基于结点优化的决策导向无环图支持向量机及其在故障诊断中的应用

易 辉 [1] 宋晓峰 [1] 姜 斌 [1] 王定成 [1,2]

摘 要 支持向量机 (Support vector machine, SVM) 是利用离在线数据自动建立故障诊断模型的智能方法, 它在多故障诊断时, 必须先进行多分类扩展. 决策导向无环图 (Decision directed acyclic graph, DDAG) 法是一种性能优秀的多分类扩展策略, 但该方法的决策结果与结点的排部密切相关, 而其结点的排部却是主观的, 影响了诊断的正确率. 本文提出一种根据故障数据的空间分布来优化结点排部的方法, 它能够提高支持向量机诊断的正确率. 采用该方法扩展的多分类支持向量机在变压器故障诊断中获得良好效果.

关键词 支持向量机, 故障诊断, 多分类, 决策导向无环图, 结点优化

DOI 10.3724/SP.J.1004.2010.00427

Support Vector Machine Based on Nodes Refined Decision Directed Acyclic Graph and Its Application to Fault Diagnosis

YI Hui[1] SONG Xiao-Feng[1] JIANG Bin[1] WANG Ding-Cheng[1,2]

Abstract Support vector machine (SVM) is an intelligent method which can create diagnostic models automatically by using off/on-line data sets, but it needs to be extended to a multi-class classifier for multi-fault diagnosis. Decision directed acyclic graph (DDAG) is an extending strategy with outstanding performance. However, its decision largely depends on the sequence of nodes which is arbitrarily selected. This affects the accuracy of diagnosis. In this paper, we proposed a method to refine the sequence of nodes according to the distribution of fault data sets, so as to improve the accuracy of SVM-based diagnosis. Multi-class SVM extended by our method has been employed as a transformer fault diagnosis, and satisfactory results have been obtained.

Key words Support vector machine (SVM), fault diagnosis, multi-class, decision directed acyclic graph (DDAG), node-refined

故障诊断技术对现代生产具有重大的意义. 目前故障诊断技术主要可分为三类: 基于解析模型的方法、基于信号的方法和基于知识的方法[1]. 由于现代工业系统变得越来越复杂, 基于解析模型的方法 (如状态估计法) 在实际运用中需要构建更复杂的精确数学模型[2], 不易操作; 基于信号处理的方法 (如小波分析法) 和一些基于知识的方法 (如神经网络法) 虽然避免了建立精确数学模型的麻烦, 但其理论基础是传统统计学, 按照大数定律, 只有训练样本接近无穷大, 其统计规律才能被精确地表达.

然而在生产中, 如果故障训练样本过多, 会导致诊断耗时较大, 不能对数据进行在线诊断. 反之, 如果训练样本过少, 那么生成的故障诊断模型性能得不到保证.

支持向量机 (Support vector machine, SVM) 是根据统计学习理论 (Statistical learning theory, SLT) 提出的一种基于知识的智能分类算法. 该方法能够在只有少量离在线故障数据的情况下[3], 自动建立优秀的故障分类模型, 是目前故障诊断的一个热点研究方向.

1 支持向量机及多分类扩展策略研究

支持向量机方法采用核函数将低维不可分数据投影至高维空间, 形成一个线性可分的数据集, 并通过构建最大间隔分类超平面将数据进行分类. 该方法通过求解一个线性约束的二次规划问题得到全局的最优解, 能够在少量样本的情况下确保得到较好的分类结果.

收稿日期 2008-11-11 录用日期 2009-04-01
Manuscript received November 11, 2008; accepted April 1, 2009
国家自然科学基金 (60874051), 中国博士后科学基金 (20070411044), 江苏省自然科学基金 (BK2007195), 江苏省博士后科研资助计划 (0701014B) 资助
Supported by National Natural Science Foundation of China (60874051), China Postdoctoral Science Foundation (20070411044), Natural Science Foundation of Jiangsu Province (BK2007195), and Jiangsu Planned Projects for Postdoctoral Research Funds (0701014B)
1. 南京航空航天大学自动化学院 南京 210016 2. 南京信息工程大学计算机与软件学院 南京 210044
1. College of Automation, Nanjing University of Aeronautics and Astronautics, Nanjing 210016 2. Institute of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044

但是, SVM 是针对二分类问题而设计的, 当类别数 $k > 2$ 时, SVM 无法直接分类. 故障诊断在实际运用中通常面临多分类问题, 将 SVM 运用到故障诊断上必须首先将二分类的 SVM 进行多分类扩展.

