基于模拟退火的SVDD特征提取和参数选择

邢红杰 赵浩鑫
(河北大学数学与计算机学院 河北省机器学习与计算智能重点实验室 保定 071002)

摘 要 支持向量数据描述(Support Vector Data Description,SVDD)被认为是用于异常检测的典型方法。众所周之,参数的设置和特征的品质是影响SVDD性能的两个关键点。将SVDD的特征提取和参数选择问题结合在一起,提出了一种基于模拟退火的SVDD特征提取和参数选择方法(SA-SVDD)。在模拟退火的过程中,自动选择最优核参数、折衷参数以及抽取特征的维数。在UCI基准数据集上的实验结果表明,与传统的参数选择方法相比,SA-SVDD取得了更优的性能。

关键词 特征提取,模拟退火,参数选择,SVDD,异常检测
中图法分类号 TP391.4 **文献标识码** A

Feature Extraction and Parameter Selection of SVDD Using Simulated Annealing Approach

XING Hong-jie ZHAO Hao-xin
(Key Laboratory of Machine Learning and Computation Intelligence, College of Mathematics and Computer Science, Hebei University, Baoding 071002, China)

Abstract Support vector data description (SVDD) is considered as a classical method for novelty detection. As is well known, the parameter setting and the quality of features are two key points to affect the performance of SVDD. Combining feature extraction and parameter selection of SVDD, this paper proposed a simulated annealing approach for feature extraction and parameter selection of SVDD (SA-SVDD). During the procedure of simulated annealing, the optimal kernel parameter, trade-off parameters, and number of extracted features are automatically selected. Experimental results on the UCI benchmark data sets demonstrate that SA-SVDD has better performance than the traditional parameter selection methods.

Keywords Feature extraction, Simulated annealing, Parameter selection, SVDD, Novelty detection

1 引言

异常检测是一类特殊的分类问题,在训练过程中只有正常数据参与训练,它通过捕获正常数据的特征来检测新的数据是否是异常点。异常检测已经成为模式识别、机器学习等领域的研究热点,在故障诊断[1]、图像分割[2]、网络入侵检测[3]等方面有着广泛的应用。迄今为止,出现了许多异常检测方法[4-7],大致可以将它们分为3类,即基于统计的方法(主要包括基于近邻的方法、参数方法、非参数方法和半参数方法)、基于神经网络的方法(包括基于有监督学习的神经网络方法和基于无监督学习的神经网络方法)、基于机器学习的方法(如决策树)。

近年来又出现了一些新的异常检测方法,如一类支持向量机(One-class Support Vector Machine,OCSVM)[8]、核主成分分析(Kernel Principal Component Analysis,KPCA)[9]、单类极小极大概率机(Single-class Minimax Probability Machine)[10]。OCSVM是一种广为流行的单类分类器,它通过寻找最优超平面,将正常数据和原点以最大间隔进行划分。然而,OCSVM仅使用正常数据进行训练,在很多情况下,异常数据确实存在,然而异常数据往往又很少,不足以构建一个两类分类器。为了解决这个问题,Gori等人提出了自相关神经网络(Auto-associative Neural Network)[11]。Tax和Duin提出了支持向量数据描述(Support Vector Data Description,SVDD)[12]。SVDD试图寻找一个最小包围球,使其包含尽可能多的正常数据和尽可能少的异常数据,它在模式识别领域得到了广泛的应用,但是SVDD的参数(包括核参数γ、正类折衷参数C_1、负类折衷参数C_2)直接影响SVDD的性能,这些参数的选择仍是一个未解的难题。在构建SVDD时,一般采用格搜索对上述参数加以选择。格搜索方法首先设定参数范围,然后把它分成若干区间,一个区间一个区间地逐次测试以寻找最优参数。格搜索方法存在以下缺陷:如何划分区间仍没有统一的标准,时间复杂度非常高,容易陷入局部最优等。

此外,一般的数据集中会经常存在冗余特征。这些冗余特征不仅浪费大量的学习时间和空间,而且还会降低SVDD的性能。传统的格搜索方法仅被用于参数选择,却不能同时

到稿日期:2012-03-02 返修日期:2012-06-02 本文受国家自然科学基金项目(60903089,61073121,61170040),河北大学基金项目(2008123,3504020)资助。

