文章编号:1001 - 9081(2006)04 - 0888 - 03

基于序列分解的复杂系统的时序预测方法

韩雪梅,徐从富,沈慧峰
(浙江大学 计算机科学与技术学院, 浙江 杭州 310027)
(zjuwendy@ yahoo. com. cn)

摘 要:现实中的时序数据,往往取自于复杂系统,表现出长记忆效应与短时不规则波动同时并存。传统的时序数据的分析和预测方法一般对不同层次的影响不加以区分,而是为其建立一个统一的模型,这使得在对复杂系统建模时需要用大量的参数予以表征,影响预测效率与精度。为此采用新的方法,将序列数据本身进行多平滑因子分解,对分解后的序列进行多尺度的采样并分别建模、预测,最后将结果整合。该方法应用于股票的实验表明,即使对起伏波动很大的时间序列,也能够得到较好的预测结果。

关键词:复杂系统; 时间序列预测; 多尺度采样; 序列分解
中图分类号: TP182 **文献标识码**:A

New time series forecasting approach for complex systems based on series decomposing

HAN Xue-mei, XU Cong-fu, SHEN Hui-feng
(*College of Computer Science*, *Zhejiang University*, *Hangzhou Zhejiang* 310027, *China*)

Abstract: Time series are often produced in complex systems which are controlled both by macroscopic level and microscopic level laws, with long memory effect and short-term irregularly fluctuations coexisting in the same series. Traditional analysis and forecasting methods didn't distinguish these multi-level influences and always made a single model for predication, which had to introduce a lot of parameters to describe the characteristics of complex systems and result in the loss of efficiency and accuracy. However, we decomposed time series into several ones with different smoothness, all the sub-time series were respectively modeled and predicated with multi-scale sampling. Then the forecasting results of sub-time series were composed to get the result of the original time series. The experiment results on the stock forecasting show that the method is efficient, even for the time series with large fluctuations.

Key words: complex systems; time series forecasting; multi-scale sampling; time series decomposing

0 引言

时间序列是常见的数据类型,对其的分析与预测常常关系到经济决策、灾害预防、疾病控制等诸多生产生活领域,是数据挖掘与知识发现的重要研究问题之一。

现实世界常见的时间序列数据,往往采自于各种复杂系统,如金融数据、气象数据等,它们受着宏观、中观、微观等不同层次因素的影响,其特点是具有大的偶然性与随机性。对复杂系统的结构剖析历来被统计学、经济学、数学等领域所关注。经济学普遍认为时间序列曲线,包含了长期的趋势、周期变化、季节性变化和不规则变化[1]。1951 年,文献[2]在对水文数据的研究中发现了时间序列所具有的长记忆性特征,第一次提出了时间序列长记忆性的问题。然后,人们发现长记忆现象在时序数据中是普遍存在的[3]。同时,在 1956 年,文献[4]提出了短范围相依过程的概念。短范围相依过程反应了时间序列的强混合性和短记忆的特点。

国内外研究者已经提出了众多时序预测的方法,如 Box-Jenkins[5],神经网络方法[6,7],遗传算法[8]和卡尔曼滤波方法[9]等。这些方法在对复杂系统的建模时往往建立一个模型,通过复杂参数来表征模型特征,对建模之前的预处理却鲜有研究。然而,对于复杂系统的序列数据,不同层次的外界因素造成系统既存在长期稳定的趋势,又有短期波动,用一个模型来刻画往往在效率与精度上难以两全:丢弃大量对预测时刻有影响的历史数据进行预测,会降低预测精度;而将历史数据赋予同样的权重去处理,又会增加计算时间,降低预测效率。

本文研究的时序数据受到宏观、中观、微观多层次因素约束,在分析与预测时也相应地采用多层次、多尺度的建模思想与方法。将时间序列分解为多个序列的"加和",对各个序列进行不同频度的采样,分别建立预测模型,并以金融数据为例进行了分析与预测。实验表明,该方法性能优越,对复杂系统拟合得较好,具有较高的预测精度。

1 序列预处理

复杂系统的观测数据往往受到从宏观到微观不同层次的因素约束,这些因素对序列波动的影响时间长短不同。有些因素会影响序列漫长的波动,例如股票的内在价值机制是股

收稿日期:2005 - 10 - 10;**修订日期**:2005 - 12 - 13
基金项目:国家自然科学基金资助项目(60402010);航天基金资助项目(No. 2003-HT-ZJDX-13)
作者简介:韩雪梅(1978 -),女,硕士研究生,主要研究方向:数据挖掘; 徐从富(1969 -),男,副教授,博士,主要研究方向:人工智能、数据挖掘、数据融合; 沈慧峰(1978 -),男,硕士研究生,主要研究方向:数据挖掘,信息融合.

