用页组拓扑平均距离改善页面聚类算法*)

林文龙　刘业政　余智学
(合肥工业大学电子商务研究所　合肥 230009)

摘　要　提出一种支持站点结构优化的页面聚类改进算法，通过引入图论中的拓扑平均距离，量化评估与挖掘站点结构中访问效率较低的内容文档集合为结构优化的兴趣页组，挖掘的页组具有更高的兴趣性，并将兴趣页组挖掘算法融入到拓扑优化算法中。实验结果表明改进算法能更好地优化站点结构，较一般算法收敛性好。

关键词　Web使用挖掘，页面聚类，频繁访问页组，自适应站点

Enhanced Algorithm for Page Clustering by Using Topology Average Distance of Web Pages Group

LIN Wen-long　LIU Ye-zheng　YU Zhi-xue
(Institute of E-Business, Hefei University of Technology, Hefei 230009, China)

Abstract　An enhanced algorithm which supports Website structure optimization was proposed for page clustering. A quantitative criteria was proposed by introducing the average distance in graph and the low access efficiency Web content pages group was discovered as interesting page group for Website structure optimization. Thus, the enhanced algorithm can find out more interesting page groups than the normal algorithm. Meanwhile the mining algorithm was integrated into the topology optimization algorithm. Experiment results show that the enhanced algorithm can improve Website structure better and it converges more rapidly.

Keywords　Web usage mining, Page clustering, Frequently visited page group, Self-adaptive Website

Web使用挖掘(Web Usage Mining)是指应用数据挖掘技术从Web使用信息中抽取兴趣模式的过程[1,2]。其中一种常见的兴趣模式是通过对站点页面聚类形成的兴趣页组，进行浏览导航服务、优化站点结构[3]等。根据挖掘目的的不同，目前存在一些兴趣页组抽取方法，其中用于站点结构优化目的的兴趣页组抽取算法主要存在以下两个问题有待改进：

1)缺乏一组能有效揭示站点结构不合理处的量化指标。文献[4]将经常被用户同时访问的页面即用户频繁访问页组作为兴趣页组，对于站点结构优化来说，频繁访问页组是有趣的，但简单地将频繁访问页组作为结构优化的兴趣页组的方式缺乏有效性与针对性。文献[5]的兴趣页组抽取算法则只抽取页面间没有链接的频繁访问页组作为兴趣页组，挖掘的页组具有很高的兴趣性，但该方法会遗漏大部分潜在的兴趣页组。文献[6]认为频繁访问页组内包含的超链接个数越少，则页组越有趣。由于用户在一组兴趣页面之间的游历是通过整个站点的超链导航体系来完成的，因此仅采用页组内的超链接个数作为指标虽然能在一定程度上刻画站点拓扑结构中的不合理之处，但缺乏全局考虑，效果有限。

2)兴趣页组抽取算法没有与拓扑结构优化算法有机集成。抽取的兴趣页组与站点结构具有相关性，这要求兴趣页组的抽取与拓扑结构的优化是一个动态集成的过程。

上述问题的存在使得已有的兴趣页组抽取算法并不能很好地适用于站点拓扑结构优化。据此，本文提出一种兴趣页组抽取的改进算法，通过引入图论中的拓扑平均距离改善兴趣页组抽取算法的性能与效果，并将兴趣页组抽取算法集成到结构优化算法中。实验分析的结果表明，与已有的方法相比，该方法能更有效地支持站点拓扑结构优化，较一般算法收敛性好。

1　兴趣页组挖掘算法的改进

1.1　页组拓扑平均距离

定义1(Web站点图结构模型)　Web站点可以表示为无权有向图$WS=(P,E)$，其中P为站点页面集合，E为页面之间的超链接集合。对于任意的节点$p_i\in P$，其意义通常有以下三种情况：p_i代表网站内包含可访问内容的内容页面，称p_i为内容节点；p_i代表网站内只起导航作用的纯导航页面，称p_i为导航节点；p_i代表网站内既包含可访问内容又起导航作用的复合页面，则称p_i为复合节点。视复合节点为一导航节点和一个与其所包含的可访问内容相对应的内容节点的复合体，记包括复合节点复合体中的内容节点在内的所有内容节点的集合为P_C。

