[文章编号] 1003－4684(2019)02-0005-05

输电线路故障分析多分类模型研究及应用

张 伟[1]，陈建峡[1]，李 超[1]，黄煜俊[1]，徐欣雨[2]
(1 湖北工业大学计算机学院，湖北 武汉 430068；2 日本东京大学工程学院，日本 东京)

[摘 要] 为了有效而准确地分析智能电网中输电线路故障和寻找诱发故障的主要因素，研发了基于二叉树的核密度逻辑回归多分类模型，以解决输电线路不对称故障分析的问题。该模型根据 Nadaraya-Watson 密度估计将训练数据映射到了特定的特征空间，根据二叉树结构特点将多个 DLR 模型组合成一个具有多分类能力的二叉树，并加以优化。实验结果表明，基于 MCDLR 的分类结果在准确率上和分类时间上明显优于已有的传统的多分类算法。

[关键词] 核密度逻辑回归；核密度估计；二叉树；输电线路；故障分析
[中图分类号] TP301.6 [文献标识码] A

随着智能电网建设的迅猛发展，各种数据挖掘技术越来越多地运用在电力系统智能诊断中。其中，输电线路大量的故障信息分布极其不对称，适合利用机器学习方法进行处理[1]。机器学习方法的应用不仅能提高电力系统实时运行信息的广度和深度，而且能够动态分析输电线路的健康状态，提高电力系统的运维决策水平。对于输电线路来说，除了故障分析算法的准确率外，最重要的就是分类算法的可解释性，即可以得知是哪些因素在诱发某一类故障的过程中起主要作用，得知这些因素可以在部署和维护电网时重点关注避免这些因素的产生，从而减少甚至避免一些故障的发生。逻辑回归具有很好的可解释性，SVM 可解释性较差。如 KNN，神经网络等非线性模型则不具备可解释性[2-3]。

逻辑回归 LR(Logistic Regression，LR)是常用的机器学习方法，用于估计某种事物的可能性。逻辑回归延伸了多元线性回归思想，用来测量分类结果与因变量之间的关系。对于某一给定的类别，逻辑回归能给出相应的类分布估计，并且在模型训练时间上占很大优势[4]。然而，逻辑回归属于线性分类器。对于非线性分类边界的输电线路来说，线性分类器的训练很难基于数据集来确定参数。基于核密度估计的逻辑回归算法 DLR(Density Estimation Logistic Regression，DLR)解决了这一问题，其主要思想是根据 Nadaraya-Watson 核密度估计将训练数据映射到特定的特征空间，然后组建优化模型优化特征权重 Nadaraya-Watson 密度估计算法的带宽。在分类精度和时间效率上，DLR 算法明显优于 LR、基于径向基函数 RBF(Radial basis function，RBF)内核的核逻辑 KLR(Kernel logistic regression，KLR)、SVM 等算法模型。

在实际应用中，二叉树分类模型在故障诊断领域中能够取得较好的预测效果。文献[4]提出一种依赖故障优先级的 SVM 二叉树分类器模型实现方法，将其应用到柴油机振动信号的故障诊断中，取得了满意的分类效果。文献[5]将改进二叉树模型应用到输电线路故障诊断中，文中结合输电线路故障的特点设计了一种二叉树分类器模型，结果表明二叉树多分类方法明显优于其它多分类方法。但是，当分类故障较多的情况下，二叉树算法不仅需要庞大的训练样本集，其分类树结构也很复杂，计算量大大增加。同时，故障类型较多时会出现样本不平衡问题，收敛困难，诊断的准确率大大降低，严重阻碍了二叉树分类器在故障诊断中的推广与应用。

针对上述问题，首先采用二叉树算法对输电线路故障样本集进行分析和聚类计算，采用优先分级判别的方法，逐步细化到所有故障都被分离。其次，采用 DLR 核逻辑密度回归算法，有效解决了样本不平衡问题，提高了分类准确度。

[收稿日期] 2018－06－22
[基金项目] 湖北省科技厅自然科学基金青年面上项目(2017CFB326)
[第一作者] 张 伟(1992－)，男，湖北黄冈人，湖北工业大学硕士研究生，研究方向为机器学习、云计算与大数据
[通信作者] 陈建峡(1971－)，女，湖北丹江口人，工学硕士，湖北工业大学副教授，研究方向为机器学习，云计算与大数据

