文章编号:1001 - 9081(2006)11 - 2554 - 04

一种选择性 GMDH 网络集成算法

曹 鹏,李金龙,张泽明,王煦法
(中国科学技术大学 计算机科学技术系,安徽,合肥 230027)
(caopeng@ mail. ustc. edu. cn)

摘 要:提出一种新的 GMDH 网络的选择性集成算法,通过对构造 GMDH 网络个体的训练样本进行惩罚性划分,产生具有差异性的 GMDH 网络集合,再利用遗传算法从中选择最优 GMDH 网络子集进行集成。实验结果与分析表明,与 GMDH 网络的整体集成和单个 GMDH 网络以及传统的 BP 神经网络集成相比,GMDH 网络的选择性集成在性能上具有明显的优势。

关键词:GMDH;惩罚性划分;选择性集成
中图分类号: TP18 **文献标识码**:A

A selective GMDH network ensemble algorithm

CAO Peng, LI Jin-long, ZHANG Ze-ming, WANG Xu-fa
(*Department of Computer Science and Technology, University of Science and Technology of China, Hefei Anhui* 230027, *China*)

Abstract: A selective GMDH network ensemble algorithm was presented. With the punitive classification of the samples for training GMDH network individuals, a group of candidate GMDH networks were developed which were different from each other. Genetic algorithm was then used to evolve the best subset of the candidates to form the ensemble. Experiments show that compared to all-candidates GMDH ensemble and GMDH network individuals as well as the traditional BP neural network ensemble, selective GMDH network ensemble improves the performance greatly.

Key words: Group Method of Data Handling (GMDH); punitive partition; selective ensemble

0 引言

GMDH(Group Method of Data Handling)是一种自组织的系统建模方法。该方法利用不完全归纳算法实现最优复杂度模型的自动选取[1]。与 BP 神经网络相比,GMDH 建模需要的训练样本数量少,模型结构不需预先设定[2]。但 GMDH 建模是一个确定性的过程并且基于对训练样本的划分。训练样本一经被划分为构造集合和选择集合后,该方法将沿着确定的方向构造网络模型。不同的划分将得到不同的 GMDH 网络[3],因此易于得到局部最优模型[4]。

Abdel-Aal 首次将集成学习方法[5]应用于 GMDH 网络[6,7],即首先在训练样本上产生一定数量的 GMDH 网络个体,对新样本的预测输出由所产生的所有 GMDH 网络个体共同决定。该方法虽然通过使用集成学习提高了系统的泛化能力,但是经验性设定 GMDH 网络的集成规模对集成预测性能的提高不一定是最优的。

本文提出一种选择性 GMDH 网络集成算法。通过对训练样本的一部分(个体训练样本)进行惩罚性划分,从而产生一组具有差异性的 GMDH 网络,再在其余训练样本(集成训练样本)上利用遗传算法选择候选 GMDH 网络集合中的最优组合,进而进行集成。实验结果和分析表明,与 GMDH 网络的整体集成和单个 GMDH 网络以及 BP 神经网络的集成相比,GMDH 网络的选择性集成在性能上有显著的提高;进一步的实验结果也验证了惩罚性划分在选择性 GMDH 网络集成中的有效性。

1 背景知识

1.1 GMDH 网络模型

GMDH 网络模型如图 1 所示,网络中每个节点都是形如(1)式的二次二项式。

$$y = a_1 x_i^2 + a_2 x_j^2 + a_3 x_i + a_4 x_j + a_5 x_i x_j + a_6 \quad (1)$$

x_i, x_j 分别表示来自于上层节点的输入,y 表示该节点的输出。GMDH 网络的构造基于外补充原理[1],即只有使用附加的外部信息才能够从给定的数据样本中筛选出最优复杂度模型,这种外部信息应该是在构造模型的过程中没有用到的。

