一种基于支持向量回归和动态特征选择的梨黑星病预测方法

辜丽川[1] 钟金琴[2] 张友华[1] 李绍稳[1]
(安徽农业大学信息与计算机学院 合肥 230036)[1] (安徽大学职业技术学院 合肥 230031)[2]

摘 要 当前作物病害预测方法存在时效性差、预测结果拟合度较低的问题。提出一种基于回归的预测方法框架,用SVM对数据向量特征进行约简,它可以重复选择密切相连的特征和构建可动态优化自身参数。用本方法对黄河故道地区砀山酥梨黑星病为例进行预测测试,与现有方法以及实测数据进行相关性统计分析。结果表明在对酥梨的黑星病预测上提出的方法,在拟合度、推理效率和准确率上具有显著的优势。

关键词 支持向量回归,特征选择,病害预测,梨黑星病
中图法分类号 TP338.6 **文献标识码** A

Method of Pre-decison on Pear Scab Based on SVR and Dynamical Feature Selection

GU Li-chuan[1] ZHONG Jin-qin[2] ZHANG You-hua[1] LI Shao-wen[1]
(School of Information and Computer, Anhui Agricultural University, Hefei 230036, China)[1]
(Vocational Technical College, Anhui University, Hefei 230011, China)[2]

Abstract At present, there are time consuming and poor effect of forecasting in the method of Pre-decison on fruit diseases. A new forecast method SVR-D1.1 based on regression was proposed in this paper, features reduction was conducted by using SVM. The method can select keen correlative features repeatedly and construct its dynamic optimizing parameters. Relativity statistical analysis was conducted between the real data and the forecasting data of Dangshansu pear scab, which show the method is more super and more valid than the current method in efficiency and precision of forecasting the occurrence tendency of Dangshansu pear scab. The experiment showed that the approach has obvious advantage on fitting degree, the reasoning efficiency and accuracy.

Keywords Support vector regression, Feature selection, Disease forecasting, Pear scab

1 引言

通过人工智能技术构建专家系统来模拟人类专家的方法是现有作物病害预测的主要途径之一。它一方面避免了传统专家经验评估法所存在的主观片面性,缺乏严格机理的弊端;另一方面,较大缓解了传统数学模型法所存在的建模因子有限、数据误差大等问题。专家系统预测法近几年来在国内外都有了很大的发展,所涉及的对象主要包括粮食作物、棉花、果树和草原病虫害等[1-4]。但目前,此类系统知识表达方法单一,多采用静态系统模型,是一种静态的系统,具有"知识获取"瓶颈以及知识库维护困难等缺陷。而在实际的农业生产中,由于病虫害的种类、抗药性及环境因子的动态变化使得病害危害发生的特点也在不断地变化,这就要求预测系统必须是动态的,能随着病害和相关预测因子的改变而不断更新。因此,结合作物病虫害危害发生的特点,研究出一种动态、开放、实用性高的病害预测方法是十分必要的。本文结合动态特征选择和支持向量回归优化模型构建了一种预测模型,将该模型用于梨黑心病的预测。实验表明该模型相比现有一些方法在时效性和准确率上都有较为显著的提高。

2 支持向量回归

支持向量机(Support Vector Machines, SVM)是基于统计学习理论框架下的一种处理非线性分类和非线性回归的有效方法。由于具有完备的理论基础和出色的学习性能,该方法已成为当前国际机器学习界的研究热点,能较好地解决小样本、高维数、非线性和局部极小点等实际问题。支持向量回归方法是从线性可分情况下的最优分类面发展而来的。所谓最优分类面就是要求分类面不但能将两类样本正确分开,而且使分类间隔最大。

设$(\vec{x}_1, y_1), \cdots, (\vec{x}_n, y_n)$训练数据集,$(x_1, y_1), \cdots, (x_l, y_l)$ $x_i \in R_n$, $y_i \in \{1, -1\}$, $i = 1, 2, \cdots, l$可以被一个最优分类超平面:

$$(\omega \cdot x) + b = 0 \quad x \in Rn, \omega \in Rn, b \in R$$

线性分开。要找到这个超平面,需要求解下面的二次规划问题:

$$\min \Psi(\omega) = 1/2(\omega \cdot \omega) \tag{1}$$

$$s.t.\ y_i[(\omega \cdot x_i) + b] \geqslant 1 \quad i = 1, 2, \cdots, l \tag{2}$$