1 MCDLRBT 多分类模型

1.1 DLR 模型

DLR 模型主要思想是根据 Nadaraya-Watson[6]密度估计将数据映射到特定的特征空间，并且组建优化模型优化特征权重 $w(x)$ 及 Nadaraya-Watson 密度估计算法的带宽 h。对于 $y_i \in \{0,1\}$，输入向量为 $x_{i,d}$，$i=1,2,\cdots,n$，$d=1,2,\cdots,D$ 时，即：类标签 y 为二分类的情况时，DLR 的概率模型如下：

$$p(y=1 \mid x)=\frac{1}{1+\exp(-w^T\varphi)}=\frac{1}{1+\exp(-\sum_{d=1}^{n} w_d\,\varphi_d(x))} \tag{1}$$

其中，$\varphi_i(x)$ 为特征映射函数，定义如下：

$$\varphi_d(x)=\ln\frac{p(y=1 \mid x_d)}{p(y=0 \mid x_d)}-\frac{n-1}{n}\ln\frac{p(y=1)}{p(y=0)} \tag{2}$$

式(2)中等式右边第一项为考虑 x_d 时 $y=1$ 的概率，第二项为数据集类别不平衡的量度。因此，$\varphi_d(x)$ 也可以认为是当 x_d 时，y 为 1 的似然概率的评价量度，w_d 的大小评价了第 d 个特征变量对于 y 被标记为 1 的贡献度[7]。

对于所有的数值变量 x_d，DLR 模型中使用 Nadaraya-Watson 估计器来估计 $p(y=k \mid x_d)$，$k=0,1$，得到

$$p(y=k \mid x_d)=\frac{\sum_{i\in Dk} K(\frac{x_d-x_{i,d}}{h_d})}{\sum_{i=1}^{n} K(\frac{x_d-x_{i,d}}{h_d})} \tag{3}$$

其中 Nadaraya-Watson 估计器中的核函数 $K(x)$ 采用的是高斯核函数，则 DLR 模型中特征映射函数 $\varphi_d(x)$ 如

$$\varphi_d(x)=\ln\frac{\sum_{i\in D1}\exp(-\frac{(x_d-x_{i,d})^2}{2h_d^2})}{\sum_{i\in D0}\exp(-\frac{(x_d-x_{i,d})^2}{2h_d^2})}-\frac{D-1}{D}\ln\frac{p(y=1)}{p(y=0)} \tag{4}$$

在某种程度上，核函数定义了两个数据实例之间的相似性。例如，高斯核函数解释基于距离的相似性。式(4)中 DLR 模型中的核函数定义了对某一维度中两个实例之间的相似度，而整体的相似度则由每一维的相似度整合来定义[7]。

DLR 模型中对于 Nadaraya-Watson 的 h 参数调整，是基于最大化训练集整体似然函数来优化参数 h 的。定义 $b_i=p(y_i=1 \mid x_i)$，则整体似然函数如

$$E(w,h)=-\sum_{i=1}^{n}\{y_i\,b_i+(1-y_i)\ln(1-b_i)\} \tag{5}$$

令 $r_d=-1/2\,h_d^2$，通过计算 $\frac{\partial E}{\partial r_d}$ 的导数，可以得到

$$\frac{\partial E}{\partial h_d}=\frac{1}{h_d^3}\sum_{i=1}^{n}(b_i-y_i)\,w_d\,\frac{\partial\varphi_d(x_i)}{\partial r_d} \tag{6}$$

令 $h=(h_1,h_2,\cdots,h_D)$，式(6)梯度函数 h 即可得到。然而，h 的二阶偏导数很难计算，最终导致在 w 和 h 的联合空间中 Newton 算法很难实现[8]。因此，DLR 模型中采用两层的分层优化框架来优化 h。第一层采用式(4)的核密度估计算法，来计算相应的特征 $\varphi_d(x)$。第二层中，先固定 h，根据训练集中的数据通过 Newton 算法来优化 w。最终根据式(6)，通过计算导数的下降方向来优化 h(一种常用的经验法则中 h 的取值[9] 设定为：$h_d=1.06\sigma n^{-1/5}$)。