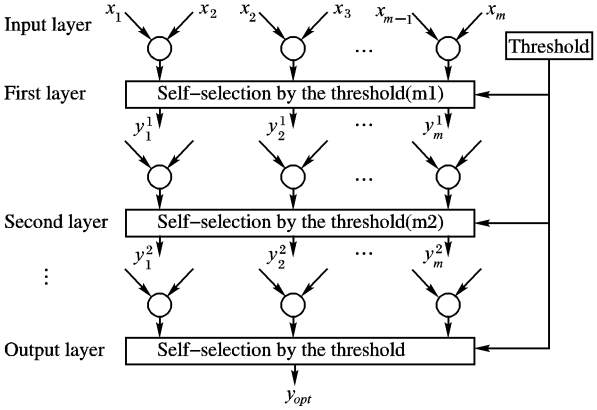

图 1 GMDH 网络模型

因此,GMDH 网络的个体训练样本 Ta 将被划分为构造集

收稿日期:2006 - 05 - 19;**修订日期**:2006 - 07 - 04

作者简介:曹鹏(1982 -),男,安徽合肥人,硕士研究生,主要研究方向:自组织建模、进化计算; 李金龙(1975 -),男,湖南怀化人,讲师,博士,主要研究方向:数据挖掘; 张泽明(1980 -),男,湖北天门人,博士研究生,主要研究方向:硬件进化; 王煦法(1948 -),男,江苏丹阳人,教授,博士生导师,主要研究方向:智能信息处理.

合 Tc 和选择集合 Ts。训练过程中不断有新的个体通过训练样本 Tc 产生,而新的个体是否为最优个体或是否可以参与下一层节点的构造是根据其在 Ts 上的表现而决定的。具体的构造过程见文献[2]。

1.2 集成学习

集成学习是一种利用多个学习器对同一问题进行学习的方法[5,8],这种思想最早由 Hansen 和 Salamon 提出并用于神经网络的集成学习[5]。他们证明,可以简单地通过训练多个神经网络并将结果进行合成,从而克服单个神经网络易陷入局部最优的不足,显著提高神经网络系统的泛化能力。典型的集成学习一般包括两个阶段,首先训练出一定数量的学习器个体,再将学习器的输出进行集成。训练样本也被相应地划分为个体训练样本和集成训练样本。Zhou 等人[9]分析并证明了在神经网络的集成学习中有选择性集成候选个体,在性能上优于候选个体的整体集成,并在此基础上提出了神经网络的选择性集成方法 GASEN[10]。

目前的研究工作表明个体之间的差异性对集成性能的提升有重要影响[11]。集成在新样本上输出的泛化误差可表示如下:

$$(f_{ens}-d)^2=\sum_i w_i(f_i-d)^2-\sum_i w_i(f_i-f_{ens})^2 \tag{2}$$

$$f_{ens}=\sum_i w_i f_i,\ \sum_i w_i=1 \tag{3}$$

其中,f_{ens} 为集成输出,f_i 为集成中个体的输出,d 为样本实际输出,w_i 为集成中个体对应的权值。(2)式表明了在集成中个体预测精度不变的前提下(等号右边第一项不变),彼此之间的差异性越大(等号右边第二项增大),则泛化误差越小,即集成的效果越好。Perrone 和 Cooper[8]认为集成中的个体陷入局部最优这一特性对集成泛化能力的提高起到重要的作用。因为如果个体互不相关,则在学习中很有可能陷入不同的局部最优,这样集成的差异度就会很大,各局部极小的负作用可以在集成中相互抵消。

2 选择性 GMDH 网络集成算法

本文提出的选择性 GMDH 网络集成算法分为 GMDH 网络个体训练阶段和选择性集成两个阶段,训练样本 T 被划分为个体训练样本 Ta 和集成训练样本 Tb。在第一阶段,通过对 Ta 进行惩罚性划分产生具有差异性的 GMDH 网络个体集合 S;在第二阶段,在 Tb 样本上使用遗传算法从集合 S 中选择出最优组合,进而进行集成。

2.1 GMDH 网络个体的训练

由于 GMDH 网络的节点系数都是在构造集合 Tc 上得到的,模型往往对该集合中的多数样本具有更高的拟合精度,而这也导致所建立的模型仅是局部最优的。

而集成中个体陷入不同的局部最优也是提高个体之间差异性的一种方法,因此对提高集成的性能是有利的。对此,为了使产生的候选 GMDH 网络之间彼此的差异性增大,我们提出一种惩罚性的样本划分方法:以个体训练样本在上一轮构造的 GMDH 网络的输出结果的平方误差为标准,定义如下:

$$e(x)=(f(x)-d(x))^2 \tag{4}$$

其中 $f(x)$ 和 $d(x)$ 分别为模型的计算输出和实际输出,进而对误差较大的训练样本进行惩罚以确定训练样本的新一轮划分。

算法 1　GMDH 网络训练样本的惩罚性划分算法

1)将 Ta 样本代入上一次构造得到的 GMDH 网络,并计算得到每个样本的平方误差 $e(x_i)$, $x_i\in Ta$;