票波动的决定力量,股票长期的价格趋势决定于股票的价值;有些影响则只是短期干扰,作用时间短暂,如股票市场中人的心理因素的作用,表现为股价的短期不规则波动。这些决定了对复杂系统建模,既要关注长期趋势,又要对中、短期影响作出判断,用一个模型很难描述多尺度的变化趋势。

本文提出基于多层次数据分解与建模的复杂系统时序预测方法。通过多平滑因子分解,序列拆解成为多个平滑程度不同,周期不同的新序列,新序列比原序列的形态更简单,易于建模。通过多尺度采样方法,对拆解序列采用不同采样频率,使得在对计算精度影响不大的前提下,减少了计算量。

1.1　相关定义

定义 1　标准采样周期

指系统观测数据的原始记录或统计间隔,记为 δ。本文假定对特定的系统,δ 的值是恒定的,其他的采样周期都是标准采样周期的整数倍。

定义 2　$\oplus$ 运算与分序列

设 T 为某个时间集,$T = \{t_1, t_2, \cdots, t_n, n \in \boldsymbol{N}\}$,并且 $t_n - t_{n-1} = \delta$。对时间序列 $X = \{x_t, t \in T\}$,定义 X 的 $\oplus$ 运算:

$$X = X(1) \oplus X(2) \oplus \cdots \oplus X(m) \tag{1}$$

其中,$X(j)(1 \leqslant j \leqslant m)$ 在本文中称为分序列,与 X 序列等势,同序号的元素采样时刻一致,且对任意 $t \in T$,分序列 $X(j)$ 中 t 时刻的元素 $x(j)_t$ 满足:

$$\sum_{j=1}^{m} x(j)_t = x_t \quad 1 \leqslant j \leqslant m$$

1.2　多平滑因子拆解

将原始序列表示成式(1)的分序列的组合,并使分序列更容易分析与建模。采用多平滑因子分解的方法,将时间序列 L 拆解成多个序列,步骤如下:

1) 设定平滑因子序列 $L = \{l_1, l_2, l_3, \cdots, l_{m-1}\}$,序列 L 中元素的值单调递减。例如,可以将平滑序列取为以 2 为底的指数序列,如 $\{2^6, 2^4, 2^2, 2^1\}$;对股票数据,可取 $\{60, 30, 15, \cdots\}$ 等。对具体领域,可根据经验或反复实验构建 L 序列。

2) 求序列 $X(1)$ 中元素 $x(1)_t, t_1 \leqslant t \leqslant t_n$,使用平滑因子 l_1:

$$x(1)_{t_p} = \begin{cases} \dfrac{1}{p}\sum_{q=1}^{p} x(t_q) & p \leqslant l_1 + 1 \\ \dfrac{1}{l_1}\sum_{q=p-l_1}^{p} x(t_q) & p > l_1 + 1 \end{cases} \tag{2}$$

3) 求序列 $X(j)$ 中元素 $x(j)_t$。序列 X 中的元素减去 $X(j-1)$ 中等序号的元素,组成新的序列 $X'(j-1)$,令序列 $X = X'(j-1)$,取平滑因子 l_j:

$$x(j)_{t_p} = \begin{cases} \dfrac{1}{p}\sum_{q=1}^{p} x(t_q) & p \leqslant l_j + 1 \\ \dfrac{1}{l_j}\sum_{q=p-l_j}^{p} x(t_q) & p > l_j + 1 \end{cases} \tag{3}$$

4) 序列 X 中的元素减去 $X(m-1)$ 中等序号的元素,即得序列 $X(m)$。

经过上述变换,原序列被拆解成多个分序列的组合,拆解的序列数 m 可由经验决定,通常可取 3 ~ 5。随着 j 的增加,平滑因子逐渐递减,分序列曲线依次由平滑到粗糙变化。