利用拉格朗日优化方法求解上述问题的对偶问题,可得

到稿日期:2008-08-29 返修日期:2009-02-25 本文受国家自然科学基金(30800663,70871033),国家863高科技计划(2007AA04Z116),安徽省十一五科技攻关项目(8010302170),安徽省高等学校省级自然科学研究项目(KJ2008B111)资助。
辜丽川(1974-),男,博士生,讲师,主要研究方向为农业信息化、机器学习、决策支持,E-mail:gulichuan@ahau.edu.cn。

到最终的分类函数

$$f(x) = \mathrm{sgn}\{(\omega \cdot x) + b\} = \mathrm{sgn}\{\sum_{i=1}^{l} = a_i y_i (x_i \cdot x) + b\} \tag{3}$$

其中，a_i 为拉格朗日乘子；b 是分类阈值。

在非线性可分情形下，SVM 方法通过非线性变换 $\Phi(x)$ 将输入空间变换到一个高维空间，在这个空间中求（广义）最优分类面。此时，相应的分类函数为：

$$f(x) = \mathrm{sgn}\{\sum_{i=1}^{l} = a_i y_i K(x_i, x) + b\} \tag{4}$$

其中，表达式 $K(x_i, x)$ 相当于 $(\Phi(x_i) \cdot (x))$，这里被当作核心函数。通过这个核心函数可以求得原始回归问题的解，而不需要用 $\Phi(x)$ 进行数据的明确转换。

在 SVM 最优分类函数确定过程中起到关键作用的仅仅是一部分样本 SV。显然，当 SV 集可以充分描述整个训练数据集的特征时，对 SV 集的划分可以认为等价于对整个数据集的分割。因此，使用 SV 集取代训练样本集进行分类学习，在不影响分类精度的前提下约简训练样本量，进而减少训练时间，提高知识发现的速度。

3 基于 SVR 的预测模型框架

本节给出了一个建立基于 SVR 预测模型的模型框架，它包含了建立预测模型所必须处理的任务，如特征选择、模型构建和模型更新等关键问题。

3.1 基于 SVR 的预测建模框架

图 1 所示为本文提出的基于回归的预测方法建模框架 SVR-D1.1，它可以重复选择密切相连的特征和构建可动态优化自身参数。该建模框架包含了建立预测模型的 6 个步骤，具有一定的通用性，说明如下。

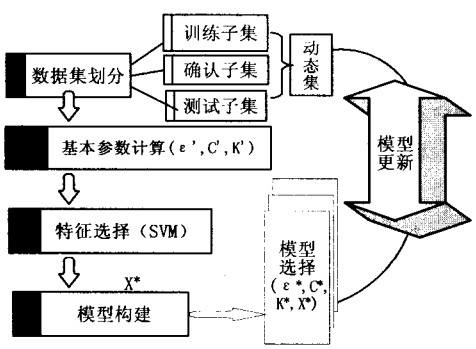

图 1 SVR-D1.1 预测建模框架

第一步（数据集划分） 这一步骤主要将数据集根据数据特征划分为训练子集、确认子集和测试子集。其中，练习数据用来构建模型，确认数据用于模型和特征的选择，测试数据是完全独立的子集，用作辅助估计现实模型的误差级别。更新模型时，这些子集可随时动态被重新定义。

第二步（基本参数计算） 主要是 ε'，C' 和核心参数 K' 的确定，一定条件下，这些参数将发挥作用。

第三步（特征的选择） 使用 SVM 方法进行特征的选择，约简不需要的特征，得到新的数据集向量[6]。

第四步（模型构建） 利用预报器 x^* 在基本参数中进行网格搜索，得到最佳的参数 ε^*，C^*，K^*。

第五步，基于核校准优化模型选择。根据上步中确定的参数 ε^*，C^*，核心函数 K^* 和预报器 x^* 确定预测模型。

第六步 SVR-D1.1 还定义了模型更新策略，可根据数据向量的特征动态更新选择情况进行模型的动态更新。

3.2 基于核校准的 SVR 优化模型选择

SVR 中的关键问题之一就是模型参数 ε' 和 C' 及把数据运用到更高维空间的核心函数的选择。如图 2 所示，SVR-D1.1 中模型构建的方法是通过计算 SVR 参数的初始值，然后对这些参数进行网格搜索来完成的。