2)对 Ta 中的样本按照 $e(x_i)$ 从小到大进行排序,得到 T'_a;

3)取 T'_a 中前 $|T'_a|/2$ 作为构造下一个 GMDH 网络的构造集合,后 $|T'_a|/2$ 作为选择集合。

通过对个体训练样本进行惩罚性划分使得当前 GMDH 网络预测较差的一部分样本直接参与下一个 GMDH 网络的构造,而当前被较好进行预测的样本将作为构造下一个 GMDH 网络的外部信息参与选择。这样使得 GMDH 方法通过沿不同的方向构造以使得到的模型能够更好地拟合不同的训练样本,从而增大 GMDH 网络之间的差异性。在集成过程中,这种差异性可以被用来提高集成的性能。

2.2 GMDH 网络个体的选择性集成

根据式(2),集成中的个体应该具有较强的泛化能力且彼此之间差异性较大。如何从候选 GMDH 网络集合中选出泛化能力强而彼此之间差异性较大的最优子集是一个组合优化问题。Zhou 在神经网络集成中采用遗传算法较好地解决了候选集合最优子集的选取问题[9]。因此本文也采用同样的方法,在产生一组差异性的 GMDH 网络个体后,利用遗传算法对候选 GMDH 网络进行选择性集成,以提高 GMDH 网络集成的性能。遗传算法的初始种群中每个染色体定义为 $1\times n$ 的向量,其中,n 为候选 GMDH 网络的个数,向量中的每一维均是在[0,1]上均匀分布的随机数,表示对应的 GMDH 网络的权值。适应度函数 f 定义如下:

$$f=\frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{N}w_i w_j C_{ij}} \tag{5}$$

其中 w_i 表示对应 GMDH 网络个体的权值,C_{ij} 表示第 i 和第 j 个 GMDH 个体的相关性度量值:

$$C_{ij}=\sum_{x\in Tb}(f_i(x)-d(x))(f_j(x)-d(x)) \tag{6}$$

$d(x)$ 为 x 的实际输出,$f_i(x)$ 为第 i 个 GMDH 网络个体对 x 的计算输出。进化操作中的交叉算子、变异算子等参数参照 GAOT Toolbox[12] 中的默认设置。进化的目标是得到候选 GMDH 网络个体对应的最优权值 ω_{best}。

在得到最优染色体 ω_{best} 后,对其进行归一化操作得到 ω。定义阈值 λ,若 $\omega(i)>\lambda(1\leqslant i\leqslant 20)$,则第 i 个候选 GMDH 网络将被选入集成。

2.3 GMDH 网络的选择性集成算法

算法 2　GMDH 网络选择性集成算法

1)训练样本 T 平均划分为两部分:Ta 和 Tb,其中 Ta 用于个体的生成,Tb 用于遗传算法中得到最优子集,定义候选个体数量 $size$;

2)将 Ta 随机划分为等规模的两部分 $Ta1$ 和 $Ta2$,$N=0$;

3)使用 $Ta1$ 和 $Ta2$ 作为构造集合和选择集合训练得到 GMDH 网络个体 $indivN$;

4)$N=N+1$,若 $N=size$,转 6);

5)将 Ta 代入 $indivN$,利用算法 1 得到 Ta 的一个新的划分,转 3);

6)使用遗传算法进化得到候选 GMDH 网络的权值,并通过该权值选定最优子集并建立集成,这里定义阈值 $\lambda=1/size$。

集成中的每个个体在进化过程中都有对应的权值,但使

用集成中个体输出的加权平均有可能导致过拟合[9],因此在将集成用于新样本时,集成的输出为集成中的每个个体输出的简单平均。

3 实验结果和分析

3.1 样本描述

为验证选择性 GMDH 网络集成的有效性,参照文献[9,14],我们使用了两个数据集合进行验证,分别是 friedman1 和 friedman2[9],各数据集合具体描述如表 1。