邢红杰(1976—),男,博士,副教授,主要研究方向为模式识别、机器学习,E-mail:hjxing@hbu.edu.cn。

消除冗余特征。最近，支持向量机(SVM)的参数选择和特征选择方法有了新的进展。一些智能优化方法，如模拟退火[13]、遗传算法[14]、粒子群算法[15]被用于SVM的参数及特征选择，取得了很好的结果。文献[13]中提出的基于模拟退火的SVM参数及特征选择(SA-SVM)能自动地选取最优参数和特征子集。

受到SA-SVM的启发，本文提出了一种基于模拟退火的SVDD参数选择和特征提取方法(SA-SVDD)。SA-SVDD主要有两方面的工作。一方面，在模拟退火的过程中，最优核参数 γ、正类折衷参数 C_1、负类折衷参数 C_2 可自动获得。另一方面，为了消除数据中的冗余特征，使用主成分分析(PCA)特征提取方法进行降维，此处，抽取特征的数目也被自动选择。

与相关方法对比，所提SA-SVDD有如下优点：

• 不易陷入局部最优：它不需要将参数空间划分成若干小区间，也不必一个格一个格跳跃式地搜索，因此，与格搜索相比，SA-SVDD有范围更大的取值空间且不易陷入局部最优；

• 有效剔除冗余特征：在完成参数选择的同时，它能自动选择抽取特征的维数并完成特征提取，而在以往的方法中，模拟退火往往仅被用于参数选择，因此不能在完成参数选择的同时剔除冗余特征。

• 性能良好：它的时间消耗比格搜索低，却有更高的分类准确率。

本文第2节简要地回顾了传统的SVDD方法；第3节简单描述了模拟退火算法；第4节详细阐述了所提SA-SVDD方法；第5节通过实验验证了所提方法在UCI标准数据集上能够取得优于其相关方法的性能；最后给出了概括性总结。

2 SVDD

SVDD是由Tax和Duin提出的一种常用的单类分类方法[12]。给定数据集 $X=\{x_1,x_2,\cdots,x_n\}$，其中正常数据 $T=\{x_1,x_2,\cdots,x_p\}$，异常数据 $O=\{x_{p+1},x_{p+2},\cdots,x_n\}$，SVDD试图寻找一个包含所有正类数据的最小包围球，球的中心设为 a，半径为 R。为了得到紧致的边界，球半径应尽可能的小。另外，在一定程度上允许某些数据落于球外，故引入松弛变量 ξ_i、ξ_l。SVDD的原问题如下：

$$\begin{aligned}&\min_{R,a,\xi_i,\xi_l} R^2+C_1\sum_{i=1}^{p}\xi_i+C_2\sum_{l=p+1}^{n}\xi_l\\&\text{s.t. } \|x_i-a\|^2\leqslant R^2+\xi_i\\&\quad \xi_i\geqslant 0,i=1,2,\cdots,p\\&\quad \|x_l-a\|^2\geqslant R^2-\xi_l\\&\quad \xi_l\geqslant 0,l=p+1,p+2,\cdots,n\end{aligned}\tag{1}$$

式中，C_1 是正类折衷参数，C_2 是负类折衷参数，即模型复杂性与准确性之间的折衷。为了求解优化问题(1)，需要使用拉格朗日乘子法，即构造如下的拉格朗日函数：

$$\begin{aligned}L(R,a,\xi_i,\xi_l,\alpha_i,\alpha_l,\gamma_i,\gamma_l)=&R^2+C_1\sum_{i=1}^{p}\xi_i+C_2\sum_{l=p+1}^{n}\xi_l-\sum_{i=1}^{p}\\&\gamma_i\xi_i-\sum_{l=p+1}^{n}\gamma_l\xi_l-\sum_{i=1}^{p}\alpha_i[R^2+\xi_i-(x_i-a)^{\mathrm{T}}(x_i-a)]-\\&\sum_{l=p+1}^{n}\alpha_l[R^2+\xi_l-(x_l-a)^{\mathrm{T}}(x_l-a)]\end{aligned}\tag{2}$$