1.3　分序列多尺度采样

一个典型的经过拆解得到的各个分序列 $X(j)$ 如图 1 所示,随着平滑因子 L 的逐渐减小,能够清楚地看出序列由平滑变得粗糙,并且,其波动周期逐渐缩短。对于较为平滑的序列,由于其表现了长期趋势的延伸,在建模的过程中需要考虑到比较长的历史;但相对于粗糙的短期波动,平滑序列上各点承载的信息更少,如果对每个新序列都用标准采样周期进行建模和预测,会大大增加不必要的计算量。近来在序列模式挖掘的处理中已有人对变频采样进行探讨[10]。其处理方式是针对同一个序列,在采样过程中采用近密远疏的方法,虽然其思路十分直观,但由于采样频度的变化使得后期处理变得复杂。

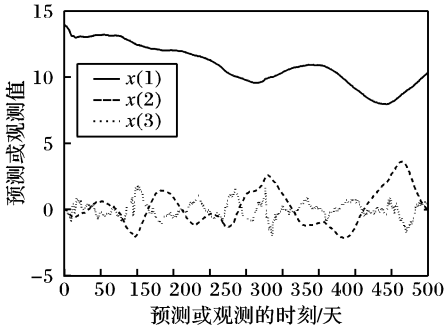

图 1　序列拆解曲线示例

本文给出了一种新的多尺度采样方式,即对拆解之后的平滑序列,根据平滑程度采用稀疏的采样的方式,而对相对粗糙的序列,依照其粗糙的程度逐渐增加采样频度。设采样周期序列为 $S = < s_1, s_2, \cdots, s_m >$,其中,$s_j (1 \leqslant j \leqslant m)$ 表示序列 $X(j)$ 的采样周期且 $s_j < s_{j-1}$,s_j 为标准采样周期 δ 的正整数倍。为便于计算,应该使 $s_{j-1}/s_j \in \boldsymbol{N}$,亦即 s_{j-1} 是 s_j 的正整数倍。序列 $X(j)$ 的采样序列记为 $\dot{X}(j)$。

2　建模与预测过程描述

2.1　新数据加入

时间序列的分析与预测是一个在线的过程,新数据在预测过程中将会不断的加入进来。每一个新读入的数据,通过 1.2 的拆解公式进行拆分。各个分序列再将拆分结果以不同的采样频率加入进来。算法如下所示:

```
BEGIN
  s1, s2, …, sm;                          //设定不同的采样周期
  cs(1) = cs(2) = … = cs(m) = 0;
          /* 不同的采样周期曲线进入点计数初始化为 0 */
  WHILE TRUE
  BEGIN
    READ xt;
    Decompose x(t) = Σ_{i=1}^{m} x(j)t;      //将新加入的点拆解
    FOR i = 1 to m
      cs(i) ++;
      IF cs(i) mod si = 0
      将 x(i)t 加入到第 i 个序列中;
    END FOR
  END
END
```

2.2　建立预测模型

对分序列数据,已有的方法如演化算法、神经网络方法都可以用来进行建模和预测。既可以对各个分序列采用不同的模型;也可以采用统一模型,但各个分序列分别使用不同的参数。本文以更具有普适性的 Box-Jenkins 方法为例描述建模及预测过程。

在一定条件下,平稳的均值为零的时间序列 $y(k)$ 可看做

是以白噪声为输入的线性定常随机系统的响应。这时 $y(k)$ 满足如下的差分方程：

$$y_k = \sum_{i=1}^{p} \phi_i y_{k-i} + \varepsilon_k - \sum_{j=1}^{q} \theta_j \varepsilon_{k-j} \quad (4)$$

式(4) 记做 ARMA(p,q) 模型[5]。其中 $\sum_{i=1}^{p} \phi_i y_{k-i}$ 为 $k-1$ 到 $k-p$ 时刻的线性加权和，$\varepsilon_k - \sum_{j=1}^{q} \theta_j \varepsilon_{k-j}$ 表示 q 个白噪声的线性加权和。通过最小二乘方法或极大似然估计法可以得出 $\phi_1, \phi_2, \cdots, \phi_p, \theta_1, \theta_2, \cdots, \theta_q$ 和 σ_ε^2 共 $p+q+1$ 个参数的估计，并依公式(4) 进行预测。

