一种基于粗糙集启发式的特征选择算法

梁 琰 何中市
(重庆大学计算机学院 重庆 400044)

摘 要 本文基于粗糙集中关于非精确集和精确集理论思想,提出了一个新的特征度量指标,即相对互信息比 RMI,由此,设计了一种基于粗糙集的启发式特征选择算法 MRMI-UC。首先利用可辨识矩阵,计算出条件属性相对于决策属性的核,以核形成当前候选特征子集作为基准点,以最大化相对互信息和不确定性系数为原则,筛选剩余特征。通过对比实验,结果表明,本文提出的算法在多数情况下能够得到较优的特征子集,算法是有效的,切实可行的。

关键词 特征选择,粗糙集理论,启发式算法,不确定性系数,互信息

A Novel Feature Selection Heuristic Algorithm Based on Rough Set Theory

LIANG Yan HE Zhong-Shi
(College of Computer Science, Chongqing University, Chongqing 400044)

Abstract In this paper, a new feature measurement RMI (Ratio of Mutual Information) is presented based on the concept of rough set theory about certain set and uncertain set. Then a novel heuristic algorithm, MRMI-UC (Algorithm based on Maximal Ratio of RMI and Uncertainty Coefficient), is proposed for Feature Selection based on rough set theory. Firstly, the Core is obtained by discernible matrix and formed as a candidate feature subset. With the starting point of Core, the rest features are filtered iteratively to maximize both RMI and Uncertainty Coefficient. Finally the algorithm is tested on the UCI datasets, experiment results show that MRMI-UC is feasible and can find a good feature subset in most cases.

Keywords Feature selection, Rough set theory, Heuristic algorithm, Uncertainty coefficient, Mutual information

1 引言

特征选择(也称属性约简)是数据挖掘中的基本问题之一。通过对样本数据进行特征选择,可以去掉不相关的和冗余的特征,使机器学习过程复杂度和时间降低。特征选择是根据某种标准,通过选择相关特征和丢弃不相关、冗余特征来形成一个优化特征子集的处理过程[1]。特征选择也是一个搜索和评价的过程。特征选择能为特定的应用在不失去数据原有价值的基础上选择最小的属性子集,去除不相关和冗余的属性;它提高了数据的质量,加快了数据挖掘的速度。特征选择算法可从搜索方向、搜索策略、评价方法和停止标准等四个方面考察特征的选择。

粗糙集(Rough Set)理论是一种处理模糊和不确定知识的数学工具,最早由波兰数学家 Z. Pawlak 在 1982 年提出[2]。它已经在数据挖掘、人工智能、模式识别与分类等领域获得了较广泛的应用。求核和属性约简是粗糙集理论研究的一个核心内容[3]。人们希望找到最佳属性约简。然而 Wong S. K. M 和 Ziarko W. 已经证明它是 NP-hard 问题[4],因而目前还没有高效的最佳属性约简算法。不过,在实际应用中,要求得到相对属性约简就可以了。其中,启发式搜索算法应用得最多。

本文提出的基于粗糙集的启发式特征选择算法,是以核形成当前候选的特征子集作为基准点,根据本文提出的新标准——以相对互信息和不确定性系数最大化为原则,筛选剩余特征。实验结果表明,本文提出的方法是有效的,切实可行的。

2 粗糙集理论

粗糙集理论从集合的视角对知识进行定义,把知识看作是关于论域的划分,从而对知识进行分析和处理。

定义 1[5] 信息系统 $L=(U,A,V,F)$。其中 $U=\{x_1,x_2,\cdots,x_n\}$ 是论域,A 是属性集合,$A=R\cup D$,R 是条件属性集,D 是决策属性集,V 是属性值集合,F 是 $U\times A\to V$ 的映射。