表 1 Friedman1,Friedman2 样本描述

Data Set	Function	Variable
Friedman1	$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$	$x_i \in U[0,1]$
Friedman2	$y = \sqrt{x_1^2 + (x_2 x_3 - 1/x_2 x_4)^2}$	$x_1 \in U[0,100]$, $x_2 \in U[40\pi, 560\pi]$, $x_3 \in U[0,1]$, $x_4 \in U[1,11]$

参照文献[13]的设置,Friedman1 和 Friedman2 两个数据集合中样本总数均为 1400,其中前 400 个样本带有正态分布 $N(0,1)$ 的随机噪声用于训练和集成过程,后 1000 个作为验证集成效果的测试样本。采用平均相对方差(Average Relative Variance, ARV)作为性能评价标准,定义如下:

$$ARV = \sum_{i=1}^{N}(y_i - f(x_i))^2 / \sum_{i=1}^{N}(y_i - \bar{y})^2 \quad (7)$$

其中 y_i 为每个样本的实际输出值,$f(x_i)$ 为样本的集成计算输出值,$\bar{y}$ 为样本实际输出的均值,$N = 1000$,表示样本个数。

3.2 实验结果

我们在每个数据集上进行 5 项实验,通过计算模型在测试数据集上的 *ARV* 值对各模型的泛化能力进行比较。第 1 项实验产生基于惩罚性样本划分的选择性 GMDH 网络集成(sel-GMDH-ensemble),即本文所提方法。实验中,首先通过对个体训练样本进行惩罚性划分,产生 20 个候选 GMDH 网络,再在集成训练样本上使用遗传算法选定一组候选个体建立集成。第 2 项实验产生 GMDH 网络个体(GMDH individual)。该实验中,400 个训练样本被随机进行划分,从而获得 GMDH 方法中的构造集合和选择集合,并用于构造得到单个 GMDH 网络。第 3 项实验产生 GMDH 网络的整体集成(all-GMDH-ensemble)。由于整体集成不需要对候选个体进行选择,因此在每一次实验中,400 个训练样本不再进行划分,而是全部用于产生 GMDH 网络个体,每次实验均将产生的 20 个 GMDH 网络作整体集成。第 4 项产生选择性 BP 神经网络集成,我们直接调用 GASEN 算法[10],并将结果与本文方法进行比较。由于 Zhou[9] 证明了与神经网络的整体集成相比,选择性集成具有更精确的预测能力,因此这里对神经网络的整体集成实验不再重复。以上 4 个实验分别独立运行 20 次。表 2 给出了每一项实验在测试数据集上的平均相对方差的结果比较。

第 5 项实验中,我们将实验一的惩罚性样本划分替换为随机划分,以验证惩罚性划分在选择性 GMDH 网络集成中的有效性,该实验同样独立运行 20 次。表 3 中给出了两种方法在测试数据集上的比较结果。

表 2 平均相对方差结果比较

Dataset	test	result					
		best	worst	mean	median	std	N
Friedman1	sel-GMDH-ensemble	0.0495	0.0911	0.0700	0.0688	0.0123	4.35
	single GMDH	0.0847	0.1329	0.1051	0.1031	0.0150	1
	all-GMDH-ensemble	0.0861	0.0999	0.0933	0.0938	0.0037	20
	GASEN	0.0530	0.1893	0.1082	0.1087	0.0315	3.55
Friedman2	sel-GMDH-ensemble	6.7418e-004	8.7411e-004	7.8065e-004	7.9012e-004	6.7513e-005	5.35
	single GMDH	8.4464e-004	0.0015	8.9672e-004	8.6058e-004	1.4527e-004	1
	all-GMDH-ensemble	0.0014	0.0015	0.0015	0.0015	2.9741e-005	20
	GASEN	0.0284	0.5888	0.2556	0.2419	0.2194	2.45

表 3 选择性 GMDH 网络集成中惩罚性划分与随机划分的平均相对方差结果比较

Dataset	test	result					
		best	worst	mean	median	std	N
Friedman1	punitive	0.0495	0.0911	0.0700	0.0688	0.0123	4.35
	random	0.0637	0.1098	0.0843	0.0831	0.0120	4.65
Friedman2	punitive	6.7418e-004	8.7411e-004	7.8065e-004	7.9012e-004	6.7513e-005	5.35
	random	8.1131e-004	0.0015	0.0011	9.4245e-004	2.9010e-004	3.30

