DOI: 10.3969/j.issn.1671-9727.2017.02.15 [文章编号] 1671-9727(2017)02-0247-06

基于 SVM 分类的边缘提取算法

张 萍, 王 琳, 游 星
(成都理工大学 管理科学学院, 成都 610059)

[摘要] 通过分析同类数据点在空间中的几何形态,从数据点集所构成几何形态的凹凸性着手,提出边界提取算法并对高维数据进行分类。针对现实生活中的高维数据,利用局部线性嵌入将数据进行降维处理,得到低维特征数据。在此基础上,对于单分类数据集,用数据集表面的点的近邻样本与过该点的切平面之间的关系寻找边界点;对于多分类数据集,利用贝叶斯后验概率来寻找边界重复的点,以此更快达到提取边界点的目的。由此可以粗略筛选出边界点。为去除不重要的边界点,降低分类误差,通过构造最优超平面和支持向量机对边界点赋予权重,并设置阈值去除不重要的边界点,由此达到用较少的边界点准确分类数据的目的。通过 100 个测试样本进行分类测试并计算其分类准确率,验证了此分类方法的可行性。

[关键词] 局部线性嵌入;近邻样本;贝叶斯后验概率;支持向量;边界提取算法
[分类号] O235 [文献标志码] A

Edge extraction algorithm based on SVM classification

ZHANG Ping, WANG Lin, YOU Xing
College of Management Science, Chengdu University of Technology, Chengdu 610059, China

Abstract: Based on the analysis on geometry of same data space and its concave-convex shape, the boundary extraction algorithm is proposed in order to classify high-dimensional data. Locally linear embedding is used to reduce high-dimensional data in real life into low dimension. For a single classification data set, relation between the neighboring points on the surface of data set with its tangent plane is used to determine the boundary point. For multi-classification data set, those overlap points are determined through Bayesian posterior probability so as to extract boundary point quickly. In this way, the boundary point is roughly dressed by screening. In order to remove unimportance boundary point and reduce classification error, weight is given to these boundary points by constructing optimal hyper-sphere and support vector machine. At the same time, a threshold is set to remove unimportance boundary point so as to obtain less boundary point data for the purpose of accurate classification. Finally, 100 test samples are classified through this way and its feasibility is verified by calculating its classification accuracy.

Key words: locally linear embedding; neighboring points; Bayesian posterior probability; support vector machine; boundary extraction algorithm

[收稿日期] 2016-03-24。
[基金项目] 四川省社科规划项目(SC16TJ004)。
[第一作者] 张萍(1961—),女,教授,主要从事统计分析研究, E-mail:297141376@qq.com。

支持向量机(support vector machine,简称SVM)通常在解决小样本分类、非线性规划问题以及高维空间模式识别等问题中具有明显优势,在实际运用过程中通常用于数据分类和回归。支持向量机在理论上是利用VC理论寻找带约束条件的最小化经验风险解,在实际运用中通常是在简单的二分类模型基础上衍生的,其本质是通过在特征空间上的间隔最大化来建立超平面,从而达到分类的目的[1]。在近几年的研究中,研究者往往通过改进目标函数和引入结构知识等方法将SVM与其他算法相结合,例如:从最小平方化指标意义上考虑分类问题的等式约束[2],由此提出了最小二乘支持向量机。由于在进行空间处理的过程中可能使结果过于稀疏,因此提出1-norm支持向量机[3]、基于贝叶斯学习以及样本之间的相关性支持向量机[4-6]。

通常,生活中所产生的数据具有一定的规律,因此映射到特征空间的数据会构成不同的几何体,这些几何体总的可以从凹凸2个方面来考虑,几何体的表面的样本点构成支持向量机超球面。由此从几何空间的角度来看,数据分类只需要考虑几何体表面的样本点[7]。因此,本文在支持向量机理论的基础上,从样本点特征空间的几何形态的角度来进行数据分类。在模式识别中,对图片的识别过程中往往会采取从像素样本点中提取几何体的边缘[8],本文利用边缘提取样本的思想进行数据分析。

