

文章编号:1001-9081(2008)08-2094-03

## 启动子序列的非均衡检测识别算法

罗泽举<sup>1</sup>,宋丽红<sup>2</sup>,陆胜<sup>1</sup>

(1.重庆工商大学 计算机科学与信息工程学院,重庆 400067; 2.重庆工商大学 经济管理实验教学中心,重庆 400067)  
(luozeju@126.com)

**摘要:**通过改进 Hessian 矩阵对角参数,调整支持向量机中超平面的位移,将数据量少的样本从两类非均衡样本中进行分离,结合隐马尔可夫随机迭代,实验发现,不能简单固定 Hessian 矩阵的对角参数,而必须加之以可调整的权系数才能控制错分的样本数. 对启动子序列进行识别,平均识别率达到 92.8%。

**关键词:**Hessian 矩阵;非均衡算法;隐马尔可夫模型;启动子识别

**中图分类号:**TP391.4;TP181   **文献标志码:**A

### Imbalanced recognition algorithm for promoter sequence

LUO Ze-ju<sup>1</sup>, SONG Li-hong<sup>2</sup>, LU Sheng<sup>1</sup>

(1. Computer Science and Information Engineering College, Chongqing Technology and Business University, Chongqing 400067, China;  
2. Economics and Management Center, Chongqing Technology and Business University, Chongqing 400067, China)

**Abstract:** By improving the Hessian matrix diagonal parameter, adjusting the hyperplane displacement and separating the few samples from two-class imbalanced samples, with reference to the hidden Markov random iterated algorithm, the experiment indicates that we cannot simply fix the diagonal parameter of Hessian matrix, and must add proper weight coefficient that can be adjusted to control the number of error-divided samples, recognize promoter sequence. The average recognition rate comes to 92.8%.

**Key words:** Hessian matrix; imbalanced algorithm; hidden Markov model; promoter sequence

### 0 引言

启动子是位于结构基因 5 端上游的一段 DNA 序列,能够指导全酶同模板正确结合,活化 RNA 聚合酶,启动基因转录。基因表达的第一步是将储存在 DNA 顺序中的遗传信息转录为 mRNA,这一步主要依靠基因上游调控顺序与转录因子之间的相互作用,这之中多聚酶结合位置的启动子序列起着重要作用,如果启动子序列失去活性,操纵子中的基因不表达,转录就不成功。因此对启动子的识别进而对基因表达进行分析有着非常重要的意义。

另一方面,由于有的启动子序列非常短小,各种生物的启动子序列长短不一,大小也不一样,一类启动子样本的数目可能明显多于另一类启动子样本数目,因而出现样本的非均衡性。这种非均衡的现象在其他领域中也常常出现,例如,信息系统领域中,一些偶尔出现的异类信号和大多数正常出现的信号;气象领域中,异常天气只占正常天气的一小部分等;这些小量的样本数量虽少,但却起着十分重要的作用,例如,最近我国南方出现的在历史上罕见的少量雨雪天气,尽管这种雨雪天气持续天数并不长,但却给我国南方造成重大直接经济损失。因此,研究这些少量样本的特征模式,将它们从多数样本中识别出来,对于及早采取有关措施进行调控减少损失,同样具有重大现实意义。

对于这些非均衡信息,由于其中可供学习的一类样本数非常少,需采用小样本学习算法进行分类和识别,支持向量机由于对小样本有非凡的处理能力而备受人们应用,Güneş S 等用标准 SVM 进行分类,但由于最优分类超平面间隔距离固定,为数多的负类混在为数少的正类中,造成分类结果不理

想<sup>[1-2]</sup>;Ertekin S 等改进了 Hessian 矩阵对角参数<sup>[3-4]</sup>,通过增加固定常数来增大最优超平面的间隔,从而使两类非均衡样本分离。实际上,加一固定常数值的弊端是不能确定是否可以真正分离出小量样本,甚至还可能错分更多的样本使识别率比标准 SVM 更低,因为算法并没有证明加入这个定常数后就可以分离出更多的样本。