针对 SVM 多分类扩展问题, 经典方法是通过建立多个二分类器, 两两组合进行决策, 达到多分类的目的. 根据不同的分类器构建策略, 可分为 1-a-r (1-against-rest) 法[4] 和 1-a-1 (1-against-1) 法[5]. 此类方法易于操作, 但运算量大, 存在划分盲区, 且可能因为正负样本量的不对称, 导致过拟合; 2002 年 Takahashi 等[6] 将决策树方法引入多分类扩展问题, 该方法采用树状结构对决策流程进行控制, 有效地削减了计算量, 并避免了划分盲区. 但是该方法在训练时仍然存在正负样本量不对称问题, 而且在决策阶段, 该方法对于不同的测试数据采用同一种决策路径 (Evaluation path), 这将严重影响决策结果的可靠性.

1.1 决策导向无环图法

在这些传统方法存在大量缺陷的情况下, 决策导向无环图 (Decision directed acyclic graph, DDAG) 引起了广泛的关注. 受图论中的有向无环图 (Directed acyclic graph, DAG) 思想启发, Platt 等[7] 提出了这种能解决样本不对称、无盲区并优化了训练和决策时间的多分类扩展策略. 对于 k 类问题, 该方法共有 $k(k-1)/2$ 个结点分布于 k 层结构中, 其顶层只含 1 个结点, 称之为根结点, 第 2 层含有 2 个结点, 依次, 第 i 层含有 i 个结点. 这些结点中, 第 i 层的第 j 个结点指向第 $i+1$ 层中的第 j 个和 $j+1$ 个结点. 采用 DDAG 法对一组 4 类数据进行决策的流程可由图 1 表示 (图中 Not i 为样本划分非 i 类).

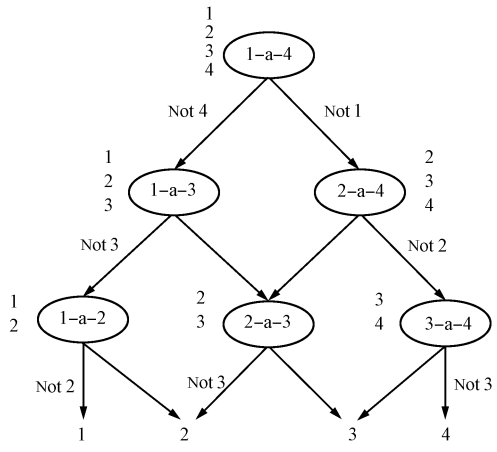


图 1 DDAG 四分类决策流程

Fig. 1 Procedures for a four-type DDAG decision

这种决策方法在不增加决策计算量的情况下, 为不同数据选取不同的决策路径, 提高了划分精度. 该方法给 SVM 多分类扩展提供了一种很有前景的解决方案.

但是, DDAG 结构不是唯一的, 不同 DDAG 结构对应不同的结点排部方式, 而划分结果却与结点的排部密切相关[8]. 一般通过重复实验的方法可选取一个较好的结点排部, 但重复实验会导致运算量加大, 且不能给予结点排部理论上的指导. 针对该问题, 本文提出了基于结点优化的 DDAG 多分类扩展策略.

