、标准方差以及均方根等 52 个指标，如表 1 所列。经过特征提取后，得到了具有 540 个样本、52 个条件属性、1 个决策属性(5 个裂纹等级)的学习样本。

5 识别实验

为了展示裂纹等级识别问题的复杂性，利用邻域属性约简的方法选取最优约简，并选取了约简中的前 3 个特征 26：(时域 y 方向均方根)、52(频域 y 方向中心)、33(频域 x 方向峰值)，画出数据特征空间分布图，如图 3 所示。可以看出，各类数据交叠在一起，给分类带来一定的难度。

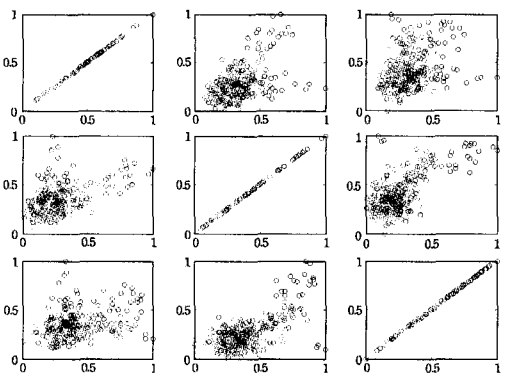

图 3 特征空间分布图

为了验证基于邻域可分子空间与间隔分布的集成学习方法(NSS_MD)的有效性，分别选择了两种基分类器，即最近邻分类器以及 RBF-SVM 分类器。分别计算了在原始数据上以及在最优约简上的分类精度，同时也计算了 Bagging 以及 Adaboost 两种集成学习方法的分类精度，如表 2 所列。表 3 和表 4 给出了 NSS_MD 集成学习方法的分类精度，其中 L 和 L1 分别代表最小化间隔损失以及正则化学习，V 和 W 分别代表简单投票以及线性加权投票。从两个表可以看出，NSS_MD 的识别率大大高于 Bagging、Adaboost 等相关方法。

表 2 识别精度比较

Method	Raw	Reduct	Bagging	Adaboost
1-NN	71.8±9.9	66.9±10.0	74.1±10.3	72.2±9.4
RBFSVM	77.5±8.8	74.1±10.2	78.3±9.2	76.7±9.7

表 3 NSS_MD 分类精度及分类器数目

	LV		L1V		LW		L1W	
	Acc.	No.	Acc.	No.	Acc.	No.	Acc.	No.
NN	88.4±9.8	21.9	88.7±8.6	24.7	82.5±8.8	82	89.3±12.4	9.1
SVM	84.4±6.2	35.2	85.8±7.0	21.9	84.4±6.1	51.3	84.7±9.2	9.9

表 4 NSS_MD 分类精度及分类器数目(权值非负)

	LV		L1V		LW		L1W	
	Acc.	No.	Acc.	No.	Acc.	No.	Acc.	No.
NN	84.5±9.5	15.2	86.9±9.1	23.1	87.6±9.9	16.1	87.8±11.0	26.9
SVM	85.1±8.4	13.3	85.6±7.3	20.4	84.7±9.2	11	85.5±7.7	12.8

对于分类器数目的选取方式，我们选择使得训练集分类精度最高的前 P 个分类器。图 4 和图 5 分别展示了在两种分类器上，最小化融合损失以及 L1 正则化学习随着分类器数目的增加，分类精度的变化。从图 6 可以看出，随着分类器数目的增加，分类精度先上升后下降，以很少的分类器便可使训练集分类精度达到最高。

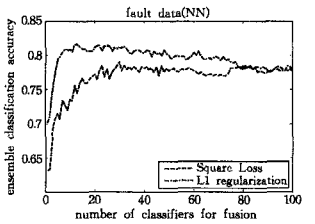

图 4 随着分类器数目增加分类精度变化

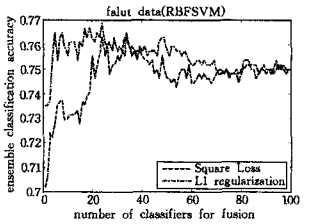

图 5 随着分类器数目增加分类精度的变化

同时我们展示了最小化融合损失以及 L1 正则化学习得到的分类器权值，如图 6 所示。从图中可以看出，L1 正则化给分类器权值带来了稀疏性。

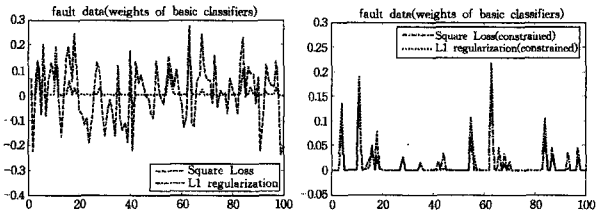

图 6 学习到的分类器的权值对比

图 7 和图 8 给出了集成学习系统在不同优化策略下的间隔累积分布。从图中可以看出，经过间隔分布优化后的集成学习的间隔分布曲线处于不进行优化的曲线的下面，这表明优化后的间隔分布能够产生更好的推广性能。