从而，可得原问题(1)的对偶问题如下：

$$\begin{aligned}\max_{\alpha}&\sum_{i=1}^{p}\alpha_i k(x_i,x_i)-\sum_{l=p+1}^{n}\alpha_i k(x_l,x_l)-\sum_{i=1}^{p}\sum_{j=1}^{p}\alpha_i\alpha_j k(x_i,x_j)\\&+2\sum_{i=1}^{p}\sum_{l=p+1}^{n}\alpha_i\alpha_l k(x_i,x_l)-\sum_{l=p+1}^{n}\sum_{m=p+1}^{n}\alpha_m\alpha_l k(x_m,x_l)\\\text{s.t. }&\sum_{i=1}^{p}\alpha_i-\sum_{l=p+1}^{n}\alpha_l=1\\&0\leqslant\alpha_i\leqslant C_1,i=1,2,\cdots,p\\&0\leqslant\alpha_l\leqslant C_2,l=p+1,p+2,\cdots,n\end{aligned}\tag{3}$$

式中，$k(x_i,x_j)$ 是核函数，本文采用径向基函数，即 $k(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|^2)$。

对于测试样本 z，如果有

$$\begin{aligned}\|z-a\|^2=&k(z,z)-2\sum_{i=1}^{p}\alpha_i k(z,x_i)+2\sum_{l=p+1}^{n}\alpha_l k(z,x_l)+\\&\sum_{m=1}^{n}\sum_{s=1}^{n}\alpha_m\alpha_s k(x_m,x_s)\leqslant R^2\end{aligned}\tag{4}$$

则 z 为正常数据，否则，z 为异常数据。

从以上的理论推导可知，SVDD的性能主要由参数 γ、C_1 及 C_2 决定。同时，如果数据中存在冗余特征，也会对SVDD的性能造成影响。为了提高SVDD的性能，需要提出同时进行参数选择和特征提取或特征选择的方法。

3 模拟退火算法

本节将对模拟退火算法进行简单的介绍。模拟退火算法(SA)是由Metropolis等人提出的一种常用的启发式随机搜索算法，它从解空间中的某个随机初始点开始，然后在初始点周围产生一个随机扰动来生成新解，若新解朝着更优的方向移动，就接收该解，否则以一定概率接受它。换言之，模拟退火算法不仅接收较优的解，在一定情况下还接收一些较差的解，这有助于模拟退火算法避免陷入局部最优，而收敛到全局最优。

文献[13]对Romeijn等人提出的“Hide-and-Seek”模拟退火方法进行了详细的介绍，本文所提方法采用了“Hide-and-Seek”模拟退火方法。图1描述了“Hide-and-Seek”模拟退火算法的主要过程。

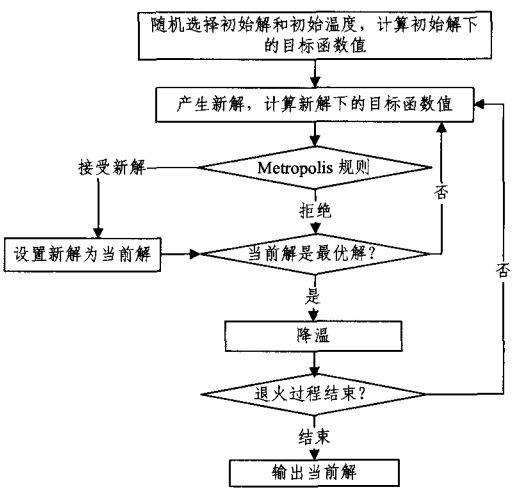

图1 “Hide-and-Seek”模拟退火的过程

它与传统的模拟退火主要有两点区别。第一，新解的产生方法。传统的模拟退火方法是在初始解的近邻空间产生新解，而“Hide-and-Seek”模拟退火方法产生的新解在所有可行解空间中，所以“Hide-and-Seek”模拟退火方法更合理，省去了不必要的迭代步骤；第二，传统的模拟退火方法需要等退火周期循环结束时才进行退火，而“Hide-and-Seek”模拟退火方法只要找到一个使目标函数更优的解，就立即退火。

4 SA-SVDD

本节将对基于模拟退火的 SVDD 特征提取和参数选择方法进行详细的描述。为了消除冗余特征，在参数选择过程中同时使用了 PCA 特征提取方法。SA-SVDD 需要将参数选择和特征提取结合起来统一处理。因此，在模拟退火的过程中需要同时确定 4 个参数，即核参数 γ、折衷参数 C_1 和 C_2，以及抽取特征的维数 d。图 2 给出了解的表示形式。在没有特征提取的情况下，仅需确定前 3 个参数。

