第34卷，第3期
2014年3月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 34, No. 3, pp638-642
March, 2014

基于遗传算法的多目标最小二乘支持向量机在近红外多组分定量分析中的应用

徐 冰[1]，王 星[1,3]，Dhaene Tom[2]，史新元[1*]，Couckuyt Ivo[2]，白 雁[3]，乔延江[1*]

1. 北京中医药大学中药信息工程研究中心，北京 100029
2. Ghent University-iMINDS, Department of Information Technology, Gent B-9050, Belgium
3. 河南中医学院，河南 郑州 450008

摘 要 近红外(NIR)定量分析通常涉及多个组分，采用遗传算法和自适应建模策略，建立了能够对多组分同时定量的多目标最小二乘支持向量机(LS-SVM)，并将其应用于玉米中四个组分和连翘中两个活性成分的NIR分析。结果表明多目标遗传算法配合自适应建模策略可保证优化收敛于全局最优解。所建玉米多目标LS-SVM模型明显优于PLS1和PLS2模型；连翘多目标LS-SVM模型与PLS模型均可取得较好的校正和预测效果。两组数据中，径向基神经网络(RBFNN)模型均出现过拟合现象。多目标LS-SVM和单目标LS-SVM性能相近，但多目标LS-SVM建模运行一次即可得到结果，在NIR多组分定量分析中具有潜在应用优势。

关键词 多目标最小二乘支持向量机；遗传算法；近红外；多组分定量；自适应建模

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2014)03-0638-05

引 言

近红外(near infrared, NIR)分析技术具有快速、无损等特点，在农业、食品、化工和医药等领域应用广泛[1,2]。近红外光谱中包含丰富的物理化学信息，可实现多组分同时定量，常用的方法是分别针对每个指标单独建立分析模型。PLS2算法和人工神经网络模型(ANN)可以处理多目标同时定量的问题[3]。但文献报道PLS2模型的性能往往低于PLS1[4,5]；而ANN在训练过程中易出现过拟合，导致其泛化能力受到一定限制[6]。

最小二乘支持向量机(least square support vector machine, LS-SVM)具有建模速度快、优化参数少、泛化能力强等优点，广泛应用于近红外定性定量分析。研究表明，通过调节LS-SVM超参数(正则化参数和核函数参数)，或对各目标组分设置权重，可构造多任务LS-SVM和多输出LS-SVM，用于近红外多组分同时定量。但多任务LS-SVM采用两步网格搜索法调节模型参数，效率较低；多输出LS-SVM要为不同组分配置权重，使超参数优化变得困难。

近年来，启发式算法被越来越多地应用于LS-SVM超参数优化，如遗传算法、粒子群优化、模拟退火等[7,8]。其共同特点是多点并行搜索，按照一定的信息传递方式，逐渐逼近全局最优解。采用LS-SVM进行近红外多组分定量，其实质是多目标建模及优化的问题，因而可尝试采用多目标启发式算法进行超参数优化，建立同时满足多个组分定量分析要求的LS-SVM，即多目标LS-SVM。本文以农作物玉米和中药材连翘为例，联合多目标遗传算法和自适应建模策略，建立了同时测定玉米中四个组分和连翘中两个活性成分的近红外多目标LS-SVM模型，并取得了满意的效果。

1 原理和方法

根据统计学习理论中的结构风险最小化原则，LS-SVM的最优决策函数可表示为以下形式

$$\hat{y} = w^{\mathrm{T}}\varphi(x) + b \tag{1}$$

式中φ表示从x样本空间到高维特征空间的映射，w为权重向量，b为偏差。式(1)的求解可转化为如下优化问题

$$\min J(w, e) = \frac{1}{2}\|w\|^2 + \frac{1}{2}\gamma\sum_{i=1}^{n} e_i^2$$
$$\text{s.t. } y_i - w^{\mathrm{T}}x_i - b = e_i \tag{2}$$