2 基于结点优化的 DDAG 多分类扩展策略

定理 1. 设有 k 层结构的 DDAG 图, 根结点的划分风险概率为 ε_1, 第 i 层结点处的划分风险概率为 ε_i (假设同层结点具有相近的划分风险概率). 测试数据对应的真实类别在第 m 层结点处首次出现. $E = \{\varepsilon_1, \cdots, \varepsilon_m, \cdots, \varepsilon_{k-1}\}$. 那么当决策系统的风险概率最小时, 必有:

$$\varepsilon_1 \geq \varepsilon_2 \geq \cdots \geq \varepsilon_{a-1} \geq \max(\varepsilon_a, \varepsilon_{a+1}, \geq \varepsilon_{k-1}),$$
$$a = \lfloor \sqrt{2(k-1)+0.25} - 0.5 \rfloor \quad (1)$$

证明. 设决策系统的风险概率最小时, 各层结点处风险概率 $E = \{\varepsilon_1, \cdots, \varepsilon_m, \cdots, \varepsilon_{k-1}\}$, $m \in \{1, 2, \cdots, k-1\}$, 则决策系统的风险概率为

$$p = 1 - (1-\varepsilon_m)(1-\varepsilon_{m+1}) \times \cdots \times (1-\varepsilon_{k-1}) \quad (2)$$

若 $\exists i \in [1, m-1], j \in [m, k-1]$, 使得 $\varepsilon_i < \varepsilon_j$, 则必然有:

$$p = 1 - (1-\varepsilon_m)(1-\varepsilon_{m+1}) \times \cdots \times (1-\varepsilon_j) \times \cdots \times (1-\varepsilon_{k-1}) \quad (3)$$

$$p' = 1 - (1-\varepsilon_m)(1-\varepsilon_{m+1}) \times \cdots \times (1-\varepsilon_i) \times \cdots \times (1-\varepsilon_{k-1}) \quad (4)$$

$$p' < p \quad (5)$$

这与假设 p 为最小值相矛盾; 所以当 p 取最小值时必有:

$$\min(E_i) \geq \max(E_j), E_i = \{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{m-1}\}$$
$$E_j = \{\varepsilon_m, \varepsilon_{m+1}, \cdots, \varepsilon_{k-1}\} \quad (6)$$

因为测试数据是多类别的, m 应在一个整数范围内选值. 通过对 DDAG 结构的观察, 可发现所有数据类别必然在前 $k-1$ 个结点中出现. 前 m 层结

点数为 $m(m+1)/2$, 可得:

$$\frac{m(m+1)}{2} \le (k-1) \Rightarrow$$
$$m \le \sqrt{2(k-1)+0.25} - 0.5 \tag{7}$$

令 $m=2$, 由式 (6) 可得:

$$\varepsilon_1 \ge \max(\varepsilon_2, \varepsilon_3, \cdots, \varepsilon_{k-1}) \tag{8}$$

令 $m=3$, 得:

$$\min(\varepsilon_1, \varepsilon_2) \ge \max(\varepsilon_3, \varepsilon_4, \cdots, \varepsilon_{k-1}) \tag{9}$$

结合式 (8), 得:

$$\varepsilon_1 \ge \varepsilon_2 \ge \max(\varepsilon_3, \varepsilon_4, \cdots, \varepsilon_{k-1}) \tag{10}$$

依次类推, 当取最小风险概率 p 时, 必有:

$$\varepsilon_1 \ge \varepsilon_2 \ge \cdots \ge \varepsilon_{a-1} \ge \max(\varepsilon_a, \varepsilon_{a+1}, \cdots, \varepsilon_{k-1}),$$
$$a = \lfloor \sqrt{2(k-1)+0.25} - 0.5 \rfloor$$

□

而结点处划分的风险概率与训练样本的间隔密切相关[9]: 设在内积空间 X 上阈值化具有单位权重向量的实值线性函数 $\hbar$ 并固定 $\gamma \in \mathbf{R}^+$, 在 $X \times \{-1, 1\}$ 上的任意概率分布 D, 在以原点为球心, 半径为 R 的球内, 在 ℓ 个随机样例集 S 上具有间隔 $m_s(f) \ge \gamma$ 的假设的误差 $err_D(f)$ 以概率 $1-\delta$ 满足:

$$err_D(f) \le \varepsilon(\ell, \hbar, \delta, \gamma) =$$
$$\frac{2}{\ell}\left(\frac{64R^2}{\gamma^2}\ln\frac{e\ell\gamma}{4R}\ln\frac{128\ell R^2}{\gamma^2} + \ln\frac{4}{\delta}\right),$$
$$\ell > \frac{2}{\varepsilon}, \frac{64\ R^2}{\gamma^2} < \ell \tag{11}$$