定义 2[5](可辨识矩阵) 可辨识矩阵由华沙大学数学家 Skowron 提出。信息系统 L,如定义 1,其中 $R=\{a_i\mid i=1,\cdots,h\}$ 和 $D=\{d\}$ 分别为条件属性集和决策属性集,$a_i(x_j)$ 是样本 x_j 在条件属性 a_i 上的值,$d(x_j)$ 是样本 x_j 在决策属性 d 上的值。可辨识矩阵定义为 $W=(w_{ij})_{n\times n}$,其中 w_{ij} 表示能够辨识样本 x_i 和 x_j 的条件属性的子集合,具体定义如下:

$$w_{ij}=\begin{cases}\{a_t\mid a_t\in R\wedge a_t(x_i)\neq a_t(x_j)\}, & d(x_i)\neq d(x_j);\\ 0, & d(x_i)=d(x_j).\end{cases}$$

在粗糙集理论研究中,有关研究人员已经通过 Skowron 提出的可辨识矩阵得到决策表的核[6],决策表的核是唯一的,它可以作为最佳属性约简起点。在可辨识矩阵中属性组合数为 1 的属性即为核,其余的有用属性可从属性不为 1 的矩阵元素中获得。属性集的核是属性集的本质部分,从中去掉任何一个属性都将影响属性集对论域中对象的区分能力。

*)国家自然科学基金资助(资助号:60173060)。梁 琰 硕士研究生,主要研究方向为模式识别。何中市 博士生导师,主要研究方向为自然语言处理、机器学习与数据挖掘。

3 特征子集的不确定性系数

设一个样本集 SM 可以表示如下$\{(S,D)_i \mid i=1,2,\cdots,m\}$，$S$ 表示条件属性集(特征集)$\{f_1,f_2,\cdots,f_p\}$，D 表示决策属性集$\{d_1,d_2,\cdots,d_q\}$)，m 表示样本集的大小。一般样本集只有一个类特征，所以 q 在这种情况下为 1，本文取 q 为 1，设这个唯一的类特征为 C。在样本集中，如果两个样本特征值相同，但分属的类不同，这两个样本成为不一致样本。在样本集 SM 中，如果将所有的样本集 S 中的特征的值相同的样本归到一个较小样本集中，则 SM 将被划分为多个较小的样本子集，这个过程称为 S 划分 SM。同理，类特征 C 划分 SM 为样本子集$\{C_i \mid i=1,2,\cdots,u\}$，$u$ 为样本子集的大小，这里也就是类数。特征集 S 划分 SM 为$\{SS_j \mid j=1,2,\cdots,v\}$，$v$ 为样本子集的多少；C 划分 SS_j 为$\{sc_{ij} \mid i=1,2,\cdots,u,\ j=1,2,\cdots,v\}$。一个特征子集 S 的不确定性系数 $U(S)$ 的计算过程如下[7]：

1)计算类特征 C 的期望信息：

$$I(C)=-\sum_{i=1}^{u}\frac{|C_i|}{m}\log_2\frac{|C_i|}{m} \tag{1}$$

其中$|X|$表示集合 X 势(元素的个数)。

2)计算 SS_j 的期望信息：

$$I(SS_j)=-\sum_{i=1}^{u}\frac{|sc_{ij}|}{|SS_j|}\log_2\frac{|sc_{ij}|}{|SS_j|} \tag{2}$$

3)计算特征集 S 的熵：

$$E(S)=\sum_{j=1}^{v}\frac{|SS_j|}{m}I(SS_j) \tag{3}$$

4)计算特征集 S 的不确定性系数：

$$U(S)=\frac{I(C)-E(S)}{I(C)} \tag{4}$$

在不确定性系数计算过程中，样本集的划分是一个基本操作。样本集划分过程中，可能会出现某些样本集为空，在计算时这样的样本集不作计算。同样，特征如果为连续值或者缺省值，将同样使划分很困难。所以，在计算进行之前，首先必须对样本集进行离散化处理和缺省值处理。