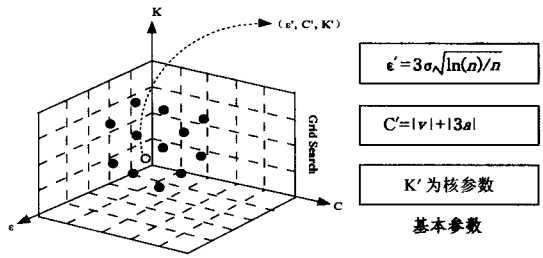

图 2 SVM 模型选择

对于 ε' 和 C' 的初始值，这里使用了由文献[7]中提出的经验规则。对于同一个 SVM 模式分类问题，一定存在一个最优的核函数使得测试错误率最小。因此，最优核函数的选择成为了建模最关键问题之一。确定最优的核函数，需要一种能够度量不同核函数之间差异的方法[8]。由于核函数之间的关系等价于核矩阵之间的关系，因此引入了核校准的概念来体现不同核函数之间的差异[9]，下面给出核校准的定义。

定义 1 设有样本集 $U = \{x_1, \cdots, x_n\}$，P_1 和 P_2，表示核函数 $\Phi(xi)$ 的 2 个核距阵。其相应内积表达式为：

$$\langle P_1, P_2 \rangle F = \sum_{i,j=1}^{l} P_1(x_i, x_j) P_2(x_i, x_j) \tag{5}$$

现采用一个标量来度量核函数 P_1 和 P_2 在样本集 U 上的差异，称其为核校准，记为 $\Pi(U, P_1, P_2)$，其计算公式为：

$$\Pi(U, P_1, P_2) = \frac{\langle P_1, P_2 \rangle}{\sqrt{\langle P_1, P_1 \rangle_F \langle P_2, P_2 \rangle_F}} \tag{6}$$

显然，作为一个度量值，核校准体现了不同核函数之间的差异关系。从几何角度来看，核校准可以看成是两个向量之间夹角的余弦值。本文即以此概念为基础，提出一种寻找最优核函数的算法。下面详细叙述该算法。

步骤一 构造一个关于参数和乘子的拉格朗日方程，如式(7)所示：

$$\mathrm{L}(\sigma, \alpha) = \mathrm{f}(\sigma) + \alpha \mathrm{Tc}(\sigma) = \mathrm{f}(\sigma) + \sum_{i=1}^{m} {}_{a}\mathrm{ic_i}(\acute{o}) \tag{7}$$

步骤二 定义函数，如式(8)所示：

$$\mathrm{f}(\sigma) = \frac{(\mathrm{K}, \mathrm{YY}^1)_F}{1\sqrt{(\mathrm{K}, \mathrm{K})_F}} \tag{8}$$

其中，l 为样本数量，K 为核矩阵，Y 为样本类标识集。

步骤三 定义函数 c(σ)，如式(9)所示：

$$\mathrm{c}(\sigma) = [(\mathrm{K}, \mathrm{K}) - 1 \leqslant 0; -\mathrm{K} \leqslant 0] \tag{9}$$

其中 K 为核距阵。

步骤四 构造二次规划子问题，如式(10)所示，以对式(7)进行二次估计。

$$\min 1/2 \Delta\sigma \mathrm{TH} \Delta\sigma + \mathrm{f}(\sigma) \mathrm{T} \Delta\sigma;$$
$$\text{s. t. } \nabla \mathrm{c_i}(\sigma)\ \mathrm{T} \Delta\sigma + \mathrm{c_i}(\sigma) \leqslant 0, \mathrm{i} = 1, \cdots, \mathrm{m} \tag{10}$$

其中 H 为正定的汉森矩阵。

步骤五 求解式(10)，并更新 H。

步骤六 重复步骤直到算法收敛到最优的 σ 值。

3.3 特征选择

合适的特征可以提高模型精度，减少建模计算时间。因此，特征选择方法对于预测建模来说极为重要[10]。

设有数据集向量 $X=(x_1,x_2,\cdots,x_m)$，分量 $x_j(j=1,2,\cdots,m)$的取值实际上体现了相应的特征 x_j 对样本 X'分类判定的贡献。如果 x_j 极小趋于 0，那么特征 x_j 对样本 X'的类别的判定几乎不起作用；此外，如果某个特征约简后，基于数据向量得到的分类函数准确率保持不变甚至有所提高，说明该特征对分类器的学习作用是可以忽略的。本文采用如下特征约简算法。