收稿日期：2013-04-16，**修订日期**：2013-07-15

基金项目：国家“重大新药创制”科技重大专项(2010ZX09502-002)资助

作者简介：徐 冰，1985年生，北京中医药大学中药学院讲师 e-mail：btcm@163.com

*通讯联系人 e-mail：yjqiao@263.net；shixinyuan01@163.com

式中 γ 为正则化参数，控制着经验风险和置信风险之间的平衡；n 为训练集样本数目；e 为训练误差。采用拉格朗日法将优化问题转化为求解线性方程组[9]，最后可得 LS-SVM 模型为

$$\hat{y}_i = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b \tag{3}$$

式中 α 代表拉格朗日乘子，$K(x, x_i)$ 为满足 Mercer 条件的核函数。由于径向基函数(radical basis function，RBF)具备较强的处理非线性问题的能力，受样本离群值的干扰较小[10]，因此本文选择 RBF 作为核函数

$$K(x, x_i) = \exp\left(\frac{\| x - x_i \|^2}{\sigma^2}\right) \tag{4}$$

在利用 LS-SVM 模型进行 NIR 定量分析时，还需考虑三个方面的问题，即模型输入、超参数优化和误差方程的选择。NIR 光谱数据中不仅包含有用信息，也包含大量噪声和冗余信息，因此在建模前应对光谱进行适当处理。常用的预处理方法包括平滑、滤波、降维和变量筛选等。其中，采用 PLS 将原始光谱降维处理，不仅可滤除噪音，保留光谱主要信息，而且将降维后得到的潜变量因子作为模型输入可提高 LS-SVM 建模的运算速度[11]，因此本文选择 PLS 算法将 NIR 光谱进行降维处理。

由式(2)和式(4)可知，为了降低 LS-SVM 模型的训练误差，需要对正则化参数 γ 和 RBF 核函数参数 σ 进行优化。本文采用多目标遗传算法在 LS-SVM 模型超参数空间内寻优搜索，以获得同时满足多个目标的参数 γ 和 σ 值。在超参数求解过程中，应用自适应建模策略(如图 1 所示)[12]，具体步骤如下：

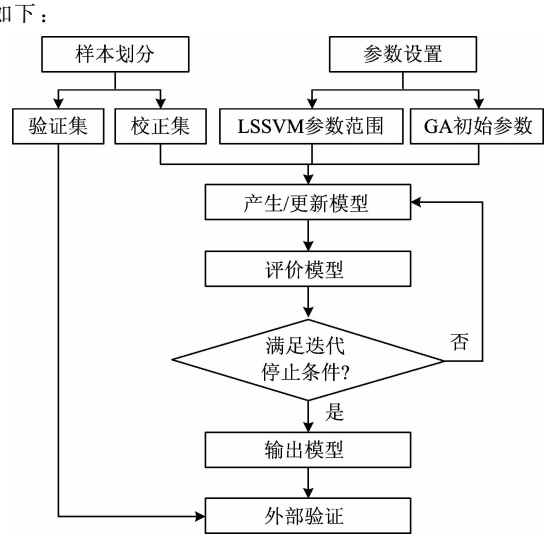

Fig. 1　Scheme of adaptive modeling

Step 1：将 NIR 数据划分为校正集和验证集，对光谱进行预处理，采用 PLS2 算法将光谱降维；

Step 2：设置正则化参数 γ 和 RBF 核函数参数 σ 的搜索区间；设定多目标遗传算法参数，包括初始种群大小、最大进化代数、交叉比例等；设定自适应建模终止条件，如使模型的训练误差低于某一数值、达到最大迭代建模次数或时间等；

Step 3：利用多目标遗传算法搜索参数 γ 和 σ 的 Pareto 前沿，用 Pareto 前沿中的超参数建立 LS-SVM 模型；

Step 4：采用交叉验证的方式计算模型训练误差，对模型性能进行评估，本文使用的误差方程为相对平方根误差(root relative squared error，RRSE)

$$\mathrm{RRSE} = \sqrt{\frac{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{m} (y_i - \bar{y}_i)^2}} \tag{5}$$

式中 m 为校正集样本数目。以各组分交叉验证 RRSE(即 RRSEcv)的均值作为该模型得分。若用 Step 3 中的某一可行解建立的模型得分小于之前所有模型，则将其存储为最优模型(best model)，并继续搜索；