定义2(访问代价)　设有节点$p_i,p_j\in P(i\neq j)$，则用户通过节点p_i访问p_j的访问代价定义为有向图WS上p_i,p_j间的最短路径长度，记为$d(p_i,p_j)$。若有向图WS上不存在节点p_i,p_j间的路径，则定义$d(p_i,p_j)=\tau$，τ为一个较大的常数，表示用户通过p_i访问p_j需要通过除点击超链接外的其它途径(如在浏览器中输入p_j的URL地址)的访问代价当量。

通常利用兴趣性评价Web日志挖掘的结果[7]，在传统的页面聚类算法中[4,6]，页组的兴趣度反映为其出现的频繁程度，即支持度。页组支持度$Support(PG)$是指包含页组中所有页面的用户会话的个数。由于在WWW浏览中，用户通常

*)国家自然科学基金项目(70672097)，国家自然科学基金重点项目(70631003)。**林文龙**　博士生，主要研究方向为Web挖掘；**刘业政**　教授，博导，主要研究方向为数据挖掘。

需要通过对特定的文档集合进行浏览以获取所需的信息，另一方面，由于以超链形式组织的 Web 信息体系结构的复杂性与零乱性以及 WWW 信息提供者预先设计的超链体系结构与用户实际访问行为模式的差异性，使得网站用户在访问到其兴趣文档集合中的一个兴趣页面后，通常会通过一条需要花费更多访问代价的路径才能到达其兴趣文档集合的其它页面，因此本文认为一个能有效支持站点结构优化的兴趣页组须满足 3 个条件：(1) 被大量用户访问；(2) 只包含用户访问的兴趣内容页面；(3) 兴趣内容页面间的平均访问代价较大。其中条件(1)可以用页组支持度来量化，条件(2)可以通过一些启发式规则来判定，如文献[8]中的参考时间长度法、参考页面链接数目法，下面通过页组拓扑平均距离量化条件(3)。

拓扑平均距离的概念来源于图论的思想[9,10]，它是有机化学中定量研究有机化合物构造关系的一个十分成功的工具，在复杂网络的分析和设计中也有广泛的应用。本文将其引入到 Web 站点有向图中[11,12]，用以量化评估站点超链体系结构，并指导兴趣页组的挖掘算法。

定义 3(页组拓扑平均距离) 设有页组 $PG \subseteq P_C$，定义其拓扑平均距离(访问代价)为：

$$\mu(PG)=\begin{cases}\dfrac{\sum_{p_i,p_j\in PG} f(p_i,p_j)\cdot d(p_i,p_j)}{|PG|(|PG|-1)} & |PG|>1\\ 0 & |PG|=1\end{cases} \quad (1)$$

其中，$|PG|$ 为页组 PG 所包含的页面数，$f(p_i,p_j)$ 为群体用户通过节点 p_i 访问 $p_j(i\neq j)$ 的频度。

页组拓扑平均距离刻画了用户从页组内的一个兴趣页面访问另一个兴趣页面平均访问代价，因此比文献[6]中的组内链接度更加合理，可以作为站点结构合理化程度的指标。若用户经常同时访问的兴趣内容页面集合的页组拓扑平均距离较大，则表明现有的站点超链体系结构不能很好地适应用户的访问模式，站点结构有待优化；反之则表明现有的站点结构可以方便用户在频繁访问页组的页面之间进行游历，这样的页组就没有必要出现在挖掘结果中。

