北京大学

硕士研究生学位论文

题目：基于复杂网络的高速公路关键路段挖掘

姓　　名：刘丹萌

学　　号：1401214385

院　　系：北京大学

专　　业：智能科学与技术 (计算机科学与技术)

研究方向：数据挖掘

导　　师：宋国杰

2017 年 5 月 10 日

版权声明

目录

第一章　高速公路关键路段识别模型

对于交通运输、水利传输、能源和通信等基础设施系统，在遭遇自然灾害或者人为灾害时，会对整个系统的性能造成显著的影响，带来重大的经济损失。所以在发生事故或者自然灾害的时候，维护这些网络的完整性至关重要。

灾难管理是一个多阶段的过程，从防灾减灾和准备，着眼于长期消除或降低风险的措施，延伸到灾后响应、恢复与重构。投资基础设施系统在缓解中起着至关重要的作用活动，它可以增强链接的稳定性。但是，将所有的路段稳定性都增强到坚不可摧，在管理人员看来是十分浪费的，甚至会达到负担不起预算的水平。本章节主要研究如何在有限的资源下，找到可以最大化网络通行效率关键路段进行管理。即将资源投放到高速公路路段集合的一个关键子集，尽量减少高速公路的期望通行时间，以达到宏观层面增强路网稳定性的目睹，实现事故前的预防，事故后的快速恢复。

这一章主要研究的是如何对高速公路关键路段挖掘问题进行建模，之后围绕着安徽、山西、北京的收费站车辆数据，求解关路段。

1.1　模型定义

高速公路具有成网性，给定一个有向图 $G = \{V\ E\}$，其中 V 代表收费站（节点）的集合；E 表示边的集合，也就是高速公路中路段的集合。对于通过高速公路出行的车辆，定义 O 为车辆的出发节点，D 作为车辆的目标节点。定义 P_e（$0<P_e<1$）为路段的损毁率，这个概率可以随着交通管理者对路段进行管理、布置资源而减小。定义管理者的决策向量 y=$\{y_1, y_2, \cdots, y_n\}$，y 是一个 n 维向量，每一维 y_i 的数值取 0 或 1，1 表示管理者进行管理，改善路段，0 表示暂时不关

注。因为每一条路段都有一定的概率损毁，所以用 C_{e_i} 来表示第 i 个路段是否损毁，当 C_{e_i} 等于 1 时，路段保持完好，当 C_{e_i} 等于 0 时，路段因为事故损毁。定义 $\boldsymbol{c}=\{C_{e_1},C_{e_2},\cdots,C_{e_n}\}$，c 表示路网的某一种拓扑结构，$\boldsymbol{C}$ 表示路网的所有拓扑结构的集合。对于行驶在高速公路上的车辆，定义车辆的出行时间为 X_i，这个出行时间由车辆的路径选择、路径车流密度决定。当高速公路路段断裂严重，车辆无法抵达目的地时，将车辆的出行时间定为一个常量 M。M 的大小代表了路网连通性的权重。为了更好的求解目标函数，在此提出两个假设：

1）路段之间的损毁概率相互独立：传统研究网络可靠性的相关文献中 [3], 都基于这个假设。

2）M>Max(X_i)：M 必须要大于连通路网中的最大出行代价，即默认断裂对路网造成的影响一定大于路段仍然连通的情况。

根据高速公路的历史事故数据，可以通过结构分析和统计调查 [23]，确定路段损毁的概率，作为本文的先验概率。这个概率可以通过在高速路段上建立基础设施，投放人力资源，或者用其他方式增强路段的稳定性来改变。假设路段以概率 p 损毁，以概率 (1-p) 保持完好。基于路段的损毁率，获取路网拓扑结构概率矩阵 Z：

$$\begin{matrix} C_{e_1}^1 & C_{e_2}^1 & \cdots & C_{e_n}^1 & P^1 \\ C_{e_1}^2 & C_{e_2}^2 & \cdots & C_{e_n}^2 & P^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C_{e_1}^m & C_{e_2}^m & \cdots & C_{e_n}^m & P^m \end{matrix}$$