1 簇边缘提取算法

1.1 单类数据边缘提取

在数据挖掘中数据分类是最基本也是最主要的分析方式,研究者常用的方法有根据样本点之间的距离分类、决策树分类、基于先验理论的贝叶斯分类以及粗糙集分类等。这些方法大都是从技术角度和经验理论两方面着手考虑,但由于同一类数据样本在空间分散程度不一样,当绝大部分数据样本紧凑地聚集在一起时往往呈现出凹凸形状的几何体,每一个几何体就是一类数据样本,因此找到几何体的边缘就是数据分类的关键。本文首先从数据样本空间形态入手,分析提取几何体的边缘点集,从而粗略估计数据样本的分类情况。

当几何体表面呈现凸状时,边缘点所有的近邻点应该在过边缘点切平面的同一边,例如最典型的凸状几何体——球体;当几何体表面呈现凹状时,绝大部分近邻点在切平面的同一边,只有极少部分点在切平面的另一边[9-10]。因此,所有边缘点的共同特征是其绝大部分点都在其切面的同一边。本文的最终目的是数据分类,通过数据样本寻找切平面是非常困难的,而刻画切平面的唯一标准是切平面的法向量,因此可以利用 k 近邻样本来计算近似法向量,并且法向量的方向是指向数据聚集的方向[11]。

设原始数据样本有 n 个,分别表示为 $x_i(i=1,2,\cdots,n)$,则 x_i 的 k 个近邻样本为 $x_{ij}(j=1,2,\cdots,k)$,由此可得

$$\vec{V}_{ij} = x_{ij} - x_i \tag{1}$$

标准化可得

$$\vec{V}'_{ij} = \frac{\vec{V}_{ij}}{|\vec{V}_{ij}|} \tag{2}$$

由此根据(1)式和(2)式过点 x_i 的切平面的法向量可估计为

$$\vec{N}_i = \sum_{j-1}^{k} \vec{V}'_{ij} = \sum_{j-1}^{k} \frac{x_{ij} - x_i}{|x_{ij} - x_i|}$$

由于法向量的方向是指向数据集内部的,根据 k 近邻样本的分布特点可知:当法向量 $\vec{N}_i$ 与向量 $\vec{V}_{ij}$ 之间的夹角在$[0, \pi/2]$时,x_{ij} 为几何体内部点;当法向量 $\vec{N}_i$ 与向量 $\vec{V}_{ij}$ 之间的夹角为钝角时,x_{ij} 为几何体外部点。由此可知

$$\theta_{ij} = \vec{N}_i^T \cdot \vec{V}_{ij}$$

当 $\theta_{ij} \geqslant 0$ 时,设点 x_i 的近邻点有 l 个,为判断 x_i 是否为边缘点,需要计算满足条件的点在 x_i 的所有近邻点中所占比例。本文通过设置阈值 γ 来说明 x_i 是否为边缘点。由此可知

$$l_i = \frac{l}{k}$$

当 $l_i \geqslant 1-\gamma$ 时,则 x_i 为边缘点。

1.2 多类数据边缘提取

在1.1中分析了根据数据空间分布形成几何体来提取数据边缘,从而达到基本分类的目的。但实际上当2类数据出现边缘重复的情况下,用单纯边缘提取的方式就会使分类误差变得很大,因此要使数据得到准确分类,就需要对边缘重复的点做进一步处理。本文利用贝叶斯后验理论对样本点进行判断。

假设所有的数据分为 m 类,其类别表示为 W_m,则贝叶斯后验理论可表示为 $P(W_m|x_i)$,其

表示点 x_i 的所有近邻点与 x_i 同类的概率。由此可得点 x_i 是重叠边缘点的条件可表示为

$$\begin{cases} n_m > 1 \\ \dfrac{1}{n_m} \leqslant \max P(W_m \mid x_i) \leqslant \dfrac{1}{n_m} + \lambda \end{cases}$$