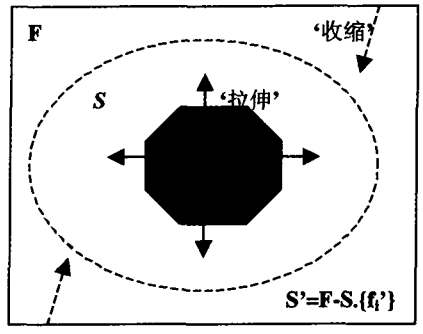
S 的不确定性系数的上界为 1，下界为 0。当样本集中没有不一致样本存在时，S 的不确定性系数为 1，表示特征集 S 能够完全清楚无误地描述类特征；当样本集完全无序，也就是当每个 SS_j 被 C 划分为大小相等的小样本集时，S 的不确定性系数为 0。S 的不确定性系数越大，特征集描述类特征的能力越强。

4 基于相对互信息比和不确定性系数的特征选择算法

4.1 指导思想

根据特征选择搜索策略，特征选择算法分为 3 类[8]：完全搜索算法，启发式搜索算法和随机搜索算法。完全搜索算法是计算每一种可能的特征子集的特征选择度量，找到符合选择判据的最优的特征子集，如分支界限法，Focus 和 ABB[9]。启发式算法是根据某种特征选择方向找到一个次优的特征子集。随机选择算法是在规定的时间或者次数内随机地选择特征子集来判断，以此来找到一个次优特征子集。后两类算法比第一类算法快，但搜索结果的质量被降低了。

当数据是高维、样本数目较大时，用完全搜索算法进行搜索是一个时间和空间都大量耗费的工程。启发式和随机选择算法则可以在较短的时间内找到一个较优的特征子集。对于启发式算法，能找到一个有效的方向并能按方向到达目标是算法的关键。本文提出的算法，是以可辨识矩阵求得的核作为基点出发，顺着相对互信息比最大这个方向，对属性子集空间进行适当的'拉伸'或'收缩'的方法，以特征子集的不确定系数最大化为原则，最终形成约简后的特征子集，示意图如图 1。



F：条件特征集(外围长方形)；CORE：核(内六边形)；S：当前候选的特征子集；S'：F-S 当前非候选特征子集，S 和 S' 是变化的。初始时 S=CORE，S'=F-CORE；从 S'中选出合适的特征加入 S，按指标，达到稳定的 S。

图 1 MRMI-UC 算法思想示意图

本文利用粗糙集精确集与非精确集的思想，先找出精确的特征子集，也就是核，其余的看作为非精确的子集，接着根据核，一方面逐步"拉伸"扩展特征子集，另一方面逐步"缩小"F 范围，最后使得形成的特征子集相对于决策属性的分类和初始条件属性所形成的相对于决策属性的分类能力一致。其中，根据核，提出了一种新的评价指标-相对互信息比和不确定性系数相结合的评价指标，指导"拉伸"或者"缩小"的动作，直至特征子集稳定。

4.2 MRMI-UC 特征选择算法描述

根据粗糙集理论，属性集的核是属性集中本质部分，从中去掉任何一属性都将会影响属性集对样本的区分能力[10]。由此可认为核是精确的特征子集，因此，以核形成当前候选的特征子集 S 作为起点和基准点，在对剩余特征 $f'\in S'$ 选择时，不仅要考虑特征 f' 与类别属性间的相关性和一致程度，还有考虑 f' 与 S 间相关性和冗余度。在此基础上，本文提出了将一个特征 f' 分别与类别属性和当前候选特征子集相关程度之间的比值，作为选择特征 f' 的一个新的评价指标，称为相对互信息比率，记为 RMI(Ratio of Relative Mutual Information)，定义如下：