Step 5：若不满足 Step 2 中的终止条件则重复 Step 3 和 Step 4，若满足 Step 2 中终止条件则停止自适应建模并输出最优模型。

2　实验部分

2.1　玉米数据

玉米近红外光谱测定数据为 Eigenvector 公司提供的开源数据[13]，广泛应用于化学计量学方法的评估测试。本文采用该数据集中 m5 近红外光谱仪测定的数据，包括 80 个样品。光谱扫描范围为 1 100～2 498 nm，扫描间隔 2 nm。参考值为玉米中水分、油脂、蛋白质和淀粉的含量。采用 Kennard-Stone(K-S)算法将所有样本划分为校正集(60 个样本)和验证集(20 个样本)，样本划分后的原始光谱直接用于建模分析。

2.2　连翘药材数据

中药连翘(*Forsythia suspensa*(Thunb.)Vahl)数据共包括 102 个样本，连翘药材经干燥、粉碎和过筛等处理后，由 Nicolet 6700 型傅里叶变换近红外光谱仪采集漫反射光谱数据，光谱范围 10 000～4 000 cm^{-1}，分辨率 8 cm^{-1}，测定条件和原始光谱参见文献[14]。参考值为连翘苷(Phillyrin)和连翘酯苷 A(Forsythoside A)的含量。采用 K-S 算法将全部样品划分为校正集(77 个样本)和验证集(25 个样本)，参考值分布范围如表 1 所示。建模前将光谱进行一阶导数处理。

Table 1　Ranges of reference values in the calibration and validation sets

成分	校正集				验证集			
	下限	上限	均值	标准差	下限	上限	均值	标准差
连翘苷/(mg·g^{-1})	0.032	5.135	1.701	1.314	0.439	4.874	1.932	1.411
连翘酯苷 A/%	0.668	5.930	2.307	1.463	0.688	5.410	2.199	1.232

2.3 软件和参数设置

本文涉及的算法均由 Matlab 7.8(MathWorks Inc., U.S.)平台实现，PLS 算法使用 PLS Toolbox 2.1(Eigenvector Research Inc., U.S.)，自适应建模过程由 SUMO Toolbox 7.0.2 完成[15]，LS-SVM 算法使用 LS-SVMlab 1.5 工具箱[10]，RBFNN 算法使用 Matlab 神经网络工具箱，多目标超参数优化使用 Matlab GADS Toolbox，其他程序自行编制。

多目标遗传算法的初始种群大小设为 10，最大进化代数为 200，交叉比例为 0.7，其他参数默认。LS-SVM 中正则化参数 lgγ 搜索区间为[−5, 5]，RBF 核参数 lgσ^2 的搜索区间为[−4, 4]。RBFNN 同样采用基于遗传算法的自适应建模策略构建，隐含层神经元(neuron)最大个数为 60，RBF 扩展系数(spread)优化范围为[0.01, 10]，设计误差目标为 0。模型交叉验证均采用留一法。自适应建模终止条件为：RRSE 值低于 0.01，或单目标最大建模迭代次数达到 25(多目标时为 50)。

3 结果与讨论

3.1 自适应建模过程分析

玉米和连翘 NIR 光谱经 PLS2 算法降维处理后，分别采用 8 和 9 个潜变量因子作为模型输入。在玉米四个组分的多目标 LS-SVM 模型构建中，多目标遗传算法搜索到 369 个 Pareto 前沿(含 3 750 个超参数组合)，并存储了 16 个最优模型；连翘中两个指标性成分的多目标 LS-SVM 模型构建中，共搜索到 251 个 Pareto 前沿(含 2 570 个超参数组合)和 5 个最优模型。绘制最优模型中不同组分的 RRSEcv 变化曲线(见图 2)，结果各组分训练误差开始逐渐降低，随后趋于平缓，表明优化过程可逐渐收敛，自适应建模终止条件设置合理。

图 3 展示了自适应建模过程中的 Pareto 前沿搜索轨迹，坐标以各组分的 RRSEcv 表示。玉米数据以油脂、蛋白质和淀粉为例，绘制了三维空间内的 Pareto 前沿轨迹[图 3(a)]；连翘数据可直接在二维平面内展示 Pareto 前沿轨迹[图 3(b)]。结果表明自适应建模过程中，多目标遗传算法可有效地推动整个种群向最优解区域移动，搜索到的最优解前沿收敛，且均匀分布。在多目标优化中，各子目标之间可能是相互冲突的，以图 3(b)为例，连翘苷 RRSEcv 值的降低会导致连翘酯苷 RRSEcv 值的升高。本文采用的决策方法是最小化各组分 RRSEcv 均值，而在不同情况下，决策者也可根据实际问题选择不同的决策方法。