其中：n_m 为 $P(W_m \mid x_i) \neq 0$ 时点 x_i 的个数；λ 为阈值，一般可取 0.25。否则当点 x_i 满足

$$\max P(W_m \mid x_i) > \frac{1}{m} + \lambda$$

点 x_i 可能为单类边缘点，因此可利用 1.1 的方法继续进行判断。

2　SVM 构造超球面

由第 1 部分简单对离散数据点进行边缘提取，可以粗略在空间图像上判断数据样本的分类[12-13]，但并没有形成具体的理论算法。因此本文将第 1 部分提取出来的边缘点构成 m 个点集，则第 m 类边缘点集表示为

$$Z_m = \{x_{m1}, x_{m2}, \cdots, x_{mt}\} \subseteq X$$

其中 t 表示第 m 类边缘点的个数。点集 Z_m 包含了特征空间中构造超球体所需的绝大部分信息，因此，本文将边缘点集作为训练样本集。在特征空间中，训练样本 x 与超球体中心 α 具有相同的距离，则

$$f(x_{m1}) = f(x_{m2}) = \cdots = f(x_{mt}) = R^2(x) \tag{3}$$

根据支持向量机理论，本文采用高斯函数作为核函数[14-15]，则有

$$\begin{aligned} R^2(x) &= \| \phi(x) - \alpha \|^2 \\ &= K(x,x) - 2\sum_{j-1}^{t} \beta_j K(x_{mj}, x) + \\ &\quad \sum_{i,j-1}^{t} \beta_i \beta_j K(x_{mi}, x_{mj}) \\ &= 1 - 2\sum_{j-1}^{t} \beta_j K(x_{mj}, x) + \\ &\quad \sum_{i,j-1}^{t} \beta_i \beta_j K(x_{mi}, x_{mj}) \end{aligned} \tag{4}$$

又因为本文选择高斯函数作为核函数，所以对每一个 $f(x_{mj})$ 都有相同的 $\sum_{i,j-1}^{t} \beta_i \beta_j K(x_{mi}, x_{mj})$，因此，(3)和(4)式可以整理化简为

$$\begin{aligned} &\min Q\beta = \theta \\ &\text{s.t.} \sum_{j-1}^{m} \beta_j = 1 \\ &\quad \forall \ j \in [1,m], 0 \leqslant \beta_j \leqslant 1 \end{aligned} \tag{5}$$

令 $\beta = [\beta_1, \beta_2, \cdots, \beta_m]^T$，$\theta = [0,0,\cdots,0]^T$，$Q = [Q_1, Q_2, \cdots, Q_{m-1}]^T$，其中

$$\begin{aligned} Q_j = [&1 - K(x_{m1}, x_{mj+1}), K(x_{m2}, x_{m1}) - \\ &K(x_{m2}, x_{mj+1}), \cdots, K(x_{mt}, x_{m1}) - \\ &K(x_{mt}, x_{mj+1})] \end{aligned}$$

(5)式存在等式约束，由此求解比较困难。为准确方便地求解系数 β，本文将此模型转化为二次规划问题并利用内点法求解[16]。

由此根据线性表达式有

$$Q\beta = \begin{bmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_{m-1} \end{bmatrix} \times \beta = \begin{bmatrix} Q_1\beta \\ Q_2\beta \\ \vdots \\ Q_{m-1}\beta \end{bmatrix} = \begin{bmatrix} (Q_1\beta)^2 \\ (Q_2\beta)^2 \\ \vdots \\ (Q_{m-1}\beta)^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{6}$$

当(6)式中每一个元素都为 0 或者接近于 0 时，每一个元素相加应该非常接近于 0，由此可得

$$\sum_{j}^{m-1} (Q_j\beta)^2 = \begin{bmatrix} Q_1\beta \\ Q_2\beta \\ \vdots \\ Q_{m-1}\beta \end{bmatrix}^T \times \begin{bmatrix} Q_1\beta \\ Q_2\beta \\ \vdots \\ Q_{m-1}\beta \end{bmatrix} = 0 \tag{7}$$

将和函数带入(7)式，整理可得

$$\begin{aligned} &\min \beta^T H \beta \\ &\text{s.t.} \sum_{j-1}^{m} \beta_j = 1 \\ &\quad \forall \ j \in [1,m], 0 \leqslant \beta_j \leqslant 1 \end{aligned} \tag{8}$$