矩阵中，C_{e_i} 表示第 i 条路段的状态，0 表示遭遇事故，已经损毁，1 表示完好无损；$C^j=\{C_{e_1}^j,C_{e_2}^j,\cdots,C_{e_n}^j\}$ 表示路网的拓扑结构，$P^j=\prod_{i=1}^{n}(P_{e_i}C_{e_i}^j+(1-P_{e_i})(1-C_{e_i}^j))$ 表示高速公路网络拓扑变成这个拓扑结构的概率。

在交通管理者进行一定的决策、处理后，路段的损毁概率发生变

化，相应的，路网拓扑结构概率矩阵 Z 也会发生变化。在此提出关路段挖掘模型：

$$L(\boldsymbol{y}) = -E(T(\boldsymbol{c}|\boldsymbol{y})) \tag{1.1}$$

其中，$T(\boldsymbol{c}|\boldsymbol{y})$：

$$T(\boldsymbol{y}) = P(K|\boldsymbol{c}) \sum_{k \in K} X_k \tag{1.2}$$

$\boldsymbol{y}$ 表示管理者想要投资维护的路段，$T(\boldsymbol{c}|\boldsymbol{y})$ 表示当路网拓扑结构为 c 的时候，高速公路的整体通行时间，对时间取负，转化为通行效率。模型的目标是研究如何选取路段，对路段增加维护，使得整个路网的通行效率得到提升。结合式1.1，式1.2，得到展开式：

$$Max(L(\boldsymbol{y})) = -Min_{\boldsymbol{y}} \sum_{\boldsymbol{c} \in C} P(\boldsymbol{c}|\boldsymbol{y}) P(K|\boldsymbol{c}) \sum_{k \in K} X_k \tag{1.3}$$

式中 $\boldsymbol{y}$ 表示关键路段集合，假设高速公路网络的路段数量为 n，则 $\boldsymbol{y}$ 为 n 维向量，对于 $\boldsymbol{y}$ 的第 i 个维度，0 表示第 i 个路段不是关键路段，1 表示第 i 个路段是关键路段；$\boldsymbol{c}$ 表示路网的拓扑结构，C 是高速公路网络所有拓扑结构的集合；$P(\boldsymbol{c}|\boldsymbol{y})$ 表示当关键路段集合为 $\boldsymbol{y}$ 时，高速路网的拓扑结构为 $\boldsymbol{c}$ 的概率；k 表示第 k 个车辆的出行路径，K 表示所有车辆的出行路径集合；$P(K|\boldsymbol{c})$ 表示当路网拓扑结构为 $\boldsymbol{c}$ 时，高速公路车辆出行路径集合为 K 的概率；X_k 表示当车辆的行驶路径为 k 时，车辆的行驶时间。

1.2 子模性分析

1.2.1 子模性定义

次模函数（submodular function）是一种具有“边际效应递减”效应的函数，即对于一个集合函数，如果 $S \subseteq V$，那么在 V 中增加一个元素所增加的收益要小于等于在 S 的子集中增加一个元素所增加的收益。形式化表述就是：对于函数 f 而言，若 $A \subseteq B \subseteq V$，且 $\varepsilon \in V - B$，则 $f(A \cup \{\varepsilon\}) - f(A) \geq f(B \cup \{\varepsilon\}) - f(B)$；或者若 $A \subseteq \Omega\ B \subseteq \Omega$，则 $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$；或者对于任意 $X \subseteq \Omega\ x_1, x_2 \in \Omega$，下面的式子一定成立：$f(X \cup x_1) + f(X \cup x_2) \geq f(X \cup x_1, x_2) + f(X)$。满足这三个条件中的任意一个，函数 f 即满足子模性。

1.2.2 子模性证明

假设 ε 是某一条路段，$\boldsymbol{y} \subseteq \boldsymbol{Y} \subseteq \boldsymbol{\Omega}$，$\boldsymbol{\Omega}$ 是关键路段集合的全集空间。$\varepsilon \in \Omega - Y$。$\{y + \varepsilon\}$ 表示对于关键路段集合 y，将 ε 作为新的关键路段加入，形成新的关键路段集合。