定义 4(页组拓扑兴趣度) 页组 PG 的拓扑兴趣度定义为：

$$Interest(PG)=1-\frac{1}{\mu(PG)} \quad (2)$$

定义 5(兴趣页组) 设页组 $PG\subseteq P_C$，若 PG 满足：(1) $Support(PG)\geqslant \min\_support$，(2) $Interest(PG)\geqslant \min\_interest$，则称为兴趣页组，其中 min_*support* 为最小支持度阈值，min_*interest* 为用户预定的最小拓扑兴趣度阈值。

综上所述，兴趣页组代表了 Web 站点中访问效率比较低的重要用户访问模式，因此可以用它作为站点结构优化的依据，使调整后的站点结构更适应站点用户的实际访问模式。

1.2 改进算法

改进算法主要从以下两个方面对传统的页面聚类算法进行改进：

(1)传统的页面聚类算法挖掘的对象是网站的所有页面，改进算法预先通过启发式规则将站点的页面分成内容页和导航页，缩减了挖掘对象的规模；

(2)在传统的页面聚类算法中，只采用支持度对候选兴趣页组集进行剪枝，改进算法则采用了支持度、拓扑兴趣度二组阈值对候选集进行剪枝，因此，改进算法可尽早地过滤掉兴趣性低的页组，使得后面迭代过程中的数据量提前减小，快速收敛；另外引入的拓扑兴趣度阈值，使改进算法抽取的兴趣页组具有较好的可理解性与可用性。由于引入拓扑兴趣度，改进算法的计算量要比文献[4]算法大，但是对于页面间的距离计算可以在预处理阶段完成，且可保留直到下一次站点内容的更新，在一定程度上减少了改进算法的计算量。

挖掘兴趣页组的改进算法是一个递归的过程，假设 FG_k 是包含 k 个页面的页组的集合，其中每个页组的支持度、拓扑兴趣度都大于预先设定的阈值，首先将 PG_1 初始化为支持度大于 min_*support* 的页面，PG_2 是在 PG_1 基础上产生的，PG_3 又是在 PG_2 基础上产生的，依此类推。该过程可以用算法 1 描述如下。

算法 1 PG_Generation

输入：站点拓扑结构，访问会话

输出：兴趣页组

过程：

```
1. classify the Web pages of WS into Pc or not based on a heuristic rule
2. initialize PG1 as the top requested single content page groups with
Support>=min_support;
3. for (i=2; i<=; i++) {
4.      Sort the pages of groups in PGi-1 in lexicographical order;
5.      for each group {x1,…,xi-1} in PGi-1{
6.           for each group {y1,…,yi-1} in PGi-1{
7.                if (x2=y1 and …and xi-1=yi-2) {
8.                     construct a new group G={x1,…,xi-1,yi-1};
9.                     if (G not already in CGi) {
10.                         test all other combinations of subgroups of
G with size (i-1);
11.                         if (all such subgroups are in PGi-1) {
12.                              if (Support(G)>? min_support and
Interest(G) >= min_interest ){
13.                                   add G into PGi;
14. } } } } } } }
```

由于兴趣页组的抽取与站点结构密切相关，我们将兴趣页组挖掘算法集成到结构优化算法中，同时为了避免破坏站点原有的超链导航体系，我们只考虑在原有的站点结构上增加一组合适的超链接对站点结构进行优化，集成兴趣页组挖掘算法的站点结构优化算法如算法 2 所示。算法中止准则采用的判断条件是：新增超链接个数超过网站管理员预定的最大新增超链接数 max_*number* 或从已有的站点结构中抽取的兴趣页组为空或搜索迭代次数超过预定阈值。

算法 2 WS_Optimization

输入：初始的站点拓扑结构 WS，访问会话 max_number，最大迭代代数

输出：优化后的站点拓扑结构

过程：

```
1. Set current optimal Website topology O(WS) = WS;
2. Callp G_Generate and output the interest Web pages groups;
3. If the output interest Web pages groups is empty, go to step10,
else, go to step4;
4. Sort the interest Web pages groups by their interest degree;
5. Add a group of new hyperlinks in the most interesting Web pages
group;
6. Update the parameter of max_number;
7. If max_number>0, go to Step 8. Otherwise, go to Step 10;
8. Replace current optimal Website topology O(WS) = WS* ;
9. Iterate the optimizing process. If no better result appears after re-
peating specified iteration time, go to Step 10, otherwise, go to step 2;
10. Stop and output the current optimal Website topology.
```