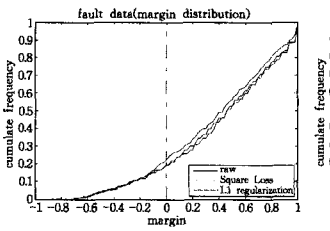

图 7 间隔分布对比

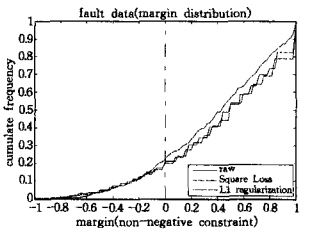

图 8 具有非负约束的间隔分布对比

结束语 齿轮裂纹等级识别在齿轮箱故障诊断中起着非常重要的作用。本文通过随机化邻域属性约简得到一系列邻域可分子空间，在各个属性约简上可得到一系列基本分类器。通过最小化融合间隔的损失或求解 L1 正则最小平方损失问题来改变间隔分布，可得到各基本分类器的权值。按权值对子分类器排序，然后选择使得集成分类精度最高的子分类器集合。实验结果表明，这种方法能够有效地进行齿轮裂纹等级识别，识别率大大优于其它方法。

参 考 文 献

[1] Zhang L, Zhou W-D. Sparse ensembles using weighted combina-

tion methods based on linear programming[J]. Pattern Recognition,2011,44(1):97-106

[2] Kittler J, Hatef M, Duin R P W, et al. On combining classifiers [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1998,20(3):226-239

[3] Zhou Z-H, Yu Y. Ensembling local learners through multimodal perturbation [J]. IEEE Trans. SMC—Part B: Cybernetics, 2005,35:725-735

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

[5] Schapire R E. The strength of weak Learnability[J]. Machine Learning,1990,5(2):197-227

[6] Hu Q-H, Yu D-R, Xie Z-X, et al. EROS: ensemble rough subspaces[J]. Pattern Recognition,2007,40:3728-3739

[7] Hu Q-H, Yu D-R, Xie Z-X. Neighborhood classifiers[J]. Expert Systems with Applications,2008,34:866-876

[8] Valentini G, Masulli F. Ensembles of learning machines[C]// Proc of Valentini02. Neural Nets WIRN Vietri-02, Series Lecture Notes in Computer Sciences. Springer-Verlag,2002:3-19

[9] Rosset S, Zhu J, Hastie T. Boosting as a Regularized Path to a Maximum Margin Classifier[J]. Journal of Machine Learning Research,2004,5:941-973

[10] 万小毛,鲍明,赵淳生. 齿轮箱故障诊断技术综述[J]. 振动、测试与诊断,1990,10(4)

[11] Kim S-J, Koh K, Lustig M, et al. A method for large-scale l1-regularized least squares[J]. IEEE Journal on Selected Topics in Signal Processing,2007,1(4):606-617

[12] Lei Y G, Zuo M J. Gear crack level identification based on weighted k neighbor classification algorithm [J]. Mechanical Systems and Signal Processing,2009,23(5):1535-1547

[13] Fumera G, Roli F. A theoretical and experimental analysis of linear combiners for multiple classifier systems[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27 (6):942-956

[14] Garg A, Roth D. Margin distribution and learning[C]// Proc. Int. Conf. Mach. Learn. ,Washington,DC,2003:210-217

[15] Schapire R E, Bartlett P, Freund Y, et al. Boosting the margin: A new explanation for the effectiveness of voting methods[J]. Annals of Statistics,1998,26(5):1651-1686

(上接第 180 页)

因此,能够获取一致性查询结果的数据库范围集为$\bigcap_{i=1}^{n} r_i$。

例 4 (例 2 续)考虑完整性约束:$IC=\{\forall(x,y,z)(Supply(x,y,z)\wedge Product(z,T_4)\rightarrow x=C)\}$和非一致数据库实例$r=\{Supply(C,D_1,It_1),Supply(D,D_2,It_2),Product(It_1,T_4),Product(It_2,T_4)\}$。