DLR 算法过程如图 1 所示。

Algorithm 1 Hierarchical optimization for DLR learning

```
Initialize h using (19)
repeat                                   ▷ outer loop: optimize h
    Assemble the feature matrix Φ under h
    repeat                               ▷ inner loop: fix h and optimize w
        w ← w − ∇_w E / ∇²_w E           ▷ on the training set
    until w converges
    for d = 1 to D do                    ▷ fix w and update h_d
        if x_d is a numerical attribute then
            h_d ← h_d − γ ∂E/∂h_d        ▷ on the validation set
        end if
    end for
until h converges
```

图 1 DLR 算法伪码

1.2 一对多分类优化模型

DLR 算法是一个具有可解释性的面向二分类问题的非线性分类算法。所以将 DLR 应用输电线路多故障识别时，需要构造多级 DLR 模型。

常见的多分类算法模型有一对一[10]、一对多[11]、二叉决策树[12]。在上述多分类模型中，基于二叉树的 DLR 多分类算法(MCDLRBT)加快了二叉决策树模型的训练速度，又解决了二叉决策树应用于多分类时数据分布不平衡的问题。

一般的二叉决策树是由多个二分类模型构建的多分类模型。本文采用的二分类算法为 DLR 核密度逻辑回归算法，随机将类别作为叶子节点构建二叉决策树，本文输电线路故障一共有 10 类，分别用 1—10 表示类别且分配给 10 个叶子节点，随机组合 2 个无父节点的叶子节点构建其上层节点，逐层组合直到出现根节点为止，构建完成的某一种一般二叉决策树模型见图 2。

假设类别标签数为 N，则构建完整的一般二叉决策树需要($N-1$)个 DLR 二分类模型，类别越多叶子节点随机组合的结果也就多，即可构建的二叉决策树模型种类也就越多，也越容易产生误差，在一般二叉决策树模型的分类过程中，如果上层节点出

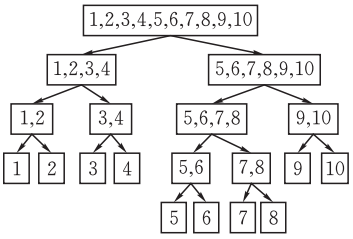

图 2　常见二叉决策树模型

现误差，那么该误差会累积传递给下一层，造成分类结果不准确。且一般二叉决策树在构建时容易出现树及其不平衡的情况，影响预测准确度。

基于准确度和模型训练时间的考虑，本文设计出一种改进一对多方法的多分类模型。该模型在构建上类似于一对多方法，第一步选出一类作为正类并记为 1，其它的各类均为负类并记为 0，第二步利用 DLR 核密度逻辑回归算法进行二分类模型训练，且已经标记为正类的数据不再参加下面其它模型的训练，从而达到逐步减少训练数据的目的，有利于减少模型的训练时间，第三步在第一步的负类中选出一类作为正类记为 1，负类中其它剩余各类标记为负类记为 0，进行二分类模型的训练。第四步重复以上步骤直到负类仅剩一类不可再分为止，最后构建成的多分类模型如图 3 所示。

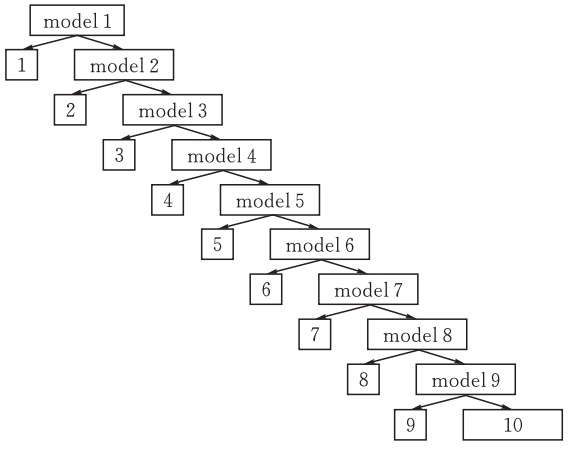

图 3　改进后的一对多分类模型

1.3　MCDLRBT 多分类模型

MCDLRBT 模型的关键在于分类层级的构建，改进一般二叉决策树多分类模型的构建方法。由于二叉决策树一般构建完成后都是固定的，如何组织训练样本来生成分类精度高的二叉决策树是本文的研究重点。在研究中发现，二叉决策树的结构对整个多分类模型的分类准确度有很大影响，由于二叉树本身的原因，基于二叉树的多分类模型的误差会跟随二叉树的节点逐级累积向下传递，比如在某一个节点预测错误，那么该节点下的所有节点就全部会出错，预测准确性会极大降低。这一现象叫做误差累积[13]，最坏的结果是根节点出现错误，那么整个分类器将失去意义，所以 MCDLR 多分类模型的关键在于让误差出现的点尽可能的远离根节点，使误差造成的影响降到最低。