定义:

$$I = L(\boldsymbol{y} + \boldsymbol{\varepsilon}) - L(\boldsymbol{y}) - (L(\boldsymbol{Y} + \boldsymbol{\varepsilon}) - L(\boldsymbol{Y})) \tag{1.4}$$

不妨假设 $Y = y + \varepsilon_2$，公式 1.4 转化为：$I = L(\boldsymbol{y} + \boldsymbol{\varepsilon_1}) - L(\boldsymbol{y}) - (L(\boldsymbol{y} + \boldsymbol{\varepsilon_1} + \boldsymbol{\varepsilon_2}) - L(\boldsymbol{y} + \boldsymbol{\varepsilon_2}))$

令 $J = L(\boldsymbol{y} + \boldsymbol{\varepsilon_1}) - L(\boldsymbol{y})$，要证明 $I \geq 0$，即证 J 单调非增。

J 属于有限离散函数，对 J 进行求导化简 [3]，得到：$\frac{dy}{dx} = \sum(\sum_{c_1|y+\varepsilon} P(c_1) - \sum_{c_2|y} P(c_2))X_k$。显然 $\sum_{c_1|y+\varepsilon} P(c_1) * X_k$ 具有单调非减性，导数恒大于 0。模型的子模性得到证明

对于具有子模性的模型，贪心求解的精度误差不会超过 $\frac{1}{e} * \boldsymbol{OPT}$

1.3　贪心求解

贪心方法在时间复杂度上比暴力枚举要少一个数量级。贪心算法步骤如下：

Algorithm 1 贪心算法求解模型

Require: 高速车辆 O-D 数据，高速公路网络拓扑结构，关键路段数量，路段损毁率
Ensure: 高速公路关键路段集合

```
function GREEDY(ODMatrix G = V, E B P_e)
    res ← 0
    Array ← []
    k ← 0
    l ← 0
    while len(Array) ≤ B do
        for i ∈ E − Array do
            if L(Array + i) > k then
                k = L(Array + i)
                l = i
            end if
        end for
        res ← k
        Array ← Array + l
    end while
    return Array
end function
```

为验证贪心算法的效果，在此引入对比方法：

算法2使用枚举方法, 获取最优解

算法3利用高速公路网络拓扑结构，抽取关键路段。算法中的 Z(i) 是计算路段 i 的中心性函数

算法4基于统计学方法，计算路段重要程度，获取关键路段。式中 f_i 表示路段 e 的流量：

1.4　实验及结果

本节针对各种方法在真实的交通数据集中进行实验，通过对比已有的关键路段挖掘方法，评估模型的效果。实验环境为：Windows Server 2008，64GB RAM，Inter(R)Xeon(R) CPU E7-4830 2.13GHz

Algorithm 2 枚举

Require: 高速车辆 O-D 数据，高速公路网络拓扑结构，关键路段数量
Ensure: 高速公路关键路段集合

function ENUMERATION($ODMatrix\ G = V, E\ B\ P_e$)
    $res \leftarrow 0$
    $Array \leftarrow []$
    $k \leftarrow 0$
    for $l \in \Omega$ **and** $len(l) \leq B$ **do**
        if $L(l) > k$ **then**
            $k = L(l)$
            $Array = l$
        end if
    end for
    return $Array$
end function

Algorithm 3 拓扑中心性

Require: 高速公路网络拓扑结构，关键路段数量
Ensure: 高速公路关键路段集合

function ENUMERATION($ODMatrix\ G = V, E\ B$)
    $res \leftarrow 0$
    $Array \leftarrow []$
    $k \leftarrow \{\}$
    for $i \in E$ **do**
        $k \leftarrow \{i, Z(i)\}$
    end for
    $SortbyValue(k)$
    $Array \leftarrow k[0 : B]$
    return $Array$
end function