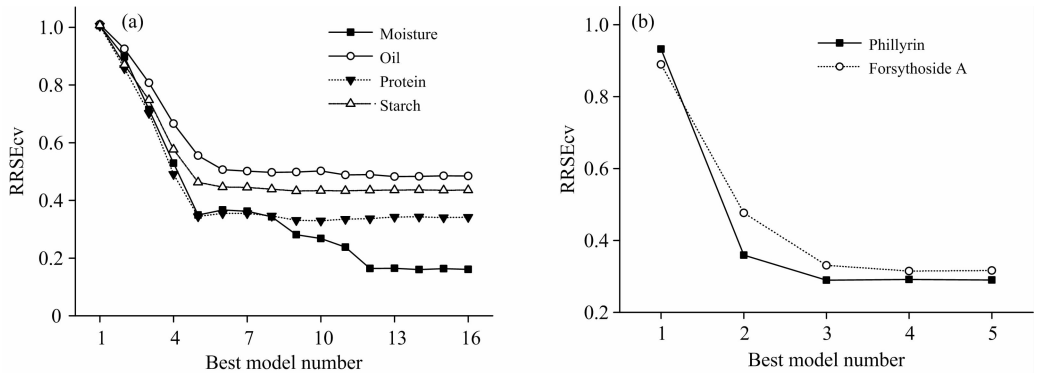

Fig. 2 RRSEcv of different components in the best model trace

(a): Corn; (b): Forsythia suspensa

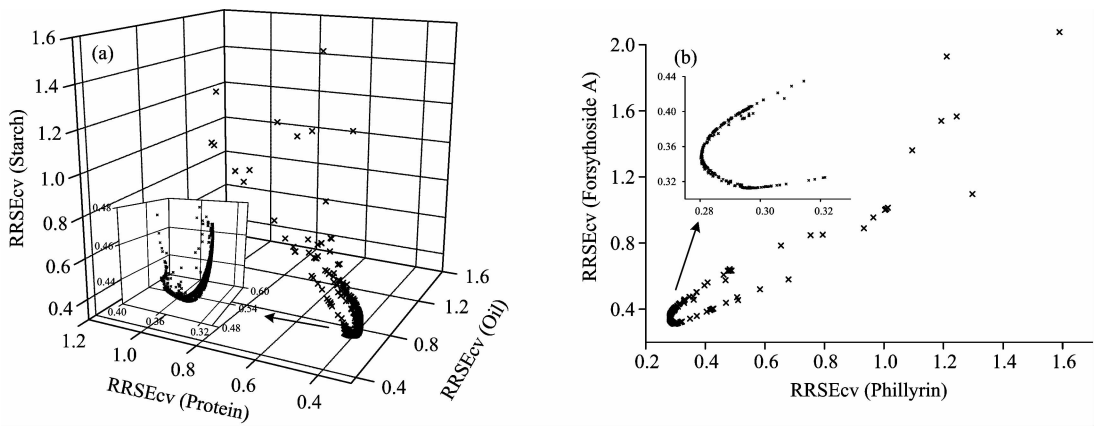

Fig. 3 Pareto front search trace

(a): Corn; (b): Forsythia suspensa

3.2 NIR 定量模型性能评价

根据自适应建模结果，玉米数据和连翘数据分别采用最优模型 16 和 5 进行多组分 NIR 定量分析。多目标 LS-SVM 模型校正和预测结果，分别与单目标模型 LS-SVM 和 PLS1，以及多目标模型 PLS2 和 RBFNN 进行了比较。模型评价指标为校正均方根误差(RMSEC)、预测均方根误差(RMSEP)和相关系数 r，结果见表 2 和表 3。其中玉米多目标 LS-SVM 模型中各指标的校正和预测性能均优于两类 PLS 模型，表现为较低的 RMSEC 和 RMSEP 值和较高的相关系数；而 RBFNN 模型校正性能虽然优于其他模型，但预测效果较差，即出现了“过拟合”现象。在连翘多成分定量模型中，多目标 LS-SVM 和两类 PLS 模型的校正和预测结果差异较小，均可取得较好的效果；但 RBFNN 模型同样出现了过拟合问题。多目标 LS-SVM 与单目标 LS-SVM 模型的性能相当，但通过进一步比较玉米和连翘数据中各类成分在相应模型中的评价指标的均值，可发现多目标 LS-SVM 模型的预测性能略优于单目标 LS-SVM，例如连翘多目标 LS-SVM 的 RMSEP 均值(0.435 3)和 r_{val} 均值(0.9402)优于单目标 LS-SVM(RMSEP 均值 0.440 5，r_{val} 均值 0.938 6)。