本文利用各个不同类别的故障样本之间的距离远近作为标准，将各故障样本逐层分离出来。最后构建出最符合当前需求的 MCDLR。

输电线路故障分为 10 种类型，本文采用 Kmeans 聚类算法[6]计算出各个聚类中心，分别计算各类别之间聚类中心的距离，对于聚类中心距离近的类，说明在特征值上差别相对较小，即这些类更加具有相似性，为了避免出现分类错误的现象，所以这些有相似性的类别越远离根节点越有利于避免误差累积，因此，本文将距离相近的类别归为正类记为 1，距离较远的类别归为负类记为 0.然后进入下一层，分别在正类和负类中重复以上过程，直到模型构建完成，最终 MCDLRBT 模型见图 4。

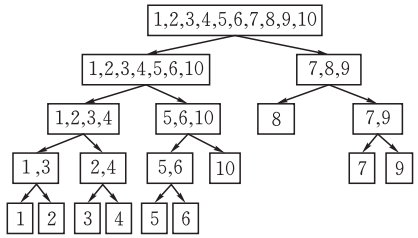

图 4　MCDLRBT 多分类模型

2　基于 MCDLRBT 的输电线路故障分析模型

MCDLRBT 模型在整个输电线路故障分析中的应用流程见图 5。首先读取数据，对数据进行预处理，第二步，利用 Kmeans 算法[6]计算出各个类别中心之间的距离，第三步，构建 MCDLR 模型，第四步，对已经构建好的 MCDLR 模型进行测试，分析模型对故障预测的准确性。

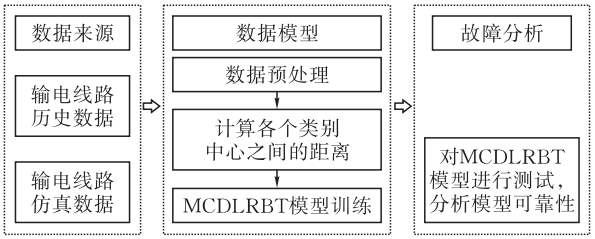

图 5　输电线路故障分析流程图

3　输电线路故障分析多分类模型实验

3.1　实验环境与数据

文中用到的实验数据是通过 MATLAB/Simulink 环境下的 PSB 模型库及 Simulink 强大的二次开发功能和丰富的工具箱，对输电系统进行仿真模拟生成。PSB（Power System Blocker）是 MAT-

LAB 软件中电力系统模块集，它主要由加拿大的 Hydro Quebec 和 TECSIM International 公司共同开发，其功能非常强大，可以用于电路、电力电子系统、电机系统、电力传输等过程的仿真，它提供了一种类似电路建模的方式进行模型绘制，在仿真前会自动将其变成状态方程描述的系统形式，然后在 Simulink 环境下进行仿真分析[14]。

文中主要对 10 种输电线路不对称故障类型进行预测分析，包括单相短路接地故障（AG，BG，CG），两相短路故障（AB，AC，BC），两相短路接地故障（ABG，ACG，BCG）和三相短路故障（ABC）。参照电力系统故障分析原理，实验中选取故障后与故障前三相电压及三相电流值[15]，故障点距离线路位置从线路总长度的 10％到 90％处，以 10％递增。数据样本为 10 条不同类型的输电线路的不同位置，共 2800 条数据。

3.2 实验结果与分析

如图 6 所示，横轴表示各个多分类模型，纵轴表示训练样本数据量，从数据样本集中随机抽取 1700 条数据进行训练，并从数据样本集中随机抽取去标签的 820 条数据进行故障分析。