1	2	3	4
r	C_1	C_2	d

图 2 解的表示形式

SA-SVDD 的构造过程可以概括为以下 4 步。

(1)数据尺度变换

文献[13]指出，数据尺度变换可以克服不同特征取值范围差异较大的缺陷。本文亦采用文献[13]中数据尺度变换的方法来将数据的特征取值范围变换到区间[−1,+1]上，即采用如下的变换公式：

$$f'_{ij}=\left(\frac{f_{ij}-\min\{f_{ij}\mid i=1,2,\cdots\}}{\max\{f_{ij}\mid i=1,2,\cdots\}-\min\{f_{ij}\mid i=1,2,\cdots\}}\right)*2-1 \tag{5}$$

式中，f_{ij} 是第 i 个样本的第 j 个特征，而 f'_{ij} 是 f_{ij} 变换后的特征值。

(2)设定退火参数

在模拟退火过程中，主要涉及 4 个参数，分别是初始温度 T_0、冷却速度、最大迭代次数 Itr、可行解空间。初始温度应选取得足够高，以防止模拟退火算法陷入局部最优的陷阱中；冷却速度是指温度下降的速率，在理论中模拟退火算法只有在对数函数 $\ln k$ 速率下降时才收敛到全局最优。但在实际中其下降速率太慢，很难满足需要，于是出现了许多近似方法，本算法中“Hide-and-Seek”模拟退火方法采用的就是其中一种，可行解空间就是所求解所在的范围。

(3)产生初始解与新解

在可行解空间中，初始解的设定采用的是随机选择的方法，然后为初始解增加较小的随机增量，来产生一个新的可行解。

(4)目标函数的设定

将 SVDD 的验证准确率作为目标函数，此处，采用几何平均数(g-means)[16]来计算验证准确率，其表达形式如下：

$$gm=\sqrt{ap*an} \tag{6}$$

式中，ap 和 an 分别表示正常数据和异常数据的验证准确率。在计算目标函数时，在训练集上采用了 5 折交叉验证，并将 5 次的平均验证准确率作为目标函数值。

在算法 1 中，对计算目标函数的过程做了总结。

SA-SVDD 不断地进行“产生新解→计算目标函数差→判断是否接受新解→接受或舍弃新解”的过程，ΔE 表示新解的目标函数值与当前解的目标函数值之间的差值，如果 $\Delta E\geqslant 0$，新解就会被接受，否则就产生一个随机数 $\gamma\in[0,1]$；若 $e^{\frac{\Delta E}{T}}<\gamma$，新解亦会被接受。若新解的目标函数值大于当前最优解的目标函数值，即将新解作为最优解，立刻降温。算法 2 展示了 SA-SVDD 的构造过程。

算法 1 SA-SVDD 的目标函数值计算

输入：训练集 $X=\{x_1,x_2,\cdots,x_k\}$ 和一个解 $S'=[r',c_1,c_2,d']$

输出：$S'=[r',c_1,c_2,d']$ 的目标函数

第 1 步 在交叉验证的每一轮中，将训练集 X 划分成训练部分 Tr 和验证部分 Va。

第 2 步 使用 PCA，将 Tr 中的数据降到 d′维。

第 3 步 使用参数 r',c_1,c_2 来训练 SVCD。

第 4 步 在验证集 Va 上使用 PCA 降维，降维后的维数为 d′。

第 5 步 在 Va 上计算验证准确率，然后回到第 1 步，开始下一轮的交叉验证，直到 5 次都执行完，转向第 6 步。

第 6 步 以第 5 步中的平均验证准确率作为目标函数值。

算法 2 基于模拟退火的 SVDD 特征提取和参数选择

输入：训练集 $X=\{x_1,x_2,\cdots,x_k\}$ 和可行解空间，最大迭代次数 Itr

输出：最优参数 $S_{opt}=[r^*,c_1^*,c_2^*,d^*]$

第 1 步 设置一个随机可行解 $S_0=[r,c_1,c_2,d_1]$ 和一个初始温度 $T=T_0$，$S_{opt}=S_0$，令 denom=1.0/(0.9^2)−1.0，chi2=自由度为 4 的 χ^2 分布 0.99 分位数，t=1