Algorithm 4 统计

Require: 高速公路网络拓扑结构，关键路段数量，高速公路路段损毁概率
Ensure: 高速公路关键路段集合

function ENUMERATION($G = V, E\ B\ P_e$)
    $res \leftarrow 0$
    $Array \leftarrow []$
    $k \leftarrow \{\}$
    for $i \in E$ **do**
        $k \leftarrow \{i, f_i * P_i\}$
    end for
    $SortbyValue(k)$
    $Array \leftarrow k[0 : B]$
    return $Array$
end function

2.13GHz (2 处理器)，后续章节的实验均在相同的实验环境下进行。特别地，实验中采用了两个国内高速公路网的数据：安徽省和山西省高速公路网数据。

1.4.1　实验数据

本节的实验数据来自于安徽省和山西省的高速公路路网，其中的数据为高速路网中车辆的行驶 O-D 数据。该路网中包含 142 个出口位置和 142 个入口位置。为了方便研究，将车辆的 O-D 数据整合为出行 O-D 矩阵 ODMatrix：

$$\begin{matrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{matrix}$$

其中，a_ij 表示以收费站 i 为起点 O，以收费站 j 为终点 D 的车辆数量。

1.4.2　实验结果

图1.1，1.2给出了在不同时间段下，几种方法的最终结果比较。图1.1是基于 2010 年 10 月 30 日一天的实验结果，纵坐标代表路网整体通行效率（路网整体通行时间取负）的绝对值，横坐标代表一天内的不同时间段，本实验中以 1 小时为一个时间段，采样八个时间点 [0,3,6,9,12,15,18,21]。由图1.1可以发现，在整体上贪心算法明显优于统计算法，同时统计算法又比直接基于高速公路拓扑结构强，应该是高速公路整体网络结构比较简单，路网拓扑结构对整体路网的影响不明显。在不同的时间段，高速公路的流量在不断变化，不同方法的效果之间的差异性也在变化，在高速公路车流最少的午夜，几种方法差

图 1.1　fig1

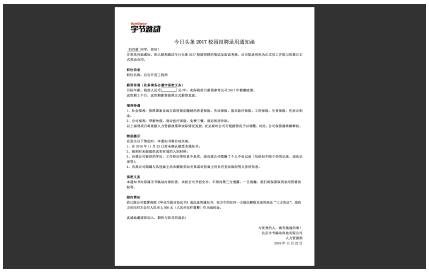
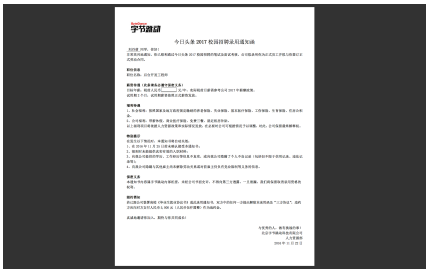


图 1.2　fig2

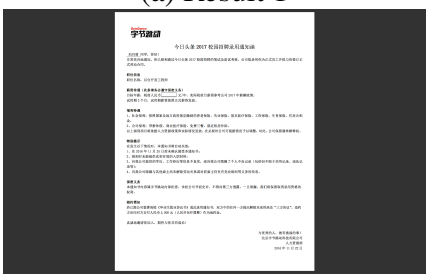
异达到最小，从六点开始，到流量最高的中午，三种方法之间的差异逐渐增大，这体现了高速公路流量对关键路段选取后的效果具有影响，流量越大，关键路段维护后造成的效益越大。图1.2是基于从 2010 年 10 月 10 日开始，到 2010 年 10 月 16 日为止的一周数据的实验，纵坐标和图1.1一样，表示网络整体的通行时间。纵坐标以一天为一个时间段，采样七天（从周日到下一个周六）。可以发现，在以一整天的 O-D 矩阵为数据集进行研究时，不同天之间的路网通行效率变化较小，不同方法之间的差异也趋于平稳。这证明了高速公路具有稳定性，以及研究有规律的静态关键路段的可行性（即他不管什么时候都是关键路段，不改变）。