在表 2 和表 3 中,best 和 worst 为独立运行 20 次实验中的最好值和最差值,mean, median 和 std 分别为 20 次试验中的均值、中值和方差,N 为集成中个体的数量。从表 2 的实验结果可以看出,GMDH 网络个体表现出泛化能力不强且稳定性较差的问题,而相对于 GMDH 网络个体,GMDH 网络的选择性集成能够提高在未知样本上的泛化能力并且增强稳定性。与 GMDH 网络的整体集成相比,GMDH 网络的选择性集成的最好值和均值都有显著提高。这是由于整体集成中并不是每个个体都对集成性能的提高有贡献,相反可能由于某些个体的存在导致了集成性能的下降。比较的结果也说明了本文提出的选择性 GMDH 网络集成方法能够进一步提高 GMDH 集成系统的泛化能力。与 GASEN 相比,选择性 GMDH 网络集成在泛化能力和稳定性上都具有明显的优势,因此可以认为 GMDH 网络作为一种集成中的个体形式能够更好地发挥集成的优势,值得进一步的研究。从表 3 给出的两种 GMDH 网络选择性集成实验结果的比较中可以看出,在各项指标上惩罚性划分都优于随机划分,这也进一步证明了本文提出的惩罚性样本划分的有效性。

4　结语

GMDH 方法建立的网络模型易收敛于局部最优,表现出泛化能力和稳定性上的不足。集成学习能够利用个体局部最优的特性,通过将个体的输出进行集成以提高泛化能力和稳定性。本文对构造 GMDH 网络的样本进行惩罚性划分,使 GMDH 沿不同的方向构造 GMDH 网络,并在此基础上使用选择性集成方法,从候选 GMDH 网络中选出最优子集作为集成。实验结果表明 GMDH 网络的选择性集成能够在克服单个 GMDH 网络泛化能力和稳定性不足的基础上进一步提高集成的效果,而与 BP 神经网络集成的比较结果也说明了 GMDH 网络作为一种集成的个体形式能够更好地发挥集成的优势,值得作进一步研究。

参考文献:

[1] MUELLER JA, LEMKE F. Self-organizing Data Mining[M]. Berlin: Libri Books, 1999.

[2] IVAKHNENKO AG. Polynomial Theory of Complex System[J]. IEEE Transactions on System, Man and Cybernetics, 1971, SMC-1(4):364-378.

[3] MORT LN. The Development of Self Organization Techniques in Modeling: A Review of the Group Method of Data Handling (GMDH)[EB/OL]. http://www.shef.ac.uk/content/1/c6/03/23/48/813.pdf, 2001.

[4] KIM D, PARK G-T. GMDH-Type Neural Network Modeling in Evolutionary Optimization[A]. Proceeding of 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE 2005)[C], 2005. 563 - 570.

[5] HANSEN LK, SALAMON P. Neural Network Ensemble[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(10):993-1001.

[6] ABDEL-AAL RE. Improved Classification of Medical Data Using Abductive Network Committees Trained on Different Feature Subsets[J]. Computer Methods and Programs in Biomedicine, 2005, 80(2):141-153.

[7] ABDEL-AAL RE. Improving Electric Load Forecasts Using Network Committees[J]. Electric Power Systems Research, 2005, 74(1):83-94.

[8] PERRONE M, COOPER L. When networks disagree: Ensemble method for neural networks[A]. Artificial Neural Networks for Speech and Vision[M]. New York: Chapman & Hall, 1993. 126-142.

[9] ZHOU Z, WU JX, TANG W. Ensembling Neural Networks: Many could be Better than All[J]. Artificial Intelligence, 2002, 137(1/2):239-263.

[10] LAMDA GROUP(Nanjing University). GASEN toolbox[CP/OL]. http://lamda.nju.edu.cn/datacode/GASEN/gasen.htm, 2006-02-16.

[11] BROWN G, WYATT J, HARRIS R, *et al*. Diversity Creation Methods: A Survey and Categorization[J]. Information Fusion, 2005, 6(1):5-20.