从上式可以得出: 对于结点处训练样本, 其离分类超平面的间隔 γ 越大, 风险概率 p 越小. 设所有样本采用 SVM 线性可分, 那么样本间隔 $d = 2\gamma$. 若使决策系统的划分风险概率最小, 则不同层结点处样本间隔应满足:

$$d_1 \le\ d_2 \cdots \le\ d_{a-1} \le \min(d_a, d_{a+1}, \cdots, d_{k-1}),$$
$$a = \lfloor \sqrt{2(k-1)+0.25} - 0.5 \rfloor \tag{12}$$

不同结点具有不同的划分风险概率, 传统 DDAG 法忽视了这一点, 也没有考虑到不同计算次数给划分结果带来的影响, 其结点的排部是随意的. 由式 (12) 可知道, 若降低决策系统划分的风险概率, 其前 a 层结点处样本分类间隔需呈升序排列. 因此本文提出了基于结点优化的 DDAG 多分类 SVM 扩展方法 (Node-refined DDAG-SVM, nrDDAG-SVM).

设有训练样本 $S = \{S_1, S_2, \cdots, S_k\}$, S_i 为第 i 类的样本数据, 采用本文所提出的 nrDDAG-SVM 多分类方法对数据进行多分类, 采用 Matlab 具体实现步骤如下:

步骤 1. k 类样本数据两两组合, 构建 $k(k-1)/2$ 个二分类 SVM. $SVM_{i,j}$ 代表用样本数据 S_i 和 S_j 训练所得分类器.

步骤 2. 根据 $SVM_{i,j}$ 的支持向量及其对应拉格朗日乘子, 计算 S_i 和 S_j 的样本间隔 $d_{i,j}$.

步骤 3. 寻找具有最小样本间隔的类别 (S_a, S_b) : $\arg\min\limits_{i,j} d_{i,j}$, $i, j \in (1, 2, \cdots, k)$; S_a, S_b 作为新流程的根结点训练样本. 即令 $(S_a, S_b) \Longrightarrow (M_1, M_k)$ (训练样本重新排列为 $M = \{M_1, M_2, \cdots, M_k\}$).

步骤 4. 寻找与 S_a, S_b 距离最近的样本数据用来训练下一层结点:

S_c : $\arg\min\limits_{i} d_{a,i}$ $(i = 1, 2, \cdots, k, i \ne a, b)$, $S_c \Longrightarrow M_2$;

S_d : $\arg\min\limits_{i} d_{b,i}$ $(i = 1, 2, \cdots, k, i \ne a, b, c)$, $S_d \Longrightarrow M_{k-1}$.

步骤 5. $L = \lfloor \frac{k}{2} \rfloor$; 采用步骤 4 中流程进行 L 次运算, 将所有训练样本据进行重新排部: $S = \{S_1, S_2, \cdots, S_k\} \Longrightarrow M = \{M_1, M_2, \cdots, M_k\}$.

步骤 6. 参照图 1, 建立样本 M 的 DDAG 决策流程, 对测试数据 X 进行分类, 得到所属类别 M_j.

步骤 7. 得到测试数据基于 M 的分类结果 M_j, 根据步骤 5 得到的 $S \Longleftrightarrow M$ 对照关系表, 得到测试数据 X 的划分类别 S_i.

通过上述流程得到的 DDAG 结构, 其结点处风险概率基本呈降序排列, 降低了决策的总体风险概率.

但是, 必须说明的是, 本文所提定理为决策系统风险最小的一个必要条件, 而非充要条件. 根据该定理设计的 nrDDAG-SVM 多分类方法, 虽然能够合理地优化结点排部, 降低决策的总体风险, 但不能确保所得结构一定具有最小的决策风险.