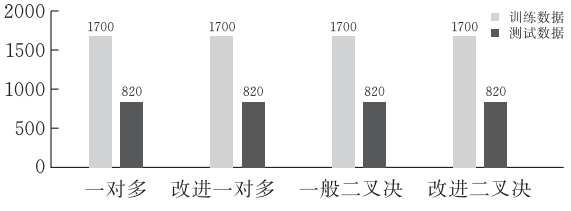

图 6 各模型使用训练与测试数据情况

MCDLRBT 算法模型中，训练阶段计算得出的训练模型权重值 W，如图 7 所示，参数 M_i 表示不同电压等级的输电线路，W_i 表示不同训练模型中的特征权重值。奇数行中权重值 W_i 的训练结果之所以为零，是因为这些维度的数据值表示的是故障发生前相同的三相电压和三相电流正常值。另外，显而易见的是每个训练模型的训练精度 P 都高于 95％，甚至达到 99.9％（表 1）。

MCDLRBT 算法模型的测试阶段中，分别带入训练阶段得到的训练模型，进行计算后获取概率结果最大值，来判断测试样本的故障类型。测试数据的类别结果三维散点图，如图 7 所示。其中，故障后 A 相、B 相、C 相电流值，分别为三维散点图的 X 轴、Y 轴和 Z 轴。各个多分类模型故障分析结果见表 2。

表 1 训练模型权重值及其精度值

	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$	$M7$	$M8$	$M9$	$M10$
$W1$	0	0	0	0	0	0	0	0	0	0
$W2$	0.29	0.18	0	0	0	0	0	0	0.41	1.00
$W3$	0	0	0	0	0	0	0	0	0	0
$W4$	0.78	0	0	0.16	0.28	0.48	0.42	0.24	1.10	1.25
$W5$	0	0	0.03	0	0	0	0	0	0	0
$W6$	0	0	0	0.58	0.47	0.41	0.75	0.22	0.17	1.16
$W7$	0	0	0	0	0	0	0	0	0	0
$W8$	1.32	2.31	2.75	0.66	0.44	0.64	0.94	1.56	0.85	0.91
$W9$	0	0	0	0	0	0	0	0	0	0
$W10$	2.84	1.20	2.50	0.59	0.68	0.46	0.96	0.76	1.53	0.95
$W11$	0	0	0	0	0	0	0	0	0	0
$W12$	2.10	1.83	1.45	0.52	0.60	0.63	1.42	0.90	0.63	1.14
p	0.98	0.98	0.96	0.96	0.96	0.96	0.95	0.95	0.94	0.98

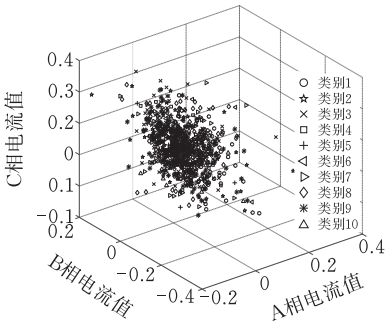

图 7 MCDLR 测试结果三维散点图

表 2 各种分类模型故障分析结果

方法	一对多	改进一对多	决策树	MCDLR
训练数据	1700	1700	1700	1700
测试数据	820	820	820	820
模型个数	10	9	9	9
训练时间/s	15.23	8.51	7.46	5.34
准确度/％	91	90	79	96

在训练数据和测试数据相同的情况下：根据表 2 可知，一对多分类模型在训练时间上开销较大，因为一对多分类模型在模型构造上所需要的二分类模型个数为故障类别数（即类别数为 N，二分类模型数为 N），比其它三个模型多一个，这是因为，传统一对多模型每次参加模型训练的数据量不变，实际类别数量也不变，其余三个模型，在每次训练后，类别记为 1 的数据被分离出来，不再参加后面的模型训练，即每次模型训练之后类别数较上一次少一个。虽然传统一对多模型每一条数据都要参加模型训练，计算量大。但是其误差累积对其整体影响不大，所以准确度较高，故一对多有较好的准确度但需要较大的时间开销。