数据库实例 r 仅有两个修复实例分别是 $r'=\{Supply(C,D_1,It_1),Product(It_1,T_4),Product(It_2,T_4)\}$ 和 $r''=\{Supply(C,D_1,It_1),Supply(D,D_2,It_2),Product(It_1,T_4)\}$。第一种查询"$Supply(x,y,z)$?",元组$(C,D_1,It_1)$是一致性结果,因为它能够从两个修复实例中被获取,但是(C,D_2,It_2)不能从 r'中获取,所以它不是一致性结果,则查询结果 $Q(\bar{t})\in(\bigcap_{i=1}^{n} r_i=\{Supply(C,D_1,It_1),Product(It_1,T_4)\})$。第二种查询"$Supply(C,D_1,It_1)$?"结果为真,而"$Supply(D,D_2,It_2)$?"为假;若查询为真,则 $Q(\bar{t})\in r_i(1\leqslant i\leqslant n)$,即 $Q(\bar{t})\in(\bigcap_{i=1}^{n} r_i=\{Supply(C,D_1,It_1),Product(It_1,T_4)\})$。一致性查询结果必须都能从它的每一个修复实例中获取到。

结束语 数据库完整性约束 IC 作为数据库模式的一部分,有效地保证了数据的完整性和有效性,使数据符合现实世界的实体规则。现有的商业 DBMS 为 IC 提供了支持,但总聚焦在发展一系列的约束关系来尽可能保证每一个数据库是合法和一致的。然而,现实世界的一个实体在数据库中常常是对应多个不一致的数据。本文利用 $TP(IC\cup r)$分支封闭和开放规则,将非一致性数据库与 tableau 推理相结合,根据结点封闭值定义来计算确定分支封闭值,然后通过开放极小分支封闭值的分支来实现非一致性数据库的修复,并将这种方法扩展到带有 I 封闭的非一致性数据库。这样可以避免通过清除方法实现修复从而导致数据信息丢失的问题。这种方法对于实现数据库修复和保证数据库的完整性具有极其重要的意义。未来,我们考虑将该方法与数据库一致性查询相结合,来提高非一致性数据库的一致数据重写查询性能。

参 考 文 献

[1] Bertossi L. Database repairing and consistent query answering [M]. Morgan & Claypool publishers,2011

[2] Arenas M, Bertossi L, Chomicki J. Consistent query answers in inconsistent databases[C]// ACM Symposium on Principles of Database Systems(ACM PODS'99). ACM Press,1999:68-79

[3] Bertossi L, Schwind C. An analytic tableaux based characterization of database repairs for consistent query answering(preliminary report)[C]// Working Notes of the IJCAI'01 Workshop on Inconsistency in Data and Knowledge. AAAI Press,2001:96-106

[4] 李娇,刘全,傅启明,等. 分布式数据库中基于局部 CON 模型的记录匹配方法[J]. 通信学报,2011,32(7):196-202

[5] Fan W, Geerts F, Ma S, et al. Detecting inconsistencies in distributed data[C]// the IEEE International Conference on Data Engineering (ICDE). 2010:64-75

[6] Fan W, Jia X, Li J, et al. Reasoning about record matching rules [J]. the International Conference on Very large Data Bases (VLDB), VLDB Endowment, 2009, 2(1):407-418

[7] Bertossi L. Consistent query answering in databases[J]. ACM SIGMOD Record,2006,35(2):68-79

[8] Chomicki J, Marcinkowski J, Staworko S. Computing consistent query answers using conflict hypergraphs[C]// ACM International Conference on Information and Knowledge Management. Washington DC: ACM Press,2004:417-426

[9] Elmagarmid A K, Ipeirotis P G, Verykois V S. Duplicate record detection: a survey[J]. IEEE Transaction on Data and Knowledge Engineering (TKDE),2007,19(1):1-16

[10] 刘全,伏玉琛,孙吉贵,等. 一种基于符号集合的自动推理扩展方法[J]. 计算机研究与发展,2007,44(8):1317-1323

[11] 刘全,孙吉贵,崔志明. 基于布尔剪枝的多值广义量词 Tableau 推理规则简化方法[J]. 计算机学报,2005,28(9):1514-1518

[12] Melvin F. First-order logic and automated theorem proving[M]. New York: Springer Verlag,1996

[13] Golab L, Karloff H, Korn F, et al. On generating near optional tableaux for conditional functional dependencies[J]. the International Conference on Very large Data Bases (VLDB), VLDB Endowment,2008,1(1):376-390