2 仿真实验

我们在 Windows2000 平台上用 MATLAB7.0 实现了本文的算法和文献[4-6]中的算法，并进行了对比。为了便于比较，文献[5]中的页面聚类算法的具体实现方法采用本文的算法 1 框架，而未采用原文的从图中挖掘子图的方法。

我们利用文献[13]中的站点结构仿真算法与 Web 日志仿真算法生成了一个有 100 个网页节的网站及其相应的访问会话作为实验的数据集。仿真的会话数据经去除长度大于 20 的过长会话与长度小于 3 的过短会话后，共得到 1356 条有效会话，平均访问步长为 9.3 步。我们从这 1356 条会话中随机抽取了 1156 条会话作为兴趣页组抽取算法与结构优化算法的学习数据集，余下的 200 条会话作为测试数据集。实验的算法中判断某一页面 p_i 是否属于 P_C 的启发式规则为文献[8]中的参考页面链接数目法。我们将链接数阈值取为 5，访问代价当量 τ 取为 5，做了三个实验。

实验一首先对兴趣页组挖掘算法的收敛性做了比较，本文改进算法的拓扑兴趣度阈值取为 0.7，实验结果如表 1 所示，表中 $|PG_i|$ 表示 PG_i 包含页组的个数。通过分析迭代过程中 $|PG_i|$ 数目的变化可以看出，文献[4]中的 SpeedTracer 算法收敛性最差，文献[5]中的 PageGather 次之，而本文的改进算法迭代收敛的速度最快，提高了计算过程的效率，其原因在于 SpeedTracer 只使用页组支持度对候选兴趣页组集合进行剪枝，PageGather 则只抽取页面间没有链接的频繁访问页组作为兴趣页组，文献[6]的改进算法则通过引入组内链接度参数改进支持度的计算公式（由于我们没有对网页大小进行仿真，此处以及下面的实验二我们忽略该文献中的范化内容链接比参数），从而提高了算法的收敛性，而本文算法的快速收敛性则主要得益于挖掘对象规模的缩减以及同时采用了支持度、拓扑兴趣度二组阈值对候选兴趣页组集合进行剪枝。

表 1　兴趣页组挖掘算法的实验结果比较

Algorithm used	min_su ort	$\|PG_1\|$	$\|PG_2\|$	$\|PG_3\|$	$\|PG_4\|$	$\|PG_5\|$	$\|PG_6\|$	$\|PG_7\|$	$\|PG_8\|$
SpeedTracer[1]	30	76	240	314	224	99	28	4	0
PageGather[2]	30	76	240	51	4	0	0	0	0
	20	94	434	383	204	60	8	0	0
	10	95	513	503	231	60	8	0	0
Enhanced[3]	30	76	240	8	0	0	0	0	0
	20	94	434	15	0	0	0	0	0
	10	95	513	91	4	0	0	0	0
Enhanced[4]	30	47	16	12	4	0	0	0	0
	20	65	59	37	10	0	0	0	0
	10	66	76	44	12	0	0	0	0