Table 2　Performance of different NIR quantitative models (Corn)

模型	组分	模型参数	校正集		验证集	
			RMSEC	r_{cal}	RMSEP	r_{val}
PLS1	水分	LVs 7*	0.286 1	0.735 8	0.280 1	0.733 0
	油脂	LVs 8	0.087 2	0.868 7	0.129 7	0.682 1
	蛋白	LVs 8	0.187 5	0.935 0	0.267 2	0.771 0
	淀粉	LVs 6	1.079 2	0.642 9	0.906 8	0.663 5
PLS2	水分	LVs 8	0.298 8	0.707 4	0.310 9	0.659 9
	油脂		0.103 5	0.815 4	0.130 3	0.682 1
	蛋白		0.211 0	0.916 0	0.300 7	0.699 8
	淀粉		1.035 9	0.556 0	0.951 6	0.565 2
RBFNN	水分	neurons=54; spread=0.60	0.048 1	0.992 4	0.177 1	0.848 5
	油脂		0.032 5	0.982 4	0.005 0	0.872 7
	蛋白		0.052 6	0.994 8	0.268 8	0.767 4
	淀粉		0.108 2	0.991 8	0.473 6	0.751 3
LS-SVM	水分	$\sigma_2=1.22\times10^3;\gamma=9.98\times10^4$	0.046 8	0.992 8	0.064 3	0.981 4
	油脂	$\sigma^2=9.99\times10^3;\gamma=9.96\times10^4$	0.069 8	0.916 2	0.072 0	0.912 7
	蛋白	$\sigma^2=31.76;\gamma=154.44$	0.112 1	0.976 4	0.194 1	0.886 5
	淀粉	$\sigma^2=40.42;\gamma=88.18$	0.264 2	0.950 0	0.376 8	0.850 4
多目标 LS-SVM	水分	$\sigma^2=990.13$ $\gamma=3.00\times10^4$	0.051 5	0.991 3	0.064 8	0.981 1
	油脂		0.064 3	0.929 2	0.067 7	0.923 1
	蛋白		0.134 1	0.965 9	0.186 2	0.896 0
	淀粉		0.277 2	0.944 8	0.380 2	0.847 5

* LVs: Latent variables used in the PLS model

Table 3　Performance of different NIR quantitative models (Forsythia suspensa)

模型	组分	模型参数	校正集		验证集	
			RMSEC	r_{cal}	RMSEP	r_{val}
PLS1	连翘苷	LVs 9	0.277 6	0.973 6	0.561 7	0.923 7
	连翘酯苷 A	LVs 8	0.403 7	0.960 7	0.302 0	0.969 5
PLS2	连翘苷	LVs 9	0.341 6	0.965 0	0.567 5	0.920 2
	连翘酯苷 A		0.395 8	0.962 8	0.325 8	0.963 7
RBFNN	连翘苷	neurons=52; spread=1.73	0.172 4	0.991 3	0.738 3	0.845 3
	连翘酯苷 A		0.152 4	0.994 5	0.619 7	0.858 2
LS-SVM	连翘苷	$\sigma^2=85.15;\gamma=583.51$	0.279 2	0.976 9	0.570 7	0.910 8
	连翘酯苷 A	$\sigma^2=9.99\times10^3;\gamma=7.38\times10^3$	0.392 1	0.962 9	0.310 3	0.966 4
多目标 LS-SVM	连翘苷	$\sigma^2=173.27$ $\gamma=227.59$	0.326 4	0.968 2	0.563 7	0.913 1
	连翘酯苷 A		0.377 8	0.965 6	0.306 9	0.967 2