(1) 对训练样本集$\hat{W}$进行训练，得到一个形式如式(4)所示的SVM 分类器，其分类准确率记为 μ_1。

(2) j = 1。

(3) 如果$|x_j|/(\Sigma n_j=1|x_j|)>\varepsilon(\varepsilon$ 为预设的贡献率下界)，则保留 x_j对应的属性 x_j，转(6)；否则，转(4)。

(4)剔除属性 x_j，重新训练一个形如式(4)的支持向量机，相应的准确率记为 μ_2。如果 $\mu_1\leqslant\mu_2$，则保留该特征，转(6)；否则，转(5)。

(5) 剔除特征 x_j，得到新属性集 X′。

(6) 赋值 j = j + 1.

(7) 如果 j≤n，转(3)；如果 j=n+1，则结束，获得约简后的特征，新的数据集向量记为($\hat{W}'$)。

上述约简算法兼顾了特征的贡献和分类准确率变化这两个因素，只有当这两方面同时满足一定条件时(算法第 3 步和第 4 步)才能把该特征当作冗余特征剔除，从而最大限度地保留对于分类准确率有作用的特征属性。

3.4 特征的动态更新

对于一些具有动态变化特征的时序问题，因为与预测相关影响因子的变化，往往会导致所构建的预测模型受到影响，导致模型的精确性降低，乃至模型的失败。一个模型构建好以后，可能在一个时期内工作得很好，但在未来的某段时间内可能因影响因子的变化而提供有误差的预测结果。为了解决这一问题，在 SVR-D1. 1 中有一个模型更新模块。这里特征的动态更新的基本方法是分两步进行数据训练：一部分是历史数据集，另一部分是最新数据。当事先所定义的新采样数据到达时，新数据就会被并入到训练集中。新数据的特征模式会采用 3. 3 所述方法进行特征选择。最后，当进行模型更新时，随着数据点从练习集合的新数据部分到有效集合的转变，属于练习和有效集合的数据的比例一直保持不变。这样既保持了模型的运算量，也实现了模型的动态更新。

4 实验

把所提出的 SVR-D1. 1 应用于真实的梨黑心病预测问题。使用本文所提出的方法进行了如下实验。

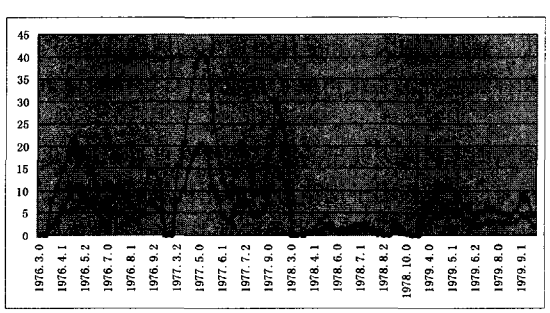

图 3 1976－1979 年病叶/果率时序

图 3 给出的是 1976－1979 年病叶率和病果率时序(图中纵坐标为病叶率和病果率，横坐标为时间，格式为年. 月. 旬，其中 0－上旬，1－中旬，2－下旬)。通过选择与梨黑心病爆发相关的初始特征，把 SVR-D1. 1 应用于病叶率与病果率 2 个时序。酥梨的黑心病相关影响因子有 6 个：旬平均气温、旬气温距平、旬降水日数、旬降水距平、病叶率、病果率，即数据集向量 $M=\{n,o,p,q,s,t\}$。预测目标向量 $R=\{s,t\}$，即预测病叶率、病果率。

结果我们为每个序列获得了一个描述各自模型的不同的参数集合。根据 3. 3 描述的特征选择方法，我们还为每个序列确定了已选特征集合。采取本文提出的利用核矩阵选择优化支持向量回归的算法，确定了核心函数，并为全部的初始特征进行了详细说明。试验发现，使用 SVR-D1. 1 方法和一个周期性索引以及一个指示月类型的二进制变量(有的是 4 周，有的是 5 周)就可以获得联系最紧密的特征。每年 3 月上旬到 10 月上旬是黄河故道砀山酥梨黑心病发病期。显然，可以看出每年 4 月发病高峰期的周期性的变化图案，这种现象同样出现在剩余的序列中。