图1.3给出了关键路段在路网中的分布图，图1.3(a) 是用贪心算法求得的关键路段集合，图1.3(b) 是高速公路统计方法获得的路段集合。图1.3(c) 是基于枚举所得的最优解集，图1.3(d) 是基于路网拓扑结构选取的关键路段集合。图1.3(a) 中颜色的变化和粗细的变化表示路段在贪心求解过程中，路段被选择的顺序；图1.3(b) 中颜色的变化表示路段的重要程度。对比两图可以发现，直观上重要的点（承载流量较大的路段，事故多发路段等）并不一定在路网中属于关键节点，需要经过一些计算才能求出；直接枚举的路段集合与贪心算法求得的路段集合十分接近。

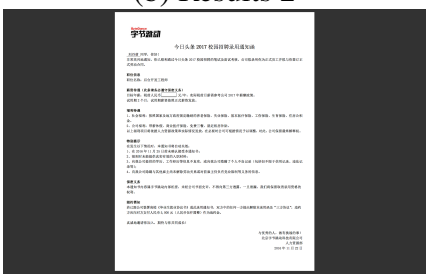
(a) Result 1



(b) Results 2



(c) Result 3



(d) Result 4

图 1.3　Example of placing a figure with experimental results.

1.4.3　时间复杂度分析

基于暴力枚举方法的时间复杂度：$O(n^B * 2^n)$

基于贪心算法的时间复杂度：$O(n * B * 2^n)$

基于统计路段重要性方法的时间复杂度：$O(n * \log(n))$

基于路网拓扑结构方法的时间复杂度：$O(n * \log(n))$

其中，后两个方法可以用大根堆将时间复杂度优化到 O(n)。

第二章　　章节

结论

pkuthss 文档模版最常见问题：

在最终打印和提交论文之前，请将 pkuthss 文档类选项中的 `colorlinks` 替换为 `nocolorlinks`，因为图书馆要求电子版论文的目录必须为黑色，且某些教务要求打印版论文的文字部分为纯黑色而非灰度打印。

`\cite`、`\parencite` 和 `\supercite` 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记：**test-en**，**[test-zh]**、**[test-en, test-zh]**。

若要避免章末空白页，请在调用 pkuthss 文档类时加入 `openany` 选项。

如果编译时不出参考文献，请参考 `texdoc pkuthss`“问题及其解决”一章“其它可能存在的问题”一节中关于 biber 的说明。

附录 A　附件

pkuthss 文档模版最常见问题：

在最终打印和提交论文之前，请将 pkuthss 文档类选项中的 `colorlinks` 替换为 `nocolorlinks`，因为图书馆要求电子版论文的目录必须为黑色，且某些教务要求打印版论文的文字部分为纯黑色而非灰度打印。

`\cite`、`\parencite` 和 `\supercite` 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记：**test-en**，**[test-zh]**、**[test-en, test-zh]**。

若要避免章末空白页，请在调用 pkuthss 文档类时加入 `openany` 选项。

如果编译时不出参考文献，请参考 `texdoc pkuthss`“问题及其解决”一章“其它可能存在的问题”一节中关于 biber 的说明。

致谢

pkuthss 文档模版最常见问题：

在最终打印和提交论文之前，请将 pkuthss 文档类选项中的 `colorlinks` 替换为 `nocolorlinks`，因为图书馆要求电子版论文的目录必须为黑色，且某些教务要求打印版论文的文字部分为纯黑色而非灰度打印。

`\cite`、`\parencite` 和 `\supercite` 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记：**test-en**，**[test-zh]**、**[test-en, test-zh]**。

若要避免章末空白页，请在调用 pkuthss 文档类时加入 `openany` 选项。

如果编译时不出参考文献，请参考 `texdoc pkuthss`“问题及其解决”一章“其它可能存在的问题”一节中关于 biber 的说明。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校在 □ 一年 / □ 两年 / □ 三年以后在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名:　　导师签名:　　日期:　　年　　月　　日