定义 3 特征 f' 的相对互信息比率 $RMI(f')$：

$$RMI(f')=\frac{r_{cf'}}{r_{Sf'}} \tag{5}$$

其中，c 表示类别属性；$S=\{f_i\}_{i=1}^{k}$ 表示当前候选特征子集，含有 k 个特征；S'=F−S 表示非候选特征子集，f' 是 S' 的任意一个特征。$r_{cf'}$ 表示特征 f' 与类别属性 c 间的相关程度；$r_{Sf'}$ 表示特征 f' 与集合 S 中属性的平均相关程度。

一般地，特征 f' 可纳入候选特征子集 S 中，应该使得 $r_{cf'}$ 尽可能最大，而 $r_{Sf'}$ 尽可能小，从而 $RMI(f')$ 尽可能最大。因此，本文提出将最大化相对互信息比率作为特征筛选的原则。

本文从互信息角度度量属性间的相关程度，采用互信息计算(5)式相对互信息比率 RMI。

如果 x 和 y 是离散的随机变量，则 x 与 y 的互信息定义为：

$$MI(x,y)=H(y)-H(y|x)$$

其中：

$$H(y)=-\sum_{y\in Y}p(y)\log_2 p(y),$$

$$H(y)=-\sum_{y\in Y}p(y)\log_2 p(y),$$

$$H(y|x)=-\sum_{x\in X}p(x)\sum_{y\in Y}p(y|x)\log_2 p(y|x)$$

用互信息 $MI(x,y)$ 衡量特征 x 与 y 间的关联程度，$MI(x,y)$ 越大，则表示 x,y 间高度相关。因此，可以用如下两式计算 $RMI(f')$ 中的 $r_{cf'}$ 和 $r_{Sf'}$：

$$r_{cf'}=MI(c,f') \tag{6}$$

$$r_{sf'}=\frac{1}{|S|}\sum_{f_i\in S}MI(f_i,f') \tag{7}$$

由此得出

$$RMI(f')=\frac{MI(c,f')}{\frac{1}{|S|}\sum_{f_i\in S}MI(f_i,f')} \tag{8}$$

根据以上分析，本文提出了一种基于粗糙集的启发式特征选择算法：

算法：MRMI-UC (Algorithm based on Maximal Ratio of Relative Mutual Information and Uncertainty Coefficient for Feature Selection)

输入：样本集 SM 可以表示如下$\{(F,c)_i \mid i=1,2,\cdots,m\}$，$F$ 表示特征集$\{f_1,f_2,\cdots,f_h\}$，C 表示类特征，m 表示样本集的大小。在粗糙中 F、C 分别称为条件和决策属性集

输出：该样子集的约简特征子集 S

过程：

(1)如果数据有连续值的，先对样本集进行离散化处理；

(2)用可辨识矩阵计算 F 相对于 C 的核 CORE=$CORE_C(F)$，$S\leftarrow$CORE，$S'\leftarrow F-S$；

(3)如果 $S=\varnothing$，对 S' 中每特征计算 $r_{cf'}$，找出使 $r_{cf'}$ 最大的 f'_j，$S\leftarrow S\cup\{f'_j\}$，$S'\leftarrow F-S$；

(4)如果 $S\neq\varnothing$，计算当前 $RMI(f')$ 和 $U(S)$ 的值：

1)对 S' 中每特征计算 RMI，按 RMI 从大到小排列 f' 和对应的 $RMI(f')$；

2)计算当前候选特征子集 S 的不确定性系数 $U(S)$；

(5)'收缩'F：

如果有 f'_x 的 $RMI(f'_x)<t_1$，$F\leftarrow F-\{f'_x\}$；

(6)'拉伸'S：

将使 RMI 值最大的 f'_y 纳入 S，$S\leftarrow S\cup\{f'_y\}$；

（若同时有 n 个属性的 f'_y 的 RMI 达到最大，则分别计算 n 个 $U(S\cup\{f'_y\})$，从中选出使 U 最大的 f'_y，$S\leftarrow S\cup\{f'_y\}$；$S'\leftarrow F-S$；）