为验证本文所提出的方法的效果，以 1976－1979 年黄河故道地区梨黑心病为测试数据集，将 1980－1983 年作为目标年份，使用 3 种预测方法对其各旬进行预测，以年为单位计算相关系数均值，并把预测值与使用神经网络实现的(NN-UP)以及 ARMAX 模型方法的预测值进行相关性统计分析。表 1、表 2 和表 3 所列为实验结果统计。

表 1、表 2 分别表示病叶率相关系数(CCPIL)和病果率相关系数(CCPIF)、均绝对比例误差(MAPE)的精确性误差测量结果。

表 1 3 种方法的系统相关系数误差

年	病叶率相关系数 CCPIL			病果率相关系数 CCPIF		
Year	SVR-D1. 1	ARMAX	NN-UP	SVR-D1. 1	ARMAX	NN-UP
1980	0. 751**	0. 701	0. 741	0. 691**	0. 600	0. 651
1981	0. 867**	0. 873	0. 807	0. 953**	0. 889	0. 803
1982	0. 807	0. 873**	0. 667	0. 807	0. 801**	0. 777
1983	0. 869**	0. 673	0. 731	0. 807	0. 787	0. 800**
1984	0. 874**	0. 873**	0. 737	0. 893	0. 893**	0. 889

表 2 均绝对误差比例

测试数据集(黄河故道砀山酥梨黑心病)			
	ARMAX	NN-UP	SVR-D1. 1
病叶率	10. 98	14. 31	11. 86
病果率	13. 64	14. 66	11. 44
平均误差	12. 31	14. 485	11. 65

表 3 各种方法的预测结果对比

测试数据集(黄河故道砀山酥梨黑心病)				
模型类别	模型准确率	命中率	覆盖率	提升系数
ARMAX	0. 8782	0. 7142	0. 1562	5. 3305
NN-UP	0. 8983	0. 7538	0. 3625	5. 6256
SVR-D1. 1	0. 9088	0. 8333	0. 4018	6. 2186

由表 1 不难看出，尽管在某个结果中出现了相互矛盾的不同的误差测量，但表 2 所列平均误差水平，表明 SVR-D1. 1 方法测量的结果比其他方法预测的结果优越。这证明了我们提出的方法可以提供更为准确的预测。表 3 所列为实验对比结果，显然本方法预测值与实测值在 5 年中的相关系数均达到极显著水平以上，说明预测值与实测值的动态曲线拟合程度是相当好的，采用本方法对酥梨黑星病预测是可行的。

结束语 本文提出了一种预测模型框架 SVR-D1. 1，该框架中使用 SVR 构建回归模型并进行动态特征选择。把所提出的方法应用于梨黑星病预测问题。将预测结果与标准的 ARMAX 方法以及 NN-UP 方法相比较。结果表明 SVR-D1. 1

(下转第 243 页)

	M1	−30665.53926	−30665.5391	−30665.53901
	M2	−30664.5	−30665.3	−30645.9
P2	M3	−30665.539	−30665.539	−30665.539
	M4	−30665.5	−30665.5	NA
	M5	−30665.539	−30665.539	−30665.539
	M1	−6961.78924	−6961.538290	−6961.084285
	M2	−6952.1	−6342.6	−5473.9
P3	M3	−6961.814	−6875.940	−6350.262
	M4	−6961.81	−6961.81	NA
	M5	−6961.814	−6961.284	−6952.482
	M1	−0.09582504	−0.09582504	−0.09582504
	M2	−0.0958250	−0.0891568	−0.0291438
P4	M3	−0.095825	−0.095825	−0.095825
	M4	−0.095825	−0.095825	NA
	M5	−0.095825	−0.095825	−0.095825
	M1	680.6315	680.634	680.6912
	M2	680.91	681.16	683.18
P5	M3	680.630	680.656	680.763
	M4	680.630	680.641	NA
	M5	680.632	680.643	680.719
	M1	0.74999918	0.74999929	0.74999960
	M2	0.75	0.95	0.75
P6	M3	0.750	0.750	0.750
	M4	0.75	0.75	NA
	M5	0.75	0.75	0.75