本文提出一种改进的非均衡支持向量机(Imbalanced Support Vector Machines, ISVM)和隐马尔可夫模型(Hidden Markov ModelS, HMMS)随机迭代学习算法,通过改进 Hessian 矩阵对角参数,让计算机根据风险控制自动调整 Hessian 矩阵对角参数,自动移动超平面的位置,使它尽量靠近小量的正类样本或尽量靠近类多的负类样本,就可实现尽可能多地分离少数样本的目的。同时结合隐马尔可夫模型的随机迭代算法,将错分的样本进行二次识别和分类,识别率比传统算法明显提高,对启动子序列进行识别,平均识别率达到了 92.8%。

### 1 非均衡支持向量机算法

假设数据点在样本空间中线性不可分,设训练样本为  $(x_1, y_1), \dots, (x_k, y_k)$ , 其中正类样本数为  $m^+$ , 负类样本数为  $n^-$ ,  $k = m^+ + n^-$ , 则正类样本占的比重是  $m^+/k$ , 负类样本占的比重是  $n^-/k$ 。为了讨论的方便,根据非均衡样本的分布情况,将为数少的样本定为正类,而将为数多的样本定为负类,惩罚常数  $C$  也因此分为两部分,一部分是针对小的正类样本的  $C^+$ , 另一部分是针对负类的  $C^-$ , 我们将优化问题变为:

$$\min_w \Phi(w) = \frac{1}{2} \|w\|^2 + C^+ \left( \sum_{i=1, y_i=+1}^k \xi_i \right) + C^- \left( \sum_{i=1, y_i=-1}^k \xi_i \right) \quad (1)$$

收稿日期:2008-02-14;修回日期:2008-03-16。

基金项目:重庆市科委自然科学基金资助项目(2007BB2205);重庆市教育委员会科学技术研究资助项目(KJ0707022)。

作者简介:罗泽举(1965-),男,重庆人,副教授,博士,主要研究方向:机器学习、模式识别、生物信息学;宋丽红(1969-),女,四川越西人,实验师,主要研究方向:机器学习、数据仓库;陆胜(1974-),男,重庆人,副教授,博士,主要研究方向:智能系统控制。

$$\text{s. t. } -[y_i(w \cdot x_i + b) - (1 - \xi_i)] \leq 0, i = 1, 2, \dots, k \\ -\xi_i \leq 0, i = 1, 2, \dots, k$$

为了解决上述优化问题,构造 Lagrange 函数:

$$L(w, b, A, \Xi, \Gamma) = \frac{1}{2} \|w\|^2 + C^+ \left( \sum_{i=1, \gamma_i=+1}^k \xi_i \right) + \\ C^- \left( \sum_{i=1, \gamma_i=-1}^k \xi_i \right) - \sum_{i=1}^k \lambda_i [y_i(w \cdot x_i + b) - (1 - \xi_i)] - \sum_{i=1}^k \gamma_i \xi_i \quad (2)$$

其中,  $\Lambda = (\lambda_1, \dots, \lambda_k)$ ,  $\Xi = (\xi_1, \dots, \xi_k)$ ,  $\Gamma = (\gamma_1, \dots, \gamma_k)$ 。对  $L(w, b, A, \Xi, \Gamma)$  求偏导数:

$$\begin{cases} \frac{\partial L(w, b, A, \Xi, \Gamma)}{\partial w} = w - \sum_{i=1}^k \lambda_i y_i x_i = 0 \\ \frac{\partial L(w, b, A, \Xi, \Gamma)}{\partial b} = \sum_{i=1}^k \lambda_i y_i = 0 \\ \frac{\partial L(w, b, A, \Xi, \Gamma)}{\partial \Xi} = C^+ - \lambda_i - \gamma_i = 0, \gamma_i = +1 \\ \frac{\partial L(w, b, A, \Xi, \Gamma)}{\partial \Xi} = C^- - \lambda_i - \gamma_i = 0, \gamma_i = -1 \end{cases} \quad (3)$$