3　基于 nrDDAG-SVM 的变压器故障诊断

为验证本文算法的有效性, 本文采用 nrDDAG-SVM 法用于变压器故障诊断. 以下实验均在 PC 机 (奔腾 2.8 G, 512 M 内存) 上 Matlab 7.1 环境下完成.

3.1　数据集准备

实验采用的变压器故障数据来源于文献 [10].

该数据采用油中溶解气体分析法对变压器运行状况进行测量. 充油电器设备内的绝缘油和有机绝缘材料在电和热的作用下会产生各种低分子烃类气体, 溶解在油内的各种气体含量与变压器的运行状况密切相关. 选用 H_2、CH_4、C_2H_6、C_2H_4、C_2H_2 5 种气体作为测量对象, 针对变压器正常工作情况、过热、低能放电、高能放电这 4 种运行状况进行故障数据选取, 进而采用 nrDDAG-SVM 法用于变压器故障诊断. 实验的训练样本集和测试样本集分布如表 1.

表 1 实验所用故障数据集

Table 1 Fault datasets for the experiments

故障类型	故障编号	训练样本数	测试样本数
正常工作	0	5	4
过热	1	25	13
低能放电	2	5	6
高能放电	3	15	2

3.2 nrDDAG 结点优化

针对训练样本, 采用本文提出方法对结点排部进行优化, 最终获得 DDAG 结构如图 2. 采用该结构对支持向量机进行多分类扩展, 然后对测试样本进行划分. 分类结果如表 2 所示, 该方法正确划分全部 25 个样本, 分类正确率为 100 %.

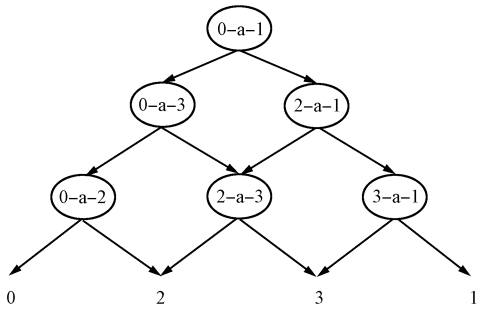


图 2 结点优化所得 DDAG 结构

Fig. 2 DDAG structure obtained by node refined method

表 2 nrDDAG-SVM 分类结果

Table 2 Classification results using nrDDAG-SVM

故障类型	故障编号	训练样本数	自检正确率	测试样本数	测试正确率
正常工作	0	5	100 %	4	100 %
过热	1	25	100 %	13	100 %
低能放电	2	5	100 %	6	100 %
高能放电	3	15	100 %	2	100 %

在对比实验中, 本文采用所有可能的 DDAG 结构对支持向量机进行多分类扩展, 并统计各结构所对应的分类正确率, 如图 3 所示.

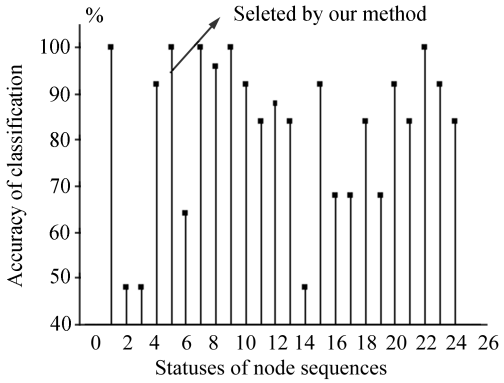


图 3 全部可能的 DDAG 结构及其分类正确率

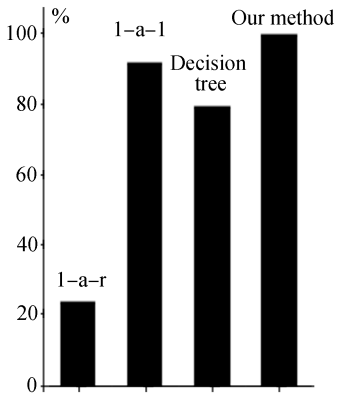
Fig. 3 Correct classification ratios for all possible DDAG structures

传统的 DDAG 其结点排部顺序是随意的, 对应的分类正确率也在不断变化. 针对该变压器故障数据, 传统方法只有 5/24 的概率得到较理想结果. 采用 nrDDAG 法, 无需进行重复计算就能够直接给出分类正确率较大的 DDAG 结构.