设序列 $\dot{X}(j)$ 已有 H 个样本点，对序列 $X(j)$ 的采样后的序列 $\dot{X}(j)$ 建模步骤如下：

1) 选择 h 个最近的样本点，$h \leqslant H$。判断样本序列的平稳性，对非平稳序列通过一阶或多阶差分的方法将其平稳化。

2) 对序列 $\dot{X}(j)$ 建立适合的 ARMA 模型，模型的合理性应通过 Box-Pierce 检验[5]。

3) 通过公式(4) 进行下一时刻的预测。

4) 新数据加入，差分后与步骤 3) 的预测值进行比较，计算预测残差。

由于复杂系统的时变性，初始建立的模型参数可能不适用于继续进行预测。当预测模型的残差不能通过检验时，选择最近 h 个样本点，重复步骤 1) 和 2)，修正预测模型。

2.3 预测结果整合

至此得到各个分序列的预测结果。由于 $\dot{X}(j)$ 采样周期不同，随着 j 的增加，采样频率逐渐增加，预测周期逐渐缩短。这时，通过观察有较长采样周期的平滑曲线的预测结果，可以了解到序列未来较长时间的可能的发展趋势。但对于短期的精确预测，需要将各个分序列预测结果进行整合。

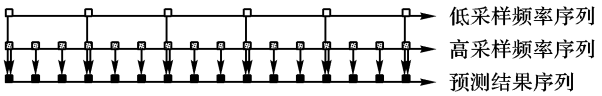

图2 多种采样频率序列与预测结果关系

如图 2 所示，最终的预测结果需要将低频率采样序列和高频率采样序列加和。设要预测时刻的值，由公式(1)，需要分别计算分序列中 τ 的值。因采样不同步，并非每一个 τ 都可以通过预测模型直接得到，但可以得到离 τ 最近的观测值和预测值。由于这些点位于平滑的分序列上，对它们的处理，可以通过简单的插值得到。至此得到最终的预测结果。

3 实验设计及结果分析

本文选取股票市场数据对方法进行验证。

图 3 ~ 图 5 为某银行某股票数据预测实例，选择共计 530 天的股票收盘价进行建模与预测。平滑序列 L 取{60,15}，将原始序列分解为三个序列，如图 3 所示，(a) 是原始序列曲线，其他子图是分解序列曲线。

序列分解之后，根据各个曲线的平滑程度，采样序列设置为{10,5,1}。结果如图 4 所示。

图 5 为该股票预测的相对误差拟合曲线，其中 92.83% 的相对预测误差小于 3%，相对预测误差的平均值为 0.0105。与使用单一参数的 ARMA 模型、神经网络、遗传算法的进行股票预测的方法相比[12]，本文的方法在预测精度上有了较大的提高。