从而得到 Lagrange 常数,如果  $y_i = -1: 0 \leq \lambda_i \leq C^-; y_i = +1: 0 \leq \lambda_i \leq C^+$ ,记  $\Lambda = (\lambda_1, \dots, \lambda_k)$ ,则问题转化为对偶:

$$F(\Lambda) = \sum_{i=1}^k \lambda_i - \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k \lambda_i \lambda_j y_i y_j K(x_i, x_j) \quad (4)$$

的下列问题:

$$\text{Maximize } F(\Lambda) = \Lambda \cdot I - \frac{1}{2} \Lambda H \Lambda^\top \quad (5)$$

$$\text{s. t. } \Lambda \cdot y = 0; 0 \leq \Lambda \leq C$$

其中,  $C$  是由相应的  $C^+$ 、 $C^-$  组成的向量。于是得到 Hessian 矩阵为:

$$H = (y_i y_j K(x_i, x_j))_{k \times k} \quad (6)$$

我们改进 Hessian 矩阵的对角调整技术,调整 Hessian 矩阵元素为:

$$H(i, j) = \begin{cases} H(i, j), & i \neq j \\ H(i, j) + \frac{m^+}{k} \delta^+, & i = j; y_i = 1 \\ H(i, j) + \frac{m^-}{k} \delta^-, & i = j; y_i = -1 \end{cases} \quad (7)$$

其中,  $\delta^+ \geq 0, \delta^- \geq 0$ , 在这里,关键是对角参数的条件  $\delta^+ m^+/k \geq 0, \delta^- m^-/k \geq 0$ ,比起以前的方法进行了很大的改进和放松<sup>[1-2]</sup>。因为我们发现,由于  $n^-/k$  比  $m^+/k$  大很多,结果是超平面明显往样本数多的负类偏移,如果仅以  $n^-/k, m^+/k$  为参数或者简单固定参数  $\delta^+ m^+/k, \delta^- m^-/k$ ,超平面甚至可能会越过某些负类样本而使伪正类样本显著增加,这样反而会使错分的样本更多。因此我们结合偏移的情况增加以权参数  $\delta^+, \delta^-$ ,让学习机器根据要求自动调整权参数,以适当调整超平面靠向负(正)类一侧,将真正正类的少数样本分离出来,从而达到提高分类效果的目的。

模型中  $C^+, C^-$  是惩罚参数,控制对两类样本错分的惩罚,由于负类样本居多,故主要是对负类偏向正类的惩罚,也就是为了减少伪正样本的数量,而不是为了减少伪负类样本的数量,因为正类样本本身就很少,因此可以取  $C^-$  大些,而取  $C^+$  小些。

## 2 隐马尔可夫离散谱变换

对于隐马氏模型  $\lambda = (S, \Sigma, A, B, \pi)$ ,重估算法最早由 Baum-Welch<sup>[5-6]</sup>提出,主要用来重新估计 HMMS 模型参数  $A, B, \pi$ ,也就是 HMMS 的训练问题,它利用最优化方法,使得由模型生成序列的值  $P(O | \lambda)$  达到最大<sup>[7-8]</sup>。算法的思想是先

得到一组参数值,再进行迭代运算,直到算法收敛到一个局部最优解。设由模型  $\lambda$  产生观察序列  $O$  的概率为  $P(O | \lambda)$ ,用其自然对数  $L = \log P(O | \lambda) = \ln P(O | \lambda)$  计算序列的离散谱范围。设观察序列是  $O = O_1^* O_2^* \cdots O_n^*$ ,相应地状态序列为  $Q = q_1 q_2 \cdots q_n$ ;定义变量:

$$\alpha_t(i) = P(O_1^* O_2^* \cdots O_t^*, q_t = S_i | \lambda) \quad (8)$$

$$\beta_t(i) = P(O_{t+1}^* O_{t+2}^* \cdots O_n^* | q_t = S_i, \lambda) \quad (9)$$

$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j | O^*, \lambda) = \\ \alpha_t(i) a_{ij} b_j(O_{t+1}^*) \beta_{t+1}(j) / P(O^* | \lambda) = \\ \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}^*) \beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) a_{ij} b_j(O_{t+1}^*) \beta_{t+1}(j)} \quad (10)$$

$$\gamma_t(i) = P(q_t = S_i | O^*, \lambda) = \sum_{j=1}^N \xi_t(i, j) \quad (11)$$