4 结 论

基于多目标 LS-SVM 模型，成功建立了同时测定玉米和连翘中多个组分含量的 NIR 分析方法。多目标遗传算法可以在 LS-SVM 超参数空间内进行高效搜索，自适应建模策略可充分保证优化收敛于全局最优解。与单目标 LS-SVM 模型相比，多目标 LS-SVM 建模，仅运行一次即可得到结果，通过单一模型即可实现多种成分的同时测定，并可取的较好的校正和预测效果，在涉及多个目标的 NIR 定量分析中具有潜在的应用优势。

References

[1] Haughey S A, Graham S F, Cancouët E, et al. Food Chemistry, 2013, 136(3-4): 1557.
[2] Xu B, Wu Z, Lin Z, et al. Analytica Chimica Acta, 2012, 720: 22.
[3] Blanco M, Peguero A. Talanta, 2008, 77(2): 647.
[4] Gallego J, Arroyo J. Analytica Chimica Acta, 2001, 437: 247.
[5] Picón Z, Martínez G, Garrido F, et al. Analyst, 2000, 125(6): 1167.
[6] Chen Q, Guo Z, Zhao J, et al. Journal of Pharmaceutical and Biomedical Analysis, 2012, 60: 92.
[7] Qu J, Zuo M J. Expert Systems with Applications, 2012, 39(5): 6089.
[8] Chen A, Wu Z, Yang G. Theory and Applications of Models of Computation. Berlin: Springer, 2006: 99.
[9] Thissen U, Üstün B, Melssen W J, et al. Analytical Chemistry, 2004, 76(11), 3099.
[10] Debruyne M, Serneels S, Verdonck T. Journal of Chemometrics, 2009, 23(9): 479.
[11] Ni Y N, Mei M H, Koko S. Chemometrics and Intelligent Laboratory Systems, 2011, 105(2): 147.
[12] Gorissen D, Couckuyt I, Laermans E, et al. Engineering with Computers-Germany, 2010, 26(1): 81.
[13] http://www.eigenvector.com/data/index.htm.
[14] WANG Xing, BAI Yan, CHEN Zhi-hong, et al(王 星，白 雁，陈志红，等). China Journal of Chinese Materia Medica(中国中药杂志), 2009, 34(16): 2071.
[15] Gorissen D, Crombecq K, Couckuyt I, et al. Journal of Machine Learning Research, 2010, 11: 2051.

Genetic Algorithm Based Multi-Objective Least Square Support Vector Machine for Simultaneous Determination of Multiple Components by Near Infrared Spectroscopy

XU Bing[1], WANG Xing[1,3], Dhaene Tom[2], SHI Xin-yuan[1*], Couckuyt Ivo[2], BAI Yan[3], QIAO Yan-jiang[1*]

1. Research Center of TCM Information Engineering, Beijing University of Chinese Medicine, Beijing 100029, China
2. Department of Information Technology, Ghent University- iMINDS, B-9050 Gent, Belgium
3. Henan College of Traditional Chinese Medicine, Zhengzhou 450008, China

Abstract The near infrared (NIR) spectrum contains a global signature of composition, and enables to predict different properties of the material. In the present paper, a genetic algorithm and an adaptive modeling technique were applied to build a multi-objective least square support vector machine (MLS-SVM), which was intended to simultaneously determine the concentrations of multiple components by NIR spectroscopy. Both the benchmark corn dataset and self-made Forsythia suspense dataset were used to test the proposed approach. Results show that a genetic algorithm combined with adaptive modeling allows to efficiently search the LS-SVM hyperparameter space. For the corn data, the performance of multi-objective LS-SVM was significantly better than models built with PLS1 and PLS2 algorithms. As for the Forsythia suspense data, the performance of multi-objective LS-SVM was equivalent to PLS1 and PLS2 models. In both datasets, the over-fitting phenomena were observed on RBFNN models. The single objective LS-SVM and MLS-SVM didn't show much difference, but the one-time modeling convenience allows the potential application of MLS-SVM to multicomponent NIR analysis.

Keywords Multi-objective least square support vector machine; Genetic algorithm; Near infrared; Multicomponent quantification; Adaptive modeling

* Corresponding author

(Received Apr. 16, 2013; accepted Jul. 15, 2013)