改进一对多模型所需要的二分类模型个数比故障类别数小 1（即类别数为 N，二分类模型数为 $N-1$），且每一层都比上一次参加训练的类别少一类，所

以每一层模型训练时间都比上一层少，但是该模型的误差累积对整体的影响较大，故改进一对多模型在模型训练的时间开销上要优于一对多算法，但是准确率上要略低于一对多算法。

一般二叉决策树模型所需要的二分类模型个数比故障类别数小 1(即类别数为 N，二分类模型数为 $N-1$)，且每一层比上一层参加训练的类别少一类或多类，所以每一层参加训练的数据量比上一层参加训练的数据少，且减少的幅度大于等于改进一对多模型，随着故障类型的增加，训练数据量减少幅度也会变大，所以整个多分类模型较改进一对多分类来说在训练时间消耗上将大大减少，但是，一般二叉决策树模型容易受误差累积影响，故一般二叉决策树模型在模型训练的时间开销上要少于一对多模型和改进一对多模型，但是其准确率要远低于一对多模型和改进一对多模型。

MCDLRBT 多分类模型是在一般二叉决策树模型上所做的改进模型，其所需要的二分类模型个数比故障类别数小 1(即类别数为 N，二分类模型数为 $N-1$)，且每一层比上一层参加训练的类别少一类或多类，所以每一层参加训练的数据量比上一层参加训练的数据少，且减少的幅度比改进一对多模型大，而且由于本文优化了在构建二叉决策树的时候的类别选取策略，使得误差的产生远离了根节点，所以，MCDLRBT 多分类模型受误差累积的影响较小，能得到远高于其它三种模型的分类准确度，。故 MCDLRBT 在模型训练的时间开销上和预测准确度上都要由于其它三种多分类方法。

综上所述，MCDLRBT 在模型训练的时间开销上和预测准确度上都要优于其它三种多分类方法。

4　总结

根据输电线路故障样本的特点，提出基于二叉树的 MCDLRBT 多分类模型优化了二叉决策树在多分类领域的应用，减少了传统的一般二叉决策树在多分类领域中常见的误差累积现象。实验结果表明，MCDLRBT 多分类模型在输电线路故障分析中取得了较高的准确度(96%)。

[参 考 文 献]

[1] Zhan Cailiang. Application of AI technology in fault diagnosis of power system[J]. Guangdong Electric Power,2011,24(9):87-92.

[2] Zhang H, Berg A C, Maire M, et al. Svm-knn: Discriminative nearest neighbor classification for visual category recognition[J]. Proc.ieee Conf.computer Vision & Pattern Recognition, 2006, 2:2126-2136.

[3] 虞和济，陈长征，张省. 基于神经网络的只能诊断[J]. 振动工程学，2000，13(2)：202-209.

[4] Mao Lin, Lu Quanhua, Cheng Tao. The research and application of ensemble logistic regression classification algorithm based on high dimensional data[J]. Bulletin of Science and Technology, 2013,29(12):64-66.

[5] Mao Yi, Chen Wenlin, Guo Baolong, Chen Yixin. A novel logistic regression model based on density estimation[J]. Acta Automatica Sinica, 2014, 40(1): 62- 72.

[6] 党华丽.Matlab/Simulink 仿真在信号与系统分析中的应用[J].信息技术，2010(3)：123-127.

[7] 马笑潇，黄席樾，柴毅.基于 SVM 的二叉树多类分类算法及其在故障诊断中的应用[J].控制与决策，2003，18(3)：272-276.

[8] 韩邦合，李永明.计算逻辑学中的误差累积理论[J]计算工程与应用，2009，45(23)：9-10.

[9] Silverman B W, Green P J. Density estimation for statistics and data analysis[M]. London: Chapman and Hall, 1986.

[10] kuronya A, Lozovanu V. Infinitesimal newton-okounkov bodies and jet separation[J]. Duke Mathematical journal 2017,166(7):1349-1376.

[11] 李昆仑，黄厚宽，田盛丰，一种基于有向无环图的多类 SVM 分类器[J].模式识别与人工智能，2003，16(2)：164-168.

[12] 袁胜发，褚福磊.SVM 多类分类算法及其在故障诊断中的应用[J].振动工程学报，2004，17(5)：419-421.

[13] 吕晓丽，李雷，曹未丰.基于二叉树的 SVM 多类分类算法[J].信息技术 2008(4)：1-3.

[14] 王成江，马新明，官云，等.基于 SVM 的改进二叉树输电线路故障分类器[J].电力系统保护与控制，2010，38(5)：39-44.

[15] Eltibi M F, Ashour W M. Initializing kmeans clusteringalgorithmusing. statistical. information[J]. International Journal of Computer Applications 2013,29(7): 51-55.

(下转第 21 页)