(7)重新计算当前 $RMI(f')$ 和 $U(S)$ 的值：

1)对 S' 中每特征计算 RMI，按 RMI 从大到小排列 f' 和对应的 $RMI(f')$；

2)计算当前候选特征子集 S 的不确定性系数 $U(S)$；

(8)判断是否结束的条件：

1)如果 $U_i-U_{i-1}>0$：

①如果 $U_i-U_{i-1}<t_1$，则 $S\leftarrow S-\{f'_y\}$，$F\leftarrow F-\{f'_y\}$，$S'\leftarrow F-S$；

②如果 $S'\neq\varnothing$，转到(4)，继续；否则输出 S，结束。

2)如果 $U_i-U_{i-1}<0$：

①如果 $abs(U_i-U_{i-1})>t_2$，则 $S\leftarrow S-\{f'_y\}$，输出 S，结束。

②如果 $S'\neq\varnothing$，转到(4)，继续；否则输出 S，结束。

其中有两个阈值，t_1、t_2，经实验，t_1 可取 0.1，t_2 可取 0.3。

从算法结构分析，在循环未结束前，每迭代一次，要么将 S' 中的特征 f' 加入 S，要么从 F 中去掉，直至 S' 为空，也就是 $F=S$，因此，该算法执行步数是有限的。下面将进一步用实验结果说明该算法的可行性与正确性。

5 实验及结果分析

为了验证本文提出的算法 MRMI-UC 的有效性，本文选择了 UCI (http://www.ics.uci.edu/~mlearn/MLRepository.html)中 5 个数据集对本文算法进行测试。分别给出了用本文算法 MRMI-UC 得到的特征选择结果与文[11]用完全搜索算法 ABB 算法得到的特征选择结果进行比较，同时，用机器学习算法的错误率来评估两种算法得到特征子集的优劣。其中分别用 Decision Table 和 Naïve-Bayes 的十折交叉验证法来验证被选择的特征子集。

表 1 为特征选择的结果，m 为样本数，p 为初始特征集大小，S 为特征子集，k 为该特征子集大小。分别给出了用本文算法 MRMI-UC 和文[11]用完全搜索算法 ABB 算法得到的特征选择结果。两种算法得到的特征子集个数相差无几。

表 1 特征选择的结果

Dataset	m	p	MRMI-UC		ABB	
			S	k	S	k
Monk1	432	6	$\{f_1,f_2,f_5\}$	3	$\{f_1,f_2,f_5\}$	3
Monk3	432	6	$\{f_2,f_5\}$	2	$\{f_2,f_4,f_5\}$	3
Vote	435	16	$\{f_1-f_6,f_{11},f_{12},f_{15}\}$	9	$\{f_1-f_4,f_9,f_{11},f_{13},f_{15},f_{16}\}$	9
Parity5+5	1024	10	$\{f_2-f_7\}$	6	$\{f_2-f_4,f_6,f_8\}$	5
Partity5+2	1024	10	$\{f_2,f_3,f_6\}$	3	$\{f_1-f_5\}$	5

表 2 和表 3 分别为用 Decision Table 和 Naïve-Bayes 两种分类算法十折交叉有效验证结果。表中 Before 和 After 分别表示用初始特征集和被选择的特征子集来进行学习。Error rate 指分类的错误率。其中在 After 项中还给出了本文提出的 MRMI-UC 和 ABB 算法得到的特征子集进行学习的结果。

表 2 基于 Decision Table 分类的十折交叉验证结果

Dataset	Before	After	
	Error rate(%)	MRMI-UC Error rate(%)	ABB Error rate(%)
Monk1	25.00	23.50	23.50
Monk3	2.77	2.57	2.57
Vote	5.91	5.52	5.10
Parity5+5	49.50	49.30	50.19
Partity5+2	49.80	49.78	50.20