我们采用如式(12)进行模型的参数重估:

$$\pi_j^* = \gamma_t(i), a_{ij}^* = \frac{\sum_{t=1}^{k-1} \xi_t(i, j)}{\sum_{t=1}^{k-1} \gamma_t(j)}, b_j^*(k) = \frac{\sum_{t=1, O_t^*=v_K}^k \gamma_t(j)}{\sum_{t=1}^k \gamma_t(j)} \quad (12)$$

Baum 等证明<sup>[5-6]</sup>得出,只要:

$$\text{Max}_{\lambda^*} Q(\lambda, \lambda^*) = \sum_Q P(Q | O, \lambda) \log \{P(O, Q | \lambda^*)\} \quad (13)$$

则有:

$$P(O | \lambda^*) \geq P(O | \lambda) \quad (14)$$

且算法收敛到一个局部最大值。

## 3 非均衡 SVM 和 HMM 混合分类迭代模型

结合支持向量机处理非均衡样本的上述算法,提出非均衡 SVM 和 HMM 混合分类模型,如图 1 所示。模型算法如下:

步骤 1 获取数据样本,对数据样本进行清除异常数据和补缺处理及标准化。

步骤 2 根据预先确定的初值  $C^+, C^-$  及权参数  $\delta^+, \delta^-$  经过变换的核函数进行 SVM 分类训练。

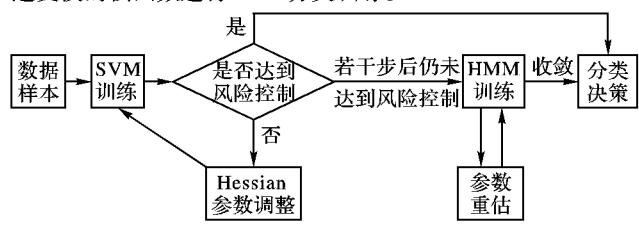


图 1 非均衡 SVM 混合识别模型

步骤 3 根据训练结果看是否达到风险要求,控制决策面和错分率,如果满足要求,则转入步骤 7,进行正式分类预测;否则转入步骤 4,让计算机自动调整 Hessian 矩阵对角参数继续训练,若干步后如果满足要求,转入步骤 7,如果若干步后仍未达到风险要求则转入步骤 5,进行隐马尔可夫离散谱变换。

步骤 4 调整 Hessian 矩阵对角参数,转入步骤 3。

步骤 5 随机生成初始模型参数。

$$\pi = (\pi_j), B = (b(k_j)) \quad (15)$$

训练隐马尔可夫离散谱范围,如果收敛则转入步骤 7;否则转入步骤 6。

步骤 6 用 EM 算法进行 HMMS 模型训练重新估算模型三个参数,直到收敛。

步骤 7 进行分类测试。

步骤 8 停止迭代。

该算法关键是步骤 4~6, 步骤 4 是进行 Hessian 参数调整以达到更加精确分离小样本的目的; 步骤 5 的初始状态概率  $\pi$  和散发矩阵  $B$  都是随机假定的, 而不是一般算法中要事先计算出来; 步骤 6 为了 HMM 收敛而进行模型参数调整。结果显示, 步骤 5 的随机假定识别效果良好。

## 4 实验结果

从瑞士实验癌研究组织生物信息组的真核生物非沉余启动子数据库(The Eukaryotic Promoter Database, EPD)(当前版本为 Release 93, 2008-02-11) 下载启动子序列数据(<http://www.epd.isb-sib.ch/>)。下载节肢动物(Arthropode)启动子序列 300 个, 原核生物玉蜀黍花柱(Zea)启动子序列 8 个, 脊椎动物选取非洲爪蛙(African clawed frog)启动子序列 28 个, 棘皮类动物(Echinoderm)启动子序列 44 个, 软体动物(Mollusc)启动子序列 3 个, 以这五类为样本进行实验, 有的 DNA 序列很少(如软体动物只有 3 个), 有的则很多(如节肢动物有 300 个), 样本明显是非均衡的, 以两类和多类分别进行实验。