第 2 步 在可行空间中，对初始解 S_0 产生一个随机扰动，生成新解 S_{new}。

第 3 步 根据算法 1 分别计算初始解和新解的目标函数值 $f(S_0)$ 和 $f(S_{new})$。

第 4 步 判断新解是否被接受

```
ΔE=f(S_new)−f(S_0);
if(ΔE≥0)
    acprate=1;
else
    acprate=e^(ΔE/T);
end
生成(0,1)之间的随机数 ε;
if (acprate≥ε)
S_0=S_new;
end
```

第 5 步 判断当前解是不是最优解，若是最优解，立即退火。

```
if (f(S_opt)≤f(S_0))
{
    fx2=f(S_opt),fx1=f(S_0),ftmp=(fx1−fx2)/denom;
    T=2*(ftmp−fx1)/chi2,S_opt=S_0;

}
end
t=t+1;
```

第 6 步 如果 t<Itr，退回第 2 步；否则，退出。

5 实验结果

为了评价所提 SA-SVDD 的分类性能，本节将它与基于模拟退火的参数选择方法(未经特征提取)、格搜索方法进行了对比。在实验比较中，使用了 6 个 UCI 数据标准数据集，即 Iris、Breast cancer (Wisconsin)、Biomed、Automobile、Sonar 和 Wine。需要指出的是，有些数据集的类别个数多于两类。为了使上述数据集适用于异常检测，将相应的数据集人为地分为两类。具体操作如下：对于 Iris 数据集，将类别为 Iris-setosa 的数据作为正类数据，而类别为 Iris-versicolor 的样本用作异常数据。对于 Breast cancer 数据集，将属于 malignant 的

数据作为正类数据，其余数据作为异常数据。Automobile 中，将第一个特征值小于等于 0 的数据用作正类数据，其余的数据作为异常数据。Sonar 中，将属于 mines 的数据用作异常。Wine 中，类别标签为 1 的样本是正类数据，其余的数据用作异常数据。

此外，随机选择 70%的正类数据和小部分的异常数据来构建训练集[16]。在本文的训练集中，91%的训练数据是正常数据，而只有 9%的训练数据是异常数据。为了保证在 SVDD 训练过程有异常数据参与训练，训练集中必须保证至少有 5 个异常样本。实验数据总结在表 1 中。对于每个数据集，#pos 表示整个数据集中正类数据的个数，#neg 是整个数据集中异常数据的个数。训练集中正常数据个数用 $m1$ 表示，$m2$ 则是训练集中异常数据的个数。

表 1　实验数据

数据集	特征个数	#pos	#neg	m1	m2
Iris	4	50	100	35	5
Breast cancer	9	241	458	169	17
Biomed	5	127	67	89	9
Automobile	25	71	88	49	5
Sonar	60	111	97	78	8
Wine	13	59	119	41	5

在实验中，将初始温度设定为 107，最大迭代次数设为 300。解空间的范围为 $\gamma=[0.001,32]$，$C_1=[0.01,500]$，$C_2=[0.01,500]$，$d=[1,nfea]$，$nfea$ 是数据维数。3 种方法在所有的数据集上均重复 5 次，并使用 5 次的平均测试准确率来衡量它们最终的分类效果。实验结果如表 2 所列。

表 2　3 种方法测试准确率的比较

数据集	格搜索	不带特征提取的模拟退火	SA-SVDD	
	平均测试准确率			抽取特征的个数
Iris	0.9298±0.0515	0.9377±0.0299	1±0	1±0
Breast cancer	0.9227±0.0263	0.9210±0.0292	0.9529±0.0327	1.4000±0.8944
Biomed	0.7445±0.0586	0.7775±0.0393	0.8158±0.0441	2.2000±0.8367
Automobile	0.6845±0.0347	0.7222±0.0575	0.7789±0.0270	16.200±6.8702
Sonar	0.6262±0.0257	0.6382±0.0242	0.6780±0.0204	28.4000±19.7434
Wine	0.8359±0.0608	0.8604±0.0180	0.9161±0.0354	6.4000±1.6733