表 3 基于 Naive Bayes 分类的十折交叉验证结果

Dataset	Before	After	
	Error rate(%)	MRMI-UC Error rate(%)	ABB Error rate(%)
Monk1	33.33	33.33	33.33
Monk3	7.64	4.39	4.39
Vote	8.05	5.51	6.91
Parity5+5	61.14	41.00	59.57
Partity5+2	61.82	56.45	60.35

ABB是特征选择算法中较为典型的完全搜索算法，耗时大；本文提出的MRMI-UC算法是启发式的特征选择算法，耗时小；从分类学习的Error rate看出，由MRMI-UC得到的结果并不比由ABB的差，结果相当。

从表中Befor和After项结果看出，经由本算法MRMI-UC特征选择后Decision Table和Naïve Bayes学习的错误率不比用初始特征集学习大，反而有些还低得多，特别是在Naive Bayes十折交叉验证中，对Parity5＋5分类的错误率由61.14%降到41.00%，降低了20%。这表明，用MRMI-UC进行特征选择数据处理，提高了Naive Bayes的学习正确率。

从两表After项中MRMI-UC和ABB算法特征选择后分类器学习错误率来看，由MRMI-UC得到的结果并不比由ABB得到的结果差，如Parity5＋5中由MRMI-UC得到的错误率41.00%比ABB得到的59.57%还低18个百分点。

实验结果表明，RMI来衡量特征间的相关程度是可行的，本文提出的基于相对互信息比和不确定性系数的特征选择算法是有关效的。

总结 本文基于粗糙集中关于非精确集和精确集理论思想，从信息论的角度来研究特征选择问题。通过研究非核属性与核，以及非核属性与类别属性间相关性关系的一些变化规律，提出了度量特征的一个新指标，即相对互信息比RMI，并由此，设计了一种基于粗糙集启发式特征选择算法。实验结果分析表明，本文提出的算法在多数情况下能够得到较优的特征子集，算法是有效的，切实可行的。但是，关于该算法对最小属性约简的完备性问题还需从理论上作进一步的探讨。

参考文献

1 Liu Huan, Motoda H. Feature Selection for Knowledge Discovery and Data Mining [M]. Kluwer Academic Publishers, 1998

2 Pawlak Z. Rough Sets-Theoretical Aspects of Reasoning about Data. Kluwer Academic Pub, 1991

3 张腾飞，肖健梅，王锡淮. 粗糙集理论中属性相对约简算法. 电子学报，2005，33(11)：2080～2083

4 Perkins C E. Ad Hoc Networking [M]. Chapter 3, ADDISON-WESLEY Press, 2000

5 王国胤. Rough集理论与知识获取[M]. 西安交通大学出版社，2001. 51

6 常犁云，王国胤，吴渝. 一种基于Rough Set理论的属性约简及规则提取方法. 软件学报，1999，10(11)：1206～1211

7 杨胜，胡福乔，施鹏飞. 一个新的特征选择判据——不确定性系数. 计算机工程，2004，30(8)：46～47

8 梁霖，徐光华. 基于克隆选择的粗糙集属性约简方法. 西安交通大学学报，2005，39(11)：1231～1235

9 Liu H, Motoda H, Dash M. A Monotonic Measure for Optimal Feature Selection. In: Proc. of ECML-98, 1998

10 于冰，阎保平. 关于粗糙集属性约简的进化算法研究和应用. 微电子学与计算机，2005，22(3)：189～194

11 Yang Sheng, GU Jun. Feature selection based on mutual information and redundancy-synergy coefficient. Journal of Zhejiang University SCIENCE, 2004, 5(11): 1382～1391

(上接第115页)

对流数据降载需要深入研究的工作有：

• 动态负载分布中负载均衡。现有的动态负载分布中负载均衡研究明显存在条件的假设，所以条件的放宽和放在更异构的环境是下一步的研究内容。因为在这种环境下，就会带来诸如带宽、能源消耗等资源的新问题，而已有的算法是否在新的环境下获得最优的结果值得深入研究。