### 4.1 实验参数调整结果

取玉蜀黍花柱和棘皮类动物进行两类非均衡样本实验,

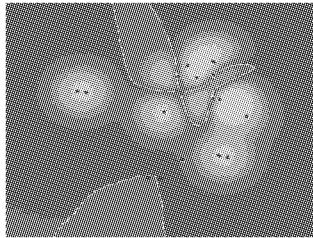


图 2 错分 4 个正类样本

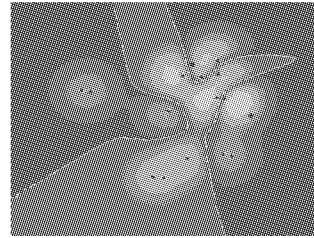


图 3 样本得到正确分类

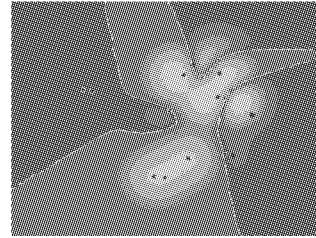


图 4 错分 4 个负类样本

### 4.2 实验指标

为了表示两类和多类的识别结果, 以伪正类(False Positive, FP)、伪负类(False Negative, FN)、真正正类(True Positive, TP)、真正负类(True Negative, TN), 平均准确率(Mean

玉蜀黍花柱 6 个(取为正类), 棘皮类动物 10 个(取为负类)作为训练, 取各条序列的 200 bps 作统一的长度, 相当于每个样本为 200 维, 通过调整 Hessian 矩阵对角参数来观察间隔面的变化, 用主成分分析法, 只取前两个主成分, 即将它投影到平面二维空间来观察其超平面的位置变化, 由于正负样本非线性可分, 我们用径向基核函数进行分类, 参数调整规则是:

$$\begin{cases} H(i,i) \leftarrow \exp[-\|x - x_i\|^2 / 2\omega^2] + \delta^+ \frac{m^+}{k}, & y_i = 1 \\ H(i,i) \leftarrow \exp[-\|x - x_i\|^2 / 2\omega^2] + \delta^- \frac{n^-}{k}, & y_i = -1 \end{cases} \quad (16)$$

其中,  $k = 16, m^+ = 6, n^- = 10$  可以看到, 当  $\delta^+ m^+ / k = 12/16 > \delta^- n^- / k = 7/16$  时, 权重太偏向正类一边, 使得 4 个样本远离正类样本, 达到了错分 4 个正类样本的事实(图 2); 当减少正类权重和增加负类权重达到  $\delta^+ m^+ / k = 8/16 \leq \delta^- n^- / k = 9/16$ , 超平面接近负类, 此时 6 个正类样本得到了正确区分, 而且全部负类样本也得到了正确区分(图 3), 边界面接近正类样本, 这正是我们需要的情况; 但如果此时继续减少正类样本的权重达到  $\delta^+ m^+ / k = 7/16 < \delta^- n^- / k = 12/16$  时, 最优分类超平面已经越过了 4 个负类样本, 使这 4 个负类样本得到错分(图 4)。

Accuracy Rate for Total, MART) 来比较节肢动物、玉蜀黍花柱、脊椎动物、棘皮类动物和软体动物这五类启动子 DNA 序列的识别情况, SVM 用一对投票策略进行两类决策, 得到表 1 所示的识别结果。

表 1 ISVM 和标准 SVM 比较

类别	样本数	训练类别	ISVM					标准 SVM				
			FP	FN	TP	TN	MAR	FP	FN	TP	TN	MAR
玉蜀黍花柱	8	P	0	1	7	0	0.875	0	2	6	0	0.750
非洲爪蛙	28	N	3	0	0	25	0.893	4	0	0	24	0.857
棘皮类动物	44	P	0	3	41	0	0.932	0	4	40	0	0.910
节肢动物	300	N	20	0	0	280	0.933	30	0	0	270	0.900

