

# On the Models of Graph Theory Based on Minimum Spanning Tree with Applications in Medicine

HUANG Yi-zhen<sup>1</sup>, ZHANG Shi-jie<sup>2</sup>, CHEN Wei<sup>3</sup>, JIN Qing-yue<sup>4</sup>

(1. Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200030, China;

2. Flash Product Group, Intel Ltd., Shanghai 200030, China;

3. Department of Mathematics, Zhejiang University, Hangzhou 310058, China;

4. Jinhua College of Profession and Technology, Jinhua 321007, China)

**Abstract:** Sample data classification is a commonly used tool in medical research. A new thought and methodology for classification is proposed in the paper. The concept of maximal  $\lambda$ -cut subgraph is introduced based on the insight into classification process and its major contradictions. As an illustration, three models based on minimum spanning tree are given as a paradigm of new classifiers upon the platform of graph theory. Then its application in research of the relationship between nutrition and disease, and gene classification is analyzed. Finally the application prospect of graph theory in medical is discussed.

**Key words** graph theory ; minimum spanning tree ; classification ; maximal  $\lambda$ -cut subgraph

## 基于最小生成树的图论模型及其在医学中的应用

黄宜真<sup>1</sup>, 张世劫<sup>2</sup>, 陈 巍<sup>3</sup>, 金庆跃<sup>4</sup>

(1. 上海交通大学计算机科学与工程系, 上海市 200030; 2. 英特尔(上海)有限公司闪存研究小组, 上海市 200030;

3. 浙江大学数学系, 浙江 杭州市 310058; 4. 金华职业技术学院, 浙江 金华 321007)

**摘要:** 样本数据分类是医学研究中常见的工具。本文提出了一种新的数据分类思想和方法。在分析分类过程及其主要矛盾的基础上, 提出了极大 $\lambda$ -截子图的概念。作为示范, 建立了三个基于最小生成树的图论模型, 并分析了其在研究营养与疾病的关系以及基因分类中的应用。最后讨论了图论在医学中的应用前景。

**关键词:** 图论 ; 最小生成树 ; 分类 ; 极大 $\lambda$ -截子图

中图分类号: R311 文献标识码: A 文章编号: 1671-3699(2004)03-0014-05

## 1 Introduction

Graph theory is a branch of mathematics that investigates the properties of graphs. A graph  $G=(V,E)$  is an abstract structure consisting of points (or vertices)  $V$  and connections (or edges)  $E$ . It implies an abstraction of the reality so it can be simplified as a set of linked nodes. Such a structure is found in many industrial applications, such as networks, production schedules and diagrams.

The most common network topology design problem is the minimum spanning tree (MST) prob-

lem in graph theory. Graham and Hell provided a comprehensive survey of various aspects of MST<sup>[1]</sup>. Efficient polynomial algorithm to the MST problem are Prim's Algorithm<sup>[2]</sup> with time complexity  $O(n^2)$  for dense graphs and Kruskal's algorithm<sup>[3]</sup> with time complexity  $O(e \log e)$  for sparse graphs, where  $n$  and  $e$  are the number of vertices and edges of the graph, respectively. Traditionally the basic MST problem and its variations, such as capacitated MST<sup>[4]</sup>, probabilistic MST<sup>[5]</sup>, and degree constrained MST problem<sup>[6]</sup>, are applied to the design of telecommunications network infrastructure.

收稿日期: 2004-08-16

基金项目: 金华市科技局基金资助项目(项目编号: 2004-1-363-199)

作者简介: 黄宜真(1957-), 男, 浙江金华人, 本科生. Electronic Publishing House. All rights reserved. <http://www.cnki.net>

This papers attempts to apply MST to classification of sample data in medicine. Sample data classification is usually a significant initialization step after which further medical research methodology may be applied. The rationality and feasibility of classification has a close relationship with the result of research.

The remainder of the paper is organized as follows. In Section 2, we present the definition of MST, the MST property and the fundamental algorithm on which the models discussed next are based. Section 3 gives three graph theory models. Section 4 discusses their applications in nutrition and the human genome, and Section 5 states some conclusions.

## 2 The basic MST problem

**Definition.** Let  $G=(V,E)$  be an undirected connected graph in which each edge  $(u,v) \in E$  has an associated cost  $C(u,v)$ . A connected subgraph  $G'=(V',E')$  satisfying  $V=V'$  is called a spanning subgraph. If a spanning subgraph is a tree, it is called a spanning tree. The cost of a spanning tree is the sum of costs on its edges. An MST of  $G$  is a spanning tree of  $G$  with a minimum cost.

**The MST Property.** Suppose  $G=(V,E)$  is an undirected connected graph with costs defined on all edges. Let  $U$  be some nonempty subset of  $V$ . If  $(u,v)$  is an edge of lowest cost such that  $u \in U$  and  $v \in V-U$ , then there exists an MST that includes  $(u,v)$  as an edge.

The MST property only holds for undirected graphs and it is the basis for MST algorithms. As to directed graphs, an algorithm for solving this problem has been provided independently by Chu and Liu<sup>[7]</sup> and Edmonds<sup>[8]</sup>.

**Definition.** A set of edges  $T$  is promising if it can be extended to produce an MST.

By definition,  $T=\emptyset$  is always promising. Also, if a promising set of edges  $T$  is a spanning tree, then  $T$  must be an MST.

**Definition.** An edge is said to leave a given set of vertices if exactly one end of this edge is in the set.

weighted undirected connected graph,  $U$  is some nonempty subset of  $V$ ,  $T$  is a promising set of edges with no edge leaves  $U$ ,  $e$  is an edge with least cost that leaves  $U$ . Then the set of edges  $T' = T \cup \{e\}$  is still promising.

Because the models in this paper all deals with complete graphs, only Prim's Algorithm is presented. It is directly based on the MST lemma. The input is a weighted undirected connected graph  $G=(V,E)$ . Assume that  $V=\{1,2,\dots,n\}$ .

### Prim's Algorithm

1.  $T \leftarrow \emptyset, U \leftarrow \{1\}$ .
2. While  $U \neq V$ :
  - 2.1 Let  $(u,v)$  be the lowest cost edge such that  $u \in