注：1 文献[4]的算法，2 文献[5]的算法，3 文献[6]的改进算法表，4 本文的改进算法，5 非集成的优化算法，6 集成的优化算法。

实验二比较了测试会话在优化前后的网站结构上的访问代价差的平均值，其中测试会话在当前网站结构下的访问代价取为用户从该会话的起点出发，逐一到达该会话所包含的目标内容页面的访问代价总和。结构优化的方法采用非集成的优先分配方法：即先由各种兴趣页组挖掘算法挖掘出兴趣页组，然后按兴趣度从高到低依次在兴趣页组内添加新超链接，当新增超链接的条数达到 max_*number* 时优化算法中止。文献[4]页组的兴趣度取为页组的支持度，本文算法的页组兴趣度取为页组的拓扑兴趣度值，实验结果如表 2 所示。从表中可以看出，本文的算法在结构优化中显著效果，其原因在于本文改进的兴趣页组抽取算法抽取的是群体用户访问代价比较高的内容页面的集合，而这类的兴趣页组对站点结构的优化具有很强的针对性，而文献[6]的改进算法的效果不甚理想，其原因可能在于我们忽略了该文献中的范化内容链接比参数，使得挖掘的兴趣页组中包含了太多的导航页面。另外我们从表中可以观察到几个有趣的结果：一是我们原以为 PageGather 方法的优化效果应该要比 SpeedTracer 好，可是实验结果却恰恰相反，二是 PageGather 方法当 max_*number*＝20 和 max_*number*＝30 的优化效果都是 0.31，为了分析其中的原因，我们着重分析并对比了这两种方法的新增超链接组，结果发现 PageGather 方法的很多新增超链接是添加在已有站点中不存在链接的导航页面之间，因此我们认为其原因可能是由于 PageGather 方法挖掘的对象是所有的页面，而把用户经常同时访问的导航页面作为兴趣度很高的兴趣页组，而实际上这类在导航页面之间新增的超链接是无益于用户从他感兴趣的一个内容页面访问其它兴趣内容页面的，这也是 PageGather 方法当 max_*number*＝10 时效果很差（仅为 0.09），当 max_*number*＝20 和 max_*number*＝30 时的优化效果没有变化的真正原因。

表 2　不同兴趣页组挖掘算法下的站点结构优化效果比较

max_number / Algorithm used	10	20	30	40	50	60	70	80	90	100
SpeedTracer[1]	0.24	0.40	0.51	0.71	0.92	1.14	1.30	1.40	1.50	1.60
PageGather[2]	0.09	0.31	0.31	0.46	0.51	0.64	0.87	1.00	1.18	1.32
Enhanced[3]	0.30	0.40	0.57	0.61	0.67	0.73	0.80	0.97	1.10	1.36
Enhanced[4]	0.34	0.54	1.00	1.23	1.44	1.66	1.84	1.98	2.14	2.57

实验三对兴趣页组的挖掘算法与结构优化算法集成和不集成的优化效果做了比较。实验中采用的挖掘算法为本文的改进的兴趣页面挖掘算法，最大迭代代数取为 200 代，实验结果如表 3 所示。从表中可以看出，集成的优化方法对优化效果有一定的改善，其原因在于集成的优化方法在优化过程中挖掘的兴趣页组能更实时反映出当前站点结构的不合理处。

表 3　集成优化与非集成优化的效果比较

max_number / Algorithm used	10	20	30	40	50	60	70	80	90	100
Without Combined[5]	0.34	0.54	1.00	1.23	1.44	1.66	1.84	1.98	2.14	2.57
Combined[6]	0.38	0.60	1.12	1.43	1.56	1.78	1.99	2.23	2.41	2.66

结束语　随着因特网的快速增长，WWW 浏览已经成为人们最主要的日常生活之一。优化 Web 站点结构有利于改善 WWW 浏览行为的质量。本文提出一个兼顾站点拓扑结构和页面内容的页面聚类改进算法，使得我们在挖掘过程中更注重用户访问代价较高的兴趣内容页之间的关系，挖掘出的兴趣页组能较好地支持站点结构优化。仿真实验分析的结果表明，经本文方法优化后的站点能有效降低用户游历其兴趣文档集合的访问代价，有助于改善 WWW 浏览中搜寻与获取有益信息的困难问题及信息搜寻行为的效率低下问题。

参 考 文 献

[1] Jaideep S, Robert C, Mukund D, et al. Web usage mining: discovery and applications of usage patterns from Web data[C]. ACM SIGKDD Explorations Newsletter, 2000, 1(2): 12-23