注:SVM 采用径向基核函数

虽然在上节画图时我们用 PCA 方法将空间投影到了二维, 但为了和标准 SVM 进行比较, 在识别阶段不进行降维而直接用 ISVM 和标准 SVM 计算。由于数据点非线性可分, 核函数选取径向基核函数, 计算时以伪正、伪负、真正正、真正负四个指标计算准确率而得出平均值。表 1 是玉蜀黍花柱和非洲爪蛙及棘皮类动物与节肢动物启动子序列的两类比较结果。玉蜀黍花柱类 8 个, ISVM 可以分离出 7 个正确样本来, 而标准 SVM 只能分离出 6 个, 非洲爪蛙 ISVM 能正确分离出 25 个, 而标准 SVM 只能分离出 24 个。

表 2 是各类启动子的平均识别率, 从表 2 可以看出, ISVM 比起标准 SVM 来, 越是样本少的类其学习算法的识别效率就越显著。例如, 对于玉蜀黍花柱, ISVM 的平均识别率

为 87.5%, 比而标准 SVM 的 75% 要高出 12 个百分点; 而对于样本数目多的样本例如节肢动物, 两种方法识别的优势并不明显, ISVM 为 91.3%, 标准 SVM 为 91%, ISVM 比标准 SVM 只高出 0.3 个百分点, 原因是因为负类中为数少的错分样本比起整个负类样本来说, 所占比重是很小的。因此 ISVM 算法是针对非均衡样本集的, 主要目的是将为数少量的样本能从混合的多数样本中分离出来。

从表 2 的多类识别结果进一步得知, 其中 ISVM 栏是用两类识别结果的平均得到的, ISVM + HMM 栏是用 ISVM 结合隐马尔可夫重估算法得到的, 当 Hessian 参数调整无法达到所需的要求时, 我们再结合 HMM 训练, 将不同序列映射到不

(下转第 2108 页)

点所收到并转发的消息数量如图 7 所示。

同样,虽然每个代理节点发起的回复次数是随机的,节点之间相差十分剧烈。即使在不均衡条件下,每个代理节点所承受的压力十分均衡。从纵向比较,可以发现节点之间压力的均衡度有随着规模上升而上升的趋势,并且节点转发消息的数量与节点发起次数仍然呈现线形关系。另外从实验结果来看,每个事件平均在网络中的转发跳数平均为 1.709,满足式(1)。综上所述,超立方体结构的引入,在广播和逆向传播时能够很好地均衡节点之间承载的压力。

### 3 结语

P/S 系统出于自身特点,它的路由算法设计比较复杂。其原因在于事件发布或者事件订阅没有指明具体的接收者,有着不确定性。所以,需要对已有的以广播协议为基础的算法进行优化,以避免不必要的消息转发。这里引入超立方体结构,提供一种高效的广播机制和搜索机制,从底层结构上增强了现有的 P/S 系统。本文涉及到超立方体结构的 P/S 系统,还有必要进一步研究其改进的策略。例如:由于在 P/S 系统中订阅事件、取消订阅事件、结构维护消息等都会频繁地在网络中传播,这将会增加网络链接的开销。其次,对于节点的语义聚集性也有必要进一步研究的方向,引入语义网的概念提高搜索效率。

#### 参考文献:

- [1] EUCSTER P T, FELBER P A, GUERRAOUI R, et al. The many faces of publish/subscribe [J]. ACM Computing Surveys, 2003,

(上接第 2096 页)

同的谱范围,采用 HMM 识别多类特有的 L 值谱映射优势进行进一步识别,得到了更为理想的识别结果,平均识别率达到了 92.8%。

表 2 ISVM、ISVM + HMM 和标准 SVM 比较

启动子类别	样本数	平均准确率(MART) /%		
		ISVM	ISVM + HMM	标准 SVM
棘皮类动物	44	90.0	93.1	88.0
软体动物	3	100.0	100.0	100.0
非洲爪蛙	28	89.3	92.8	86.0
玉蜀黍花柱	8	87.5	100.0	75.0
节肢动物	300	91.3	91.6	91.0

注: 正类准确率为  $A_T = TP/m^+$ ; 负类准确率为  $A_N = TN/n^-$ ; 总体准确率为  $A_{all} = (TP + TN)/k$ ;  $m^+ = TP + FN$ ,  $n^- = TN + FP$ ,  $k = m^+ + n^-$ 。