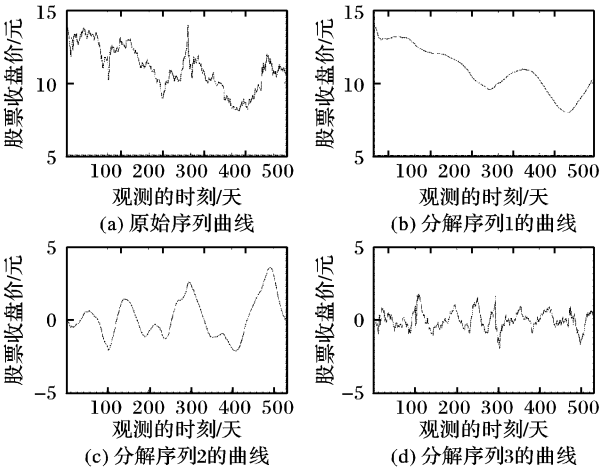

图3 原始序列与各分序列曲线

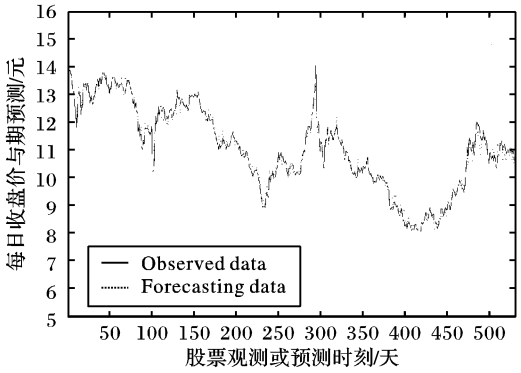

图4 民生银行股票数据的拟合与预测曲线

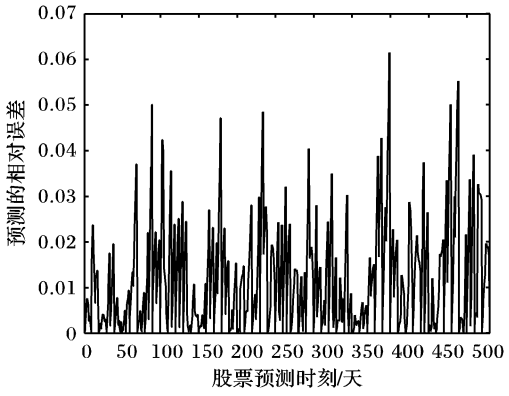

图5 相对误差拟合曲线

4 结语

本文提出了基于多层次数据分解与建模的复杂系统的时序预测方法。实验证明，该方法对复杂系统的建模与预测有较好的效果，即使是诸如股票一类的复杂系统，也能够较好的拟合和预测结果。

参考文献：

[1] BOWERMAN BL, O'CONNELL RT. Forecasting and Time Series: An Applied Approach, Third Edition[M]. Florence KY: Thomson Learning, 1993. 3 - 8.

[2] HURST HE. Long-term storage capacity of reservoirs[J]. Transactions of the American Society of Civil Engineers, 1951, 116: 770 - 808.

(下转第 894 页)

(tell: language KQML: ontology rules of learning: in-reply-to (struct s1): content(val(input-layer 3) (hide-layer 5) (out-layer 2)))

(tell: language KQML: ontology rules of learning: in-reply-to (struct s2): content (val (collection in-sample) (collection out-sample) (collection ref-sample) (collection error-control)))

3)Agent A 根据双方认可的适应度函数和学习样本计算学习规则,得出最优规则 r,则:

(ask-if: language KQML: ontology rules of learning: reply-with m1: content(val(rules r))

Agent B 可以接收和否定,如果接受,则返回一个 tell 消息,如果否定的话就返回一个 untell 消息,Agent 接受后,也可以用 insert 原语把该规则插入自己的知识库。

6 结语

针对模型库系统提出了一个全新的模型表示与复合方法,即基于 Agent 的模型表示与复合。通过引入 Agent 理论与技术,解决了传统模型库系统在自适应和自学习方面的局限,而且把模型的复合转换成一个 Agent 之间的协作问题求解,这样可以通过越来越成熟的 Agent 协作理论与技术来实现模型库系统中最复杂的模型表示与模型复合技术。另外,提出了一个全新的模型定义,充分考虑了模型应该具备知识本体的特征,也为如何定义模型与模型库提供了一个全新的视野,当然,本文所引用的一些理论和知识是否能够形成一个统一严密的体系还需要进一步的完善和应用实践的不断检验。

下一步的工作主要有:1)如何把模型表示与传统的模型论紧密结合起来,在形式化方面提供一个严密和完整的描述;2)模型库全局知识维护的工程化实现技术;3)基于 Agent 的模型库系统架构与实现。

参考文献:

[1] 张维明,刘忠,肖卫东,等. 信息系统建模[M]. 北京:电子工业出版社, 2002.

[2] 黄梯云. 智能决策支持系统[M]. 北京:电子工业出版社, 2001.

[3] 陈文伟. 智能决策技术[M]. 北京:电子工业出版社,1998.

[4] 张文修, 梁怡, 吴伟志. 信息系统与知识发现[M]. 北京:科学出版社, 2003.

[5] BLANNNING RW. An entity-relationship approach to model management[J]. Decision Support Systems, 1986, 2(1):65 - 72.

[6] BASU A, BLANNING RW. Enterprise modeling using metagraphs [M]. Decision Support Systems: Experiences and Expectations, 1992. 183 - 200.

[7] LENARD. An object-oriented approach to model management[J]. Decision Support Systems, 1993, 9(1): 67 - 73.

[8] MA J. An Object-Oriented framework for model management[J]. Decision Support Systems, 1995, 13(2):133 - 139.