[2] 陈新中，李岩，杨炳儒，等. Web 日志挖掘技术进展[J]. 系统工程与电子技术，2003，25(4)：492-495

[3] 刘业政，林文龙，焦宁，等. Web 站点结构优化仿真[J]. 系统仿真学报，2007，19(20)

[4] Wu K L, Yu P S, Ballman A. SpeedTracer: a Web usage mining and analysis tool[J]. IBM System Journal(S0018-8670), 1998, 37(1): 89-105

[5] Perkowitz M, Etzioni O. Adaptive Web sites: automatically syn-

thesizing Web pages // Fifteenth National Conference on Artificial Intelligence. Madison, 1998
[6] 杨怡玲，管旭东，尤晋元. 基于页面内容和站点结构的页面聚类挖掘算法[J]. 软件学报，2002，13(3)：467-469
[7] Fayyad U M, Piatetsky S G, Smyth P. The KDD process for extracting useful knowledge from volumes of data[J]. Communications of the ACM, 1996, 39(11): 27-34
[8] Fu Y, Creado M, Shih M Y. Adaptive Web Sites by Web Usage Mining // International Conference on Internet Computing 2001. Las Vegas, USA, 2001
[9] Doyle J K, Graver J E. Mean distance in a graph[J]. Discrete Math., 1977, 17: 147-154
[10] West D B. Introduction to Graph Theory (2nd Edition) [M]. Published by Prentice Hall 1996, 2001: 67-107
[11] Broder A, Kumar R, et al. Graph structure in the Web[C]. Amsterdam, Netherlands, 2000
[12] Kumar R, Raghavan P, Rajagopalan S, et al. The Web as a graph [C] // Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Dallas, Texas, United States, 2000: 1-10
[13] Borges J, Levene M. A Clustering-Based Approach for Modelling User Navigation With Increased Accuracy // Proceedings of the Second International Workshop on Knowledge Discovery from Data Streams (IWKDDS) in conjunction with PKDD 2005. Porto, Portugal, Outubro, 2005

(上接第 187 页)

其中 N 为测试集样本个数，$\tilde{y}_j$ 为预测值。

表 2　分类错误率统计结果(%)

数据集	All Feature	GA-FS	H-MTL	GA-MTL	e-GA-MTL
Colon	29.0±9.2	27.5±9.1	28.4±9.7	26.8±7.0	25.0±7.8
Breast	23.2±2.6	22.7±4.6	23.1±3.5	22.5±5.7	21.5±2.7
Leukemia	30.0±3.2	24.4±3.3	23.3±5.2	23.0±3.9	22.7±4.2
Ovarian	25.8±3.3	23.3±4.3	21.3±4.8	21.2±3.5	21.1±3.1
平均值	27.0±4.6	24.5±5.3	24.0±5.8	23.4±5.0	22.56±4.5

从表 2 中可以看出，5 个算法的分类错误率在平均值上从小到大依次是 e-GA-MTL＜ GA-MTL ＜ GA-FS＜ H-MTL＜ All Feature。除了 Breast 数据集上 GA-FS 结果比 H-MTL 好之外，其它所有数据集上这个结论也都成立。这个结果表明，1)使用了特征选择的 e-GA-MTL，GA-MTL，H-MTL 和 GA-FS 比没有做特征选择的要好；2)使用了多任务学习的 e-GA-MTL，GA-MTL，H-MTL 比没有使用 GA-FS 的要好；3)在多任务学习的基础上，在用来确定哪些特征作为输入/输的方法中，使用遗传算法的 e-GA-MTL 和 GA-MTL 要比预先定义阈值的 H-MTL 方法要好；4)去掉不相关特征的方法 e-GA-MTL 比没有去掉 GA-MTL 的要好。

3.3　讨论

从实验结果中可以看到，本文提出的算法 e-GA-MTL 要好于以前的算法：All Feature，GA-FS，H-MTL 和 GA-MTL。特别是要好于 GA-MTL。