从表 2 可以看出，对于问题 P4，该算法在最优值方面优于 HM，平均值、最差值方面均优于其它算法。对于问题 P2，P3，P5，该算法在最优值方面与其它算法一样达到最优，在平均值、最差值方面明显优于 HM，与其它算法效果相同。对于问题 P1，P4，P6，该算法在最优值方面优于其它算法，在平均值、最差值方面略低于 SR，SMES，比其它的算法效果好。这些实验结果表明该算法是非常有效的。

结束语 提出了一种新的罚函数模型来处理约束优化问题。本文首先引入了函数值满意度函数与约束满意度函数，将这两个函数的乘积构造为适应度函数，并将反映种群质量信息参数的当前种群的可行解与不可行解的比值设置在适应度函数中，此模型不但能解决罚因子难以确定的问题，而且在进化过程中能有效地区分可行解与不可行解，从而使算法更加有效。

参 考 文 献

[1] Coello C A C. Constraint-handling using an evolutionary multiobjective optimization technique[J]. Civil Engineering and Environmental Systems, 2000, 17: 319-346

[2] Smith A E, Coit D W. Constraint handling techniques-Penalty functions[M]. Back T, Fogel D B, Michalewicz Z, eds. in Handbook of Evolutionary Computation, New York: Oxford Univ. Press and Inst. Physics, 1997

[3] Koziel S, Michalewicz Z. Evolutionary algorithms, homomorphous mappings, and constrained parameter optimization[J]. Evolutionary Computation, 1999, 7: 19-44

[4] Runarsson T P, Yao X. Stochastic ranking for constrained evolutionary optimization[J]. IEEE Trans. on Evolutionary Computation, 2000, 4: 284-294

[5] Hamida S B, Schoenauer M. ASCHEA: New results using adaptive segegational constraint handling[C]// Proc. Congr. Evolutionary Computation. vol, 1. May 2002: 884-889

[6] Mezura-Montes E, Coello C A. A Simple Multimembered Evolution Strategy to Solve Constrained Optimization Problems[J]. IEEE Transsactions on Evolutionary Computation, 2005, 9(1): 1-17

(上接第 217 页)

执行得较好且能选择出最重要的特征。SVR-D1.1 优点不仅局限于预测结果的拟合度好，而且随着测试环境相关因子的改变，该方法可以动态选择最相关特征和允许预测模型的调整，表现出周期性更新模型的能力。这大大提高了它解决实际事务的预测问题的潜力。

参 考 文 献

[1] Gu Lichuan, Ni Zhiwei, Li shw. Forecasting Method on Pear Scab Based on Fusion Reasoning[C]// CIS2007 Proceedings. Harbin, China, IEEE Press, 2007: 401-404

[2] 高灵旺，陈继光，等. 农业病虫害预测预报专家系统平台的开发[J]. 农业工程学报，2006(10)：154-158

[3] 刘莉，李绍稳，等. 模糊聚类在砀山酥梨黑星病预测专家系统中的应用[J]. 农业网络信息，2004(2)：12-14

[4] Xiong J T, Xiong F L, Tu R S. Case-based reasoning in forecasting insect pests[A]// Xiong F L, Lee J K, Mizoguchi R, eds. Proc. of PACES'95 [C]. Beijing: Publishing House of Electronics Industry, 1995: 627-630

[5] Vapnik V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995

[6] Niu Dongxiao, Li Jinchao, Li Jinying, et al. Daily load forecasting using support vector machine and case-based reasoning[C]// Second IEEE Conference on Industrial Electronics and Applications, ICIEA2007. 2007: 1271-1274

[7] Cherkassky V, Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression[J]. Neural Networks, 2004, 17(1): 113-126

[8] 刘向东，骆斌，陈兆乾. 支持向量机最优模型选择的研究[J]. 计算机研究与发展，2005，42(4)：576-581

[9] Muller K-R, Mika S, Ratsch G, et al. An introduction to kernel-based learning algorithms[J]. Neural Networks, IEEE Transactions, 2001, 12(2): 181-201

[10] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003(3): 1157-1182