3.3 诊断能力测试

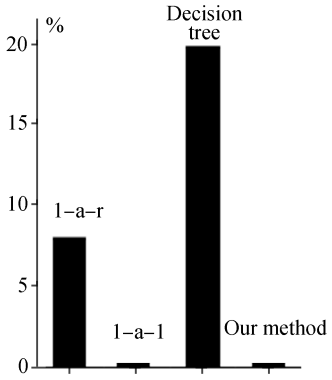
分别采用 1-a-r 法、1-a-1 法、决策树法和 nrDDAG 法对变压器故障数据进行诊断划分. 因前两种方法存在分类盲区, 在本实验中分别对各算法的正确分类率和错误分类率进行统计, 以比较算法在针对变压器故障数据时的划分能力.

如图 4 所示, 本文所提方法与 1-a-1 方法具有最小的错误划分率, 但本文方法的运算结果不存在盲区点, 所以其正确分类率大于 1-a-1 法. 本文方法具有最高的正确分类率.



(a) 正确划分率

(a) Ratios of correct classification

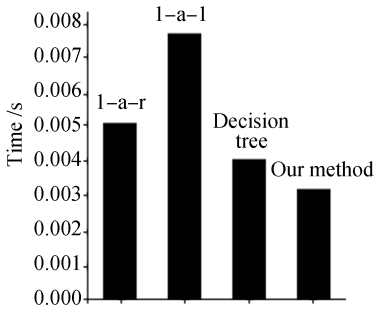


(b) 误判比率

(b) Ratios of incorrect classification

图 4　诊断能力测试

Fig. 4　Testing for the capability of diagnosis



(b) 测试时间

(b) Time for testing

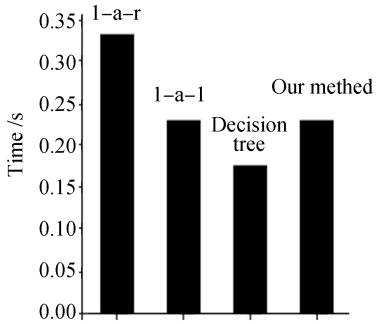
图 5　运算速度测试

Fig. 5　Testing for the speed of computing

3.4　运算速度测试

运算速度是衡量算法性能的重要指标. 训练速度决定构建分类器所需时间, 决策速度决定分类器对样本分类所需时间. 在本实验中, 分别采用 1-a-r 法、1-a-1 法、决策树法和本文提出的 nrDDAG 法对变压器故障数据进行处理, 并记录各算法的训练时间和决策时间.

采用 Matlab 自带计时工具, 对 4 种方法的训练时间和决策时间分别进行记录, 其运行时间如图 5 所示. 通过图 5 可以发现, 本文提出的 nrDDAG 法与决策树法具有最小的决策时间. 在训练时间上, 因为训练样本量较少, 样本量对训练时间的影响被削弱, 而决策树法构建的二分类器最少, 所以该方法需要最少的训练时间. 而 nrDDAG 法需要构建二分类器较多, 所以该方法所需训练时间略长. 综合训练时间和决策时间, nrDDAG 时间消耗仅高于决策树法, 具有比较满意的运算速度.



(a) 训练时间

(a) Time for training

在以上变压器故障数据实验中, 通过 nrDDAG 法进行多分类扩展的支持向量机, 在保持算法效率的基础上, 比传统多分类支持向量机具有更高的分类准确性.