其中：$H = Q^TQ$ 为 $R^{m \times m}$ 的 Hessian 矩阵。根据凸内点算法可以求解上式，得到系数 β。由于 $\beta = [\beta_1, \beta_2, \cdots, \beta_m]^T$ 中各个元素实际上可以看作点的权重系数，同时构成超球体的点越多，越能准确描述超球体轮廓；但对构造超球体函数来说容易出现过拟合现象[1]，因此本文设置了一个阈值 β_0 来控制过拟合现象，当 β_j 小于阈值时，移除对应点。根据支持向量聚类原理，并通过大量实验表明，对于具有 n 个样本的数据集来说，通常将阈值设置为 n 的长度所对应的幂级数。

3　数据分类实例

为验证上述算法，本文利用现实生活中的香水数据，共有 5 个属性指标，200 个训练数据样本，100 个测试样本，通过局部线性嵌入将数据进行降维处理[17-18]。利用 MATLAB 软件编程，通过对上述边缘提取算法编程可得到近邻样本 k、阈值 r 分别与支持向量占比以及利用 Libsvm 分类器得到的分类正确率的关系图(图 1、图 2)。

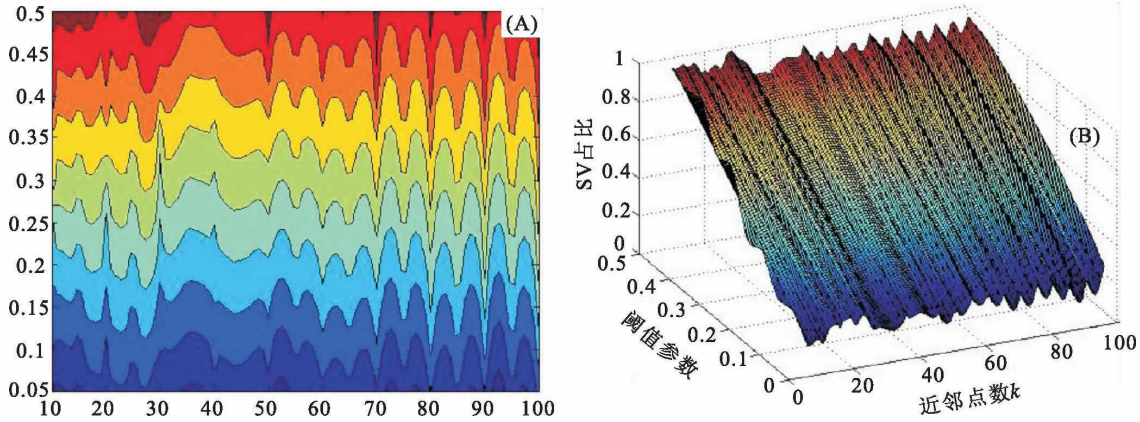

图 1 k、阈值 r 与 SV 占比关系

Fig. 1 The relationship between the k, threshold r and proportion of SV

(A)SV 占比等高线；(B)空间关系图

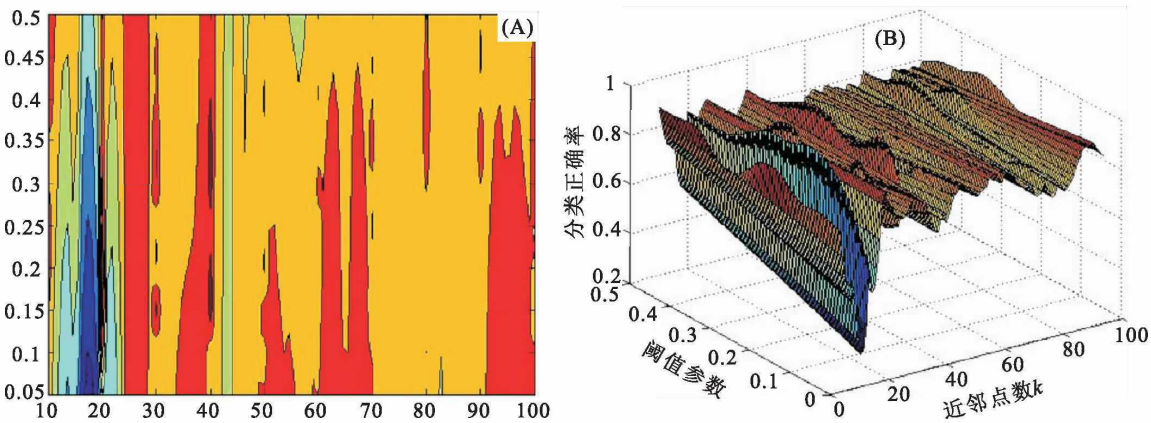

图 2 k、阈值 r 与分类正确率关系

Fig. 2 The relationship among k, threshold r and correct classification ratio

(A)分类正确率等高线；(B)空间关系图