• 自适应策略在DSMS中的应用。对于在DSMS中引入控制理论，在DSMS系统中加入质量控制器，只是涉及一些较简单的问题。对于更复杂流数据系统与环境进行建模和控制仍需进行研究，特别是分布式流数据管理的动态资源配置，是目前研究的热点。

• 全局控制。要保持低的等待时间，算法也必须确保查询结果的质量不能任意地降级。一个降载计划能在某个节点发送高质量的结果，但在全局水平上并不一定有相同的质量，所以，对单点的周期性控制要与全局的控制统一起来。

• 拓扑结构的一般化。Borealis系统中元数据的合并与传播工作是在基于树结构的拓扑服务器环境下进行的。在这种情形下，来自子节点的元数据在其父节点以增量模式合并、修改和向前传播。因此，对每个节点来说足可以和它直接邻近的上下节点通讯。对于更一般的拓扑结构环境下，一个节点可能有多个父节点，要完成全局的协作，不直接相邻的节点间也需通讯。在这种情况下，这些节点中一个节点的降载决策将会影响其它节点的决策。

参考文献

1 Motwani R, Widom J, Arasu A, et al. Query Processing, Resource Management, and Approximation in a Data Stream Management System. In: Proc. of CIDR 2003, Jan. 2003

2 Carney D, Cetintemel U, Cherniack M, et al. Monitoring streams: A new class of data management applications. In: VLDB, 2002

3 Krishnamurthy S, Chandrasekaran S, et al. TelegraphCQ: An Architectural Status Report. IEEE Data Engineering Bulletin, 2003, 26(1): 11～18

4 Tatbul N, Cetintemel U, Zdonik S, et al. Load shedding in a data stream manager. In: Proc. of the 29th Intl Conf. on Very Large Databases (VLDB'03), 2003

5 Babcock B, Datar M, Motwani R. Load shedding for aggregation queries over data streams. In: 20th International Conference on Data Engineering, 2004

6 Tu Yi-Cheng, Hefeeda M, Xia Yuni, et al. Control-based Quality Adaptation in Data Stream Management Systems. In: the Proc. of the International Conference and Workshop on Database and Expert Systems Applications (DEXA), August 2005

7 Chi Yun, Yu P S, Wang Haixun, et al. Loadstar: A load shedding scheme for classifying data streams. In: SIAM International Conference on Data Mining (SDM), 2005

8 Reiss F, Hellerstein J. Data Triage: An Adaptive Architecture for Load Shedding in TelegraphCQ. In: IEEE ICDE Conference, Tokyo, Japan, April 2005

9 Tatbul N, Zdonik S. Dealing with Overload in Distributed Stream Processing Systems. In: IEEE International Workshop on Networking Meets Databases (NetDB'06), Atlanta, GA, April 2006

10 Abadi D J, Carney D, Cetintemel U, et al. Aurora: a new model and architecture for data stream management. The VLDB Journal, 2003, 2(2): 120～139

11 Abadi D, Ahmad Y, Balazinska M, et al. The Design of the Borealis Stream Processing Engine. In: CIDR Conference, Asilomar, CA, January 2005

12 Xing Ying, Zdonik S, Hwang Jeong-Hyon. Dynamic Load Distribution in the Borealis Stream Processor. In: 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan, April 2005

13 Gupta D, Bepari P. Load sharing in distributed systems. In: Proc. of the National Workshop on Distributed Computing, January 1999

14 Shirazi B A, Hurson A R, Kavi K M. Scheduling and load balancing in parallel and distributed systems. IEEE Computer Science Press, 1995

15 Tu Yi-Cheng, Liu Song, Prabhakar S, et al. Load Shedding in Stream Databases: A Control-Based Approach. In: Proc. of the 32th Intl Conf on Very Large Databases, 2006