4　结论

支持向量机是一种具有良好前景的故障诊断方法, 但因其本质为两分类器, 必须对其进行多分类扩展才可应用于故障诊断领域. 常规的多分类扩展策略存在样本不对称、样本划分盲区、运算效率低等诸多问题. 本文对其中具有较好应用前景的 DDAG 法进行改进. 首先证明了风险概率最小的 DDAG 结构, 其前 $a\,(a = \lfloor\sqrt{2(k-1)+0.25} - 0.5\rfloor)$ 层结点处样本间隔必然呈升序排列. 根据这一特性, 提出结点排部优化的 DDAG 多分类扩展策略支持向量机 (nrDDAG-SVM). 通过对结点的优化排部, 在不增加运算量的前提下, 降低了支持向量机决策系统错误划分的风险概率.

References

1 Zhou Dong-Hua, Ye Yin-Zhong. *Modern Fault Diagnosis and Fault-Tolerant Control*. Beijing: Tsinghua University Press, 2000
(周东华, 叶银忠. 现代故障诊断与容错控制. 北京: 清华大学出版社, 2000)

2 Jiang B, Staroswiecki M, Cocquempot V. Fault accommodation for nonlinear dynamic systems. *IEEE Transactions on Automatic Control*, 2006, **51**(9): 1578−1583

3 Zhang Xue-Gong. Introduction to statistical learning theory and support vector machines. *Acta Automatica Sinica*, 2000, **26**(1): 32−42
(张学工. 关于统计学习理论与支持向量机. 自动化学报, 2000, **26**(1): 32−42)

4 Vapnik V N. *Statistical Learning Theory*. New York: Wiley, 1998

5 Knerr S, Personnaz L, Dreyfus G. Single-layer learning revisited: a stepwise procedure for building and training a neural network. *Neurocomputing: Algorithm, Architectures and Applications*. New York: Springer-Verlag, 1990

6 Takahashi F, Abe S. Decision-tree-based multiclass support vector machines. In: Proceedings of the 9th International Conference on Neural Information. Washington D. C., USA: IEEE, 2002. 1418−1422

7 Platt J C, Cristianini N, Shawe-Taylor J. Large margin DAG′s for multiclass classification. In: Proceedings of Neural Information Processing Systems. Massachusetts: MIT Press, 2000. 547−553

8 Kijsirikul B, Ussivakual N. Multiclass support vector machines using adaptive directed acyclic graph. In: Proceedings of IEEE International Joint Conference on Neural Networks. Honolulu, USA: IEEE, 2002. 980−985

9 Cristinanini N, Shawe-Taylor J. *An Introduction to Support Vector Machines*. Cambridge: Cambridge University Press, 2000

10 Ganyun L V, Cheng H Z, Zhai H B, Dong L X. Fault diagnosis of power transformer based on multi-layer SVM classifier. *Electric Power Systems Research*, 2005, **74**(1): 1−7

易 辉 南京航空航天大学自动化学院博士研究生. 主要研究方向为智能分类和故障诊断. E-mail: jsyihui@126.com
(**YI Hui** Ph. D. candidate at the College of Automation Engineering, Nanjing University of Aeronautics and Astronautics. His research interest covers intelligent classification and fault diagnosis.)

宋晓峰 南京航空航天大学自动化学院副教授. 主要研究方向为智能计算.
E-mail: xfsong@nuaa.edu.cn
(**SONG Xiao-Feng** Associate professor at the College of Automation Engineering, Nanjing University of Aeronautics and Astronautics. His main research interest is intelligent computation.)

姜 斌 南京航空航天大学自动化学院教授. 主要研究方向为故障诊断与容错控制. 本文通信作者.
E-mail: binjiang@nuaa.edu.cn
(**JIANG Bin** Professor at the College of Automation Engineering, Nanjing University of Aeronautics and Astronautics. His research interest covers fault diagnosis and fault tolerant control. Corresponding author of this paper.)

王定成 南京信息工程大学计算机与软件学院副研究员. 主要研究方向为人工智能和智能控制.
E-mail: dcwang2005@126.com
(**WANG Ding-Cheng** Associate professor at the Institute of Computer and Software, Nanjing University of Information Science and Technology. His research interest covers artifical intelligence and intelligent control.)