[9] 李京, 孙颖博, 刘智深,等. 模型库管理系统的设计与实现[J]. 软件学报, 1998,9(8): 613 - 618.

[10] 周宽久, 黄梯云. 面向对象的模型表示与模型复合[J]. 哈尔滨工业大学学报,1997,29(4):18 - 20.

[11] DUTTA A, BASU A. An artificial approach to model management in decision support systems[J]. IEEE computer, 1984, 17(9):89 - 97.

[12] GROFFRION AM. The SML language for structured modeling[J]. Operation Research, 1992, 40(1): 38 - 75.

[13] LENARD. Representing models as data[A]. Proceedings of the Nineteenth Annual Hawaii International Conference on System Science[C]. 1986. 389 - 396.

[14] LEE K-W, HUH S-Y. Financial model-base construction for flexible model manipulation of models and solvers[A]. Annual Hawaii International Conference on System Sciences[C]. 2003. 84 - 91.

[15] MA J. Type and inheritance theory for model management[J]. Decision Support Systems, 1997, 19(1): 53 - 60.

[16] 王保江, 怀进鹏, 夏万强. 基于构件的模型库和方法库的设计和实现[J]. 北京航空航天大学学报. 1998, 24(4): 418 - 421.

[17] LI M, PENG H. Research on building and management of model base of decision Support System[A]. IEEE Proceedings of WCICA 2004[C]. 2004. 1895 - 1899.

[18] 周永林,潘云鹤. 面向 Agent 分析与建模[J]. 计算机研究与发展. 1999, 36(4): 410 - 416.

[19] DAM KH, WINIKOFF M. Comparing Agent-oriented methodologies [J]. Lecture Notes in Computer Science, 2005, 3382(1): 62 - 72.

[20] GUARION N, GIARETTA P. Ontology: a Knowledge Bases towards a terminological clarification[M]. Towards very large Knowledge Bases-Knowledge Building and Knowledge Sharing, ISO Press, 1995. 25 - 32.

[21] DRAFT Specification of the KQML Agent - Communication Language[EB/OL]. http://www.cs.umbc.edu/kqml/papers/kqml-spec.ps, 1993 - 06 - 15.

[22] 朱剑英. 智能系统非经典数学方法[M]. 武汉:华中科技大学出版社, 2001. 311 - 316.

(上接第 890 页)

[3] 张世英, 樊智等. 协整理论与波动模型[M]. 北京: 清华大学出版社, 2004.

[4] ROSENBLATT M. A central limit theorem and a strong mixing condition[A]. Proceedings of the National Academy of Sciences[C]. 1956. 43 - 47.

[5] SHAUN - INN WU, RUEY - PYNG LU. Combining artificial neural networks and statistics for stock-market forecasting[A]. Proceedings of ACM conference on Computer science[C]. 1993. 257 - 264.

[6] SAAD EW, PROKHOROV DV, WUNSCH DC II. Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks[J]. IEEE Transactions on Neural Networks, 1998, 9(6):1456 - 1470.

[7] LEE RST. iJADE stock advisor: an intelligent Agent based stock prediction system using hybrid RBF recurrent network[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2004, 34(3):421 - 428.

[8] HITOSHI IBA, TAKASHI SASAKI. Using genetic programming to predict financial data[A]. IEEE Proceedings of Congress on Evolutionary Computation[C]. 1999. 244 - 251.

[9] MCGONIGAL D, IONESCU D. An outline for a Kalman filter and recursive parameter estimation approach applied to stock market forecasting[A]. IEEE Conference on Electrical and Computer Engineering[C]. 1995. 1148 - 1151.

[10] PALPANAS T, VLACHOS M, KEOGH EJ, *et al*. Online Amnesic Approximation of Streaming Time Series[A]. IEEE Conference on ICDE[C]. 2004. 338 - 349.

[11] BAR-SHALOM Y, CHEN H, MALLICK M. One-Step Solution for the Multistep Out-of-Sequence-Measurement Problem in Tracking [J]. IEEE Transactions on Aerospace and Electronic Systems, 2004, 40(1): 27 - 37.

[12] BAO R. A study of non-periodic short-term random walk forecasting based on RBFNN, ARMA, or SVR-GM(1,1|/spl tau/) approach[A]. IEEE Conference on Systems, Man and Cybernetics [C]. 2003. 254 - 259.