我们知道，不同的特征对于提高学习器的性能会有不同的贡献[3]。有的特征对于学习器的性能的提高效果很明显，其它特征不能替代，可以归类为强相关特征；有的特征效果一般，有替代的特征，但是对学习器性能也有所提高，可以归类为弱相关特征；还有一些特征可以说对于学习器的性能没有提高，反而会损害学习器性能，这种特征就是不相关特征。

e-GA-MTL 能够通过遗传算法自动地把特征分为上述三类，即二进制染色体对应位是 00 的特征是不相关特征，在训练模型之前会被去掉；二进制染色体上对应基因是 10 的特征是强相关特征，在训练模型时作为 MTL 的输入端特征；二进制染色体上对应基因是 01 的特征是弱相关特征，在训练模型时作为 MTL 的输出端；二进制染色体上对应基因是 11 的特征既是强相关特征，又是弱相关特征，也就是临界于强、弱相关特征之间的，这种特征在训练模型时既作为 MTL 的输入，又作为输出。

GA-MTL 表现得没有 e-GA-MTL 好，是由于 GA-MTL 没有考虑不相关特征的影响，而不相关特征会降低多任务学习的泛化能力和鲁棒性。

结束语　本文提出了一个基于遗传算法的多任务学习算法，e-GA-MTL，较之已有的搜索算法，有效提高了多任务学习的预测精度。实验表明，e-GA-MTL 在预测精度上要高于 GA-MTL 以及之前的其他算法，这是由于 e-GA-MTL 不仅能够自动地决定哪些特征作为输入，哪些特征作为输出，而且它能够自动地去掉那些不相关的特征。不相关的特征是 GA-MTL 没有考虑到的。

本文仅探讨了 e-GA-MTL 在分类问题上的有效性，对于回归问题还有待评测。另外，由于特征是被划分为 3 类：输入、输出、删除，而算法中对特征的指派却有 4 个状态：00，01，10，11。特别是对状态为 11 的特征，我们目前的处理是既作为输入，又作为输出，这个问题还可以做进一步的探讨。

参 考 文 献

[1] Liu Huan , Yu Lei . Toward integrating feature selection algorithms for classification and clustering. IEEE Trans on Knowledge and Data Engineering, 2005, 17(3): 1-12
[2] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research, 2003, 3: 1157-1182
[3] Yu Lei, Liu Huan. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 2004, 5 (10): 1205-1224
[4] Mitra P, Murthy C A, Pal S K. Unsupervised feature selection Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301-312
[5] Carnana R, de Sa V R. Benefiting from the variables that variable selection discards. Journal of machine learning research, 2003, 3: 1245-1264
[6] Li G Z, Yang J, Lu J, et al. On multivariate Calibration problems // Lecture Notes on Computer Science 3173. Springer, August 2004: 389-394
[7] Yang J, Honavar V. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems, 1998, 13: 44-49
[8] Li Guo-zheng, Liu Tian-yu. Improving Generalization Ability of Neural Networks Ensemble with Multi-Task Learning. Journal of Computational Information Systems, 2006, 2(4): 1235-1239
[9] Verikas A, Bacauskiene M. Feature selection with neural networks. Pattern Recognition Letters, 2002, 23: 1323-1335
[10] Li Guo-zheng, Yang Jie, Liu Guo-ping, et al. Feature selection for multi-class problems using support vector machines // Lecture Notes on Artificial Intelligence 3173. Auckland, New Zealand, Springer, 2004: 292-300
[11] Foresee F D , Hagan M T . Gauss - newton approximation to Bayesian regularization // Proceedings of the 1997 International Joint Conference on Neural Networks. 1997: 1930-1935
[12] Li J, Liu H. Kent Ridge Biomedical Data Set Repository. Available at: http://sdmc-lit. org. sg/GEDatasets/Datasets. html, 2002