[12] North Carolina State University. GAOT[CP/OL]. http://www.ie.ncsu.edu/mirage/GAToolBox/gaot/gaotv5.zip, 1998-06-12.

[13] 吴建鑫,周志华,沈学华, 等. 一种选择性神经网络集成构造方法[J]. 计算机研究与发展,2000, 37(9).

[14] CHANDRA A, YAO X. Ensemble Learning Using Multiobjective Evolutionary Algorithms[J/OL]. Journal of Mathematical Modeling and Algorithms, 2006-03.

(上接第 2543 页)

[16] LI Z, ZHANG Z, WANG L. A novel QoS routing scheme for MPLS traffic engineering[A]. Proceedings of International Conference on Communication Technology[C], 2003, 1.

[17] SZETO BW, IRAQI Y. DORA: Efficient Routing for MPLS Traffic Engineering[J]. Journal of Networks and Systems Management, 2002, 10(3).

[18] ELWALID A, JIN C, LOW S, *et al*. MATE: MPLS Adaptive Traffic Engineering[A]. INFOCOM2001[C], 2001.

[19] CUI B, YANG Z, DING W. A Load Balancing Algorithm Supporting QoS for Traffic Engineering in MPLS Netwoks[A]. The Fourth International Conference on Computer and Information Technology[C], 2004.

[20] LONG K, ZHANG Z, CHENG S. Load balancing algorithms in MPLS traffic engineering[A]. 2001 IEEE Workshop on High Performance Switching and Routing[C], 2001.

[21] LIM SH, YAACOB MH, PHANG KK, *et al*. Traffic engineering enhancement to QoS-OSPF in DiffServ and MPLS networks[J]. IEE Proceedings-Communications, 2004, 151(1).

[22] LEE Y, SEOK Y, CHOI Y, *et al*. A Constrained Multipath Traffic Engineering Scheme for MPLS Networks[EB/OL]. http://networks.cnu.ac.kr/publication/icc.pdf, 2006.

[23] CHO HY, LEE JY, KIM BC. Multi-path Constraint-Based Routing Algorithms for MPLS Traffic Engineering[A]. IEEE International Conference on Communications[C], 2003, 3.

[24] HASKIN D, KRISHNAN R. A Method for Setting an Alternative Label Switched Paths to Handle Fast Reroute[EB/OL]. draft-haskin-mpls-fast-reroute-05.txt, 2006.

[25] SALVADORI E, BATTITI R, ARDITO F. Lazy Rerouting for MPLS Traffic Engineering[EB/OL]. http://eprints.biblio.unitn.it/archive/00000368/01/011.pdf, 2006.

[26] YETGINER E, KARASAN E. Robust Path Design Algorithms for Traffic Engineering with Restoration in MPLS Networks[A]. Seventh International Symposium on Computers and Communications[C], 2002. 933-938.

[27] YOON S, LEE H, CHOI D, *et al*. An Efficient Recovery Mechanism for MPLS-based Protection LSP[A]. Joint 4th IEEE International Conference on ATM (ICATM 2001) and High Speed Intelligent Internet Symposium[C], 2001. 75-79.

[28] DOGAR FR, UZMI ZA, BAQAI SM. CAIP: a restoration routing architecture for DiffServ aware MPLS traffic engineering[A]. Proceedings of 5th International Workshop on Design of Reliable Communication Networks[C], 2005.

[29] AMIN M, KIN-HON H, PAVLOU G, *et al*. Improving survivability through traffic engineering in MPLS networks[A]. Proceedings of 10th IEEE Symposium on Computers and Communications[C], 2005. 758-763.

[30] BARTOS M, RAMAN R. A Heuristic Approach to Service Restoration in MPLS Networks[A]. IEEE International Conference on Communications[C], 2001, 1. 117-121.

[31] BREMLER-BARR A, AFEK Y, KAPLAN H, *et al*. Restoration by Path Concatenation: Fast Recovery of MPLS Paths[EB/OL]. http://www.cs.tau.ac.il/~natali/PODCMPLS.ps, 2001.

[32] SUBRAMANIAN S, MUTHUKUMAR V. Alternative Path Routing Algorithm for Traffic Engineering[A]. The Proceedings of the 15th ICSENG 02[C], 2002.