由此阈值参数与 k 值并不是越大越好，支持向量的占比与 k 值关系不大，但在一定程度上会影响分类正确率。根据分类正确率等高线图可以看出颜色越深代表分类准确率越高，因此可确定阈值在[0.1，0.25]，并且近邻样本数在[20，35]以内分类正确率较高。所以本文取阈值 r 为 0.15，k 为 25。利用 MATLAB 编程，可得边缘提取结果(图 3)。

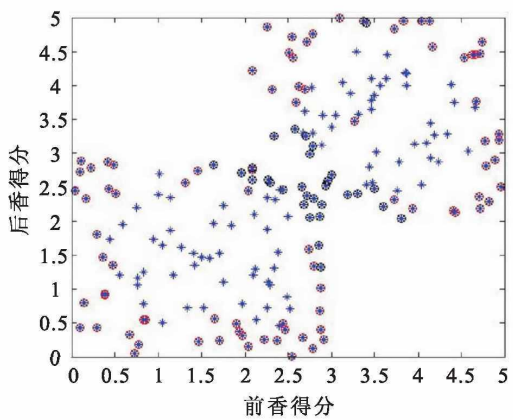

图 3 边界提取结果

Fig. 3 The results of boundary extraction

由图 3 可以看出，粗略提取出来的结果中选出的边界点显得较为重复，共提取边界点 134 个，因此可能存在样本点过多的情况。本文利用 SVM 构造超球面，其中高斯核函数的函数宽度取 $q=5$[19]，求出边界点的权重向量 β，部分权重系数如表 1。

为降低因样本点过度使用而导致的误差，本文通过设置阈值 $\beta_0=0.001$，将不重要的边界点去掉，可以得到简易分类结果(图 4)。

根据图 4 分类结果，将最终的边界点作为支持向量，对剩下的 100 个测试样本进行分类测试，

表 1　权重系数
Table 1　Weight coefficient

前香得分	后香得分	权重系数	前香得分	后香得分	权重系数
0.036	2.453	0.00156	2.296	2.571	0.000247
2.740	1.586	0.00041	0.291	0.437	0.00015
2.747	2.988	0.00534	2.872	0.687	0.00034
0.107	2.886	0.00022	2.873	2.068	0.00365
0.139	0.791	0.00171	2.296	2.571	0.000761
⋮	⋮	⋮	⋮	⋮	⋮
2.895	2.244	0.00032	2.912	0.251	0.01265

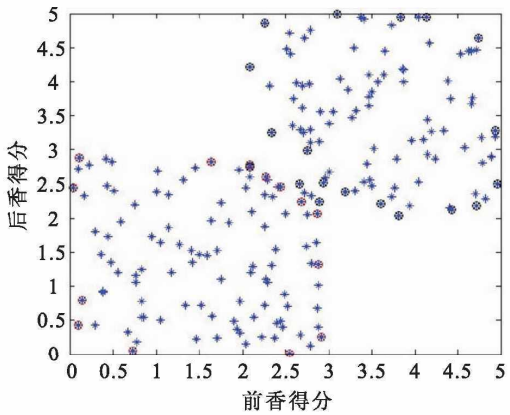

图 4　去掉不重要的边界点分类结果
Fig. 4　Diagram showing the removal of unimportant boundary point