### 5 结语

进一步的实验还发现,超平面始终偏向对角权重系数大的类,而且权重系数可以是任意大于等于零的实数,于是我们将 Hessian 矩阵对角参数范围进一步拓宽到  $[0, \infty)$ 。这一结论与传统结论不同<sup>[1-2]</sup>,我们认为要根据风险控制进行自动迭代调整而不是将这些参数值固定,这将有利于提高非均衡样本的识别率。对于多类识别,ISVM 算法经过若干步后如果仍未达到风险控制,结合隐马尔可夫离散谱映射进行第二次识别,使得识别效果比单独使用非均衡算法更显著。

35(2): 114 - 131.

- [2] 马建刚, 黄涛, 汪锦岭, 等. 面向大规模分布式计算发布订阅系统核心技术[J]. 软件学报, 2006, 17(1): 134 - 147.
- [3] SCHLOSSER M, SINTEK M, DECKER S, et al. Shaping up peer-to-peer networks[EB/OL]. [2007-08-23]. <http://infolab.stanford.edu/~schloss/docs/HyperCuP-DISC2002.pdf>.
- [4] RATNASAMY S, SHENKER S, STOICA I. Routing algorithms for DHTs: Some open questions[EB/OL]. [2007-08-26]. <http://www.cs.rice.edu/Conferences/IPTPS02/174.pdf>.
- [5] CARZANIGA A A, ROSENBLUM D, WOLF A. Design and evaluation of a wide-area event notification service[J]. ACM Transactions on Computer Systems, 2001, 19(3): 332 - 383.
- [6] STROM R, BANAVAR G, CHANDRA T, et al. Gryphon: An information flow based approach to message brokering[C]// International Symposium on Software Reliability Engineering'98. Washington D C, USA: IEEE Computer Society, 1998: 89 - 94.
- [7] CUGOLA G, NITTO E D, FUGGETTA A. The JEDI event-based infrastructure and its application to the development of the OPSS WFMS[J]. IEEE Transaction on Software Engineering, 2001, 27(9): 827 - 850.
- [8] ROWSTRON A, KERMARREC A M, CASTRO M, et al. SCRIBE: The design of a large-scale event-based notification infrastructure[C]// Proceeding of the 3rd International Workshop on Networked Group Communication. London: Springer-Verlag, 2001: 30 - 43.
- [9] PIETZUCH P R. Hermes: A scalable event-based middleware [D]. UK: University of Cambridge, 2004.

#### 参考文献:

- [1] SONNENBURG S, ZIEN A, RATSCH G. ARTS: Accurate recognition of transcription starts in human[J]. Bioinformatics, 2006, 22(14): 472 - 480.
- [2] POLAT K, GUNES S, ARSLAN A. A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square Support Vector Machine[J]. Expert Systems with Applications: An International Journal, 2008, 34(1): 482 - 487.
- [3] YU T, DEBENHAM J, JAN T, et al. Combine vector quantization and support vector machine for imbalanced datasets[C]// Artificial Intelligence in Theory and Practice. Boston: Springer, 2006, 217: 81 - 88.
- [4] ERTEKIN S, HUANG J, BOTTOU L, et al. Learning on the border: active learning in imbalanced data classification[C]// Proceedings of The ACM 16th Conference on Information and Knowledge Management. New York: ACM press, 2007: 127 - 136.
- [5] BAUM L E, SELL G R. Growth functions for transformations on manifolds[J]. Pacific Journal of Mathematics, 1968, 27(2): 211 - 227.
- [6] BAKER J K. The dragon system - an overview [J]. IEEE Acoustic Speech Signal Processing, 1975, 23(1): 24 - 29.
- [7] SHA F, SAUL L K. Large margin Gaussian mixture modeling for phonetic classification and recognition[C]// Proceedings of ICASSP 2006. Toulouse: IEEE Signal Processing Society, 2006: 265 - 268.
- [8] SHA F, SAUL L K. Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models[C]// Proceedings of ICASSP, 2007. Hawaii: IEEE Signal Processing Society, 2007, 4: 313 - 316.