注：表中数据全部是 5 次结果的平均值±标准差的形式，故 SA-SVDD 抽取特征的个数出现了小数。

从表 2 的结果可以发现，本文所提 SA-SVDD 在所有数据集上的分类性能均优于未进行特征提取的模拟退火方法以及格搜索方法。具体而言：

(1)在所有数据集上，带特征提取的 SA-SVDD 都比格搜索要好。SA-SVDD 不仅平均测试准确率高，而且只使用了较少的特征。

(2)SA-SVDD 比基于模拟退火的 SVDD(未经特征提取)性能要好，这表明在参数选择过程中同时进行特征提取的必要性。此外，SA-SVDD 不仅得到了最好的平均测试准确率，而且也达到了最高测试准确率，如表 3 所列。

表 3　3 种方法的最高测试准确率

数据集	格搜索	不带特征提取的模拟退火	SA-SVDD	
	最高测试准确率			抽取特征的个数
Iris	0.9661	0.9661	1	1
Breast cancer	0.9440	0.9361	0.9794	1
Biomed	0.8218	0.8315	0.8733	2
Automobile	0.7285	0.7901	0.8174	18
Sonar	0.6627	0.6667	0.7029	38
Wine	0.9048	0.8819	0.9043	6

结束语　本文中提出了基于模拟退火的 SVDD 特征提取和参数选择方法(SA-SVDD)。所提方法结合了参数选择和特征提取的优点。该文的贡献主要有两方面：第一，SA-SVDD 为 SVDD 提供了一种自动的参数选择方法，它比传统的格搜索效率更高。第二，在模拟退火的过程中，同时加入特征提取方法，可以有效地剔除数据集中的冗余特征。据此，与基于模拟退火参数选择的 SVDD(未经特征提取)方法相比，SA-SVDD 能够在减少计算消耗的同时提高分类性能。实验结果验证了所提方法的有效性。

参 考 文 献

[1] Worden K. Structural fault detection using a novel measure[J]. Sound and Vibration, 1997, 201(1): 85-101

[2] Singh S, Markou M. An approach to novelty detection applied to the classification of image regions [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(4): 396-407

[3] 周颖杰，胡光岷，贺伟淞. 基于时间序列图挖掘的网络流量入侵检测 [J]. 计算机科学，2009，36(1)：46-50

[4] Markou M, Singh S. Novelty detection: a review—part 1: statistical approaches [J]. Signal Processing, 2003, 83(12): 2481-2497

[5] Markou M, Singh S. Novelty detection: a review—part 2: neural network based approaches [J]. Signal Processing, 2003, 83(12): 2499-2521

[6] 潘志松，陈斌，缪志敏，等. One-Class 分类器研究 [J]. 电子学报，2009，37(11)：2496-2503

[7] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey [J]. ACM Computing Surveys, 2009, 41(3)

[8] Schölkopf B, Williams R C, Smola A J, et al. Support vector method for novelty detection [J]. Massachusetts Institute of Technology Press, 2000, 12(3): 582-588

[9] Hoffmann H. Kernel PCA for novelty detection [J]. Pattern Recognition, 2007, 40(3): 863-874

[10] Lanckriet G R G, Ghaoui L E, Jordan M I. Robust novelty detection with single-class MPM [C]// Advance in Neural Information Processing Systems. 2003: 905-912

[11] Gori M, Lastrucci L, Soda G. Autoassociator-based models for speaker verification [J]. Pattern Recognition Letters, 1996, 17(3): 241-250

[12] Tax D M J, Duin R P W. Support vector data description [J]. Machine Learning, 2004, 54(1): 45-66

[13] Lin S W, Lee Z J, Chen S C, et al. Parameter determination of support vector machine and feature selection using simulated annealing approach [J]. Applied Soft Computing, 2008, 8(4): 1505-1512

[14] 刘东平，单甘霖，张岐龙，等. 基于改进遗传算法的支持向量机参数优化 [J]. 微计算机应用，2010，31(5)：11-15

[15] 朱凤明，樊明龙. 混沌粒子群算法对支持向量机模型参数的优化 [J]. 计算机仿真，2010，27(11)：183-186

[16] Wu M R, Ye J P. A small sphere and large margin approach for novelty detection using training data with outliers [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2088-2092