其分类准确率高达 98.91%，由此说明通过提取数据集边界样本结合支持向量机建立超球面进行数据分类的方法是可行的。

4　结 论

本文通过分析单类数据集在空间所呈现的几何形态，从几何体的凹凸 2 种情况进一步分析数据集表面的点的近邻点与过该点的切平面之间的几何位置关系，将切平面用近似法向量刻画，并通过设置阈值来提取边界点。考虑到现实生活中所产生的数据类与类之间必定有部分边界重复的区域，因此本文在单分类的基础上提取利用 SVM 分类器训练数据样本结合贝叶斯后验概率进行边界提取。为更准确提取出比较重要的边界点，本文通过构造 SVM 超球体，将边界点进行训练同时赋予权重，通过设置阈值去掉部分不重要的边界点。最后，通过剩余的 100 个测试样本进行分类测试，其分类准确率为 98.91%，由此说明该方法的可行性。

[参 考 文 献]

[1] 黄光鑫. 支持向量机数据描述与支持向量机及其应用[D]. 成都：电子科技大学档案馆，2011：12－19.
Huang G X. Support Vector Description and Support Vector Machine and Their Application[D]. Chengdu: The Archive of University of Electronic Science and Technology of China, 2011: 12－19. (in Chinese)

[2] Li Y F, Zhou Z H. Imporoving semi-supervised support vector machines through unlabeled instances selection[C]//Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'11). Washington: AAAI Press, 2011: 386－391.

[3] Zhu J, Rosset S, Hastie T, *et al*. 1-norm support vector machines [C]//Advances in Neural Information Processing Systems 16. Cambridge of Massachusetts: MIT Press, 2003: 49－56.

[4] Psraokis I, Damoulas T, Girolami M A. Multiclass relevance vector machines: Sparsity and accuracy [J]. IEEE Transactions on Neural Networks, 2010, 21(10): 1588－1598.

[5]Tipping M. Sparse Bayesian learning and the relevance vector machine [J]. Journal of Machine Learning Research, 2001, 1(9): 2011－244.

[6] 吕常魁，姜澄宇，王宁生. 一种支持向量机聚类的快速算法[J]. 华南理工大学学报(自然科学版)，2005，33(1)：6－8.
Lu C K, Jiang C Y, Wang N S. A fast algorithm for support vector clustering[J]. Journal of South China University of Technology (Natural Science Edition), 2005, 33(1): 6－8. (in Chinese)

[7] 张战成，王士同，邓赵红，等. 支持向量机的一种快速分类算法[J]. 电子与信息学报，2011，33(9)：2181－2185.
Zhang Z C, Wang S T, Deng Z H, *et al*. Fast decision using SVM for incoming samples [J]. Journal of Electronics & Information Technology, 2011, 33(9): 2181－2185. (in Chinese)

[8] Burges C J. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121－167.

[9] LI Y H, Maguire L. Selecting critical patterns based on local geometrical and statistical information [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(6): 1189－1201.

[10] LI Y H. Selecting training points for one-class support vector machines [J]. Pattern Recognition Letters, 2011, 32(11): 1517－1522.

[11] 韩德强,韩崇昭,杨艺. 基于 K-最近邻的支持向量预选取方法[J]. 控制与决策,2009,24(4):494－498.
Han D Q, Han C Z, Yang Y. Approach for pre-extracting support vectors based on k-NN[J]. Control and Decision, 2009, 24(4): 494－498. (in Chinese)

[12] Wilson D R, Martinez T R. Reduction techniques for instance-based learning algorithms [J]. Machine Learning, 2000, 38(3): 257－286.

[13] Huang K Z, Zheng D N, Sun J, *et al*. Sparse learning for support vector classification[J]. Pattern Recognition Letters, 2010, 31(13): 1944－1951.

[14] Cristianini N, Scholkopf B. Support vector machines and kernel methods: The new generation of learning machines [J]. AI Magazine, 2002, 23(3): 31－41.

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

[16] Scholkopf B, Smola A J, Milliamson R C, *et al*. New support vector algorithms[J]. Neural Computation, 2000, 12(5): 1207－1245.

[17] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323－2326.

[18] Marchiori E. Hit miss networks with applications to instance selection [J]. Journal of Machine Learning Research, 2008, 9(6): 997－1017.

[19] 张翔,肖小玲,徐光祐. 一种确定高斯核模型参数的新方法[J]. 计算机工程,2007,33(12):52－54.
Zhan X, Xiao X L, Xu G Y. A new method for determining the parameter of Gaussian Kernel [J]. Computer Engineering, 2007, 33(12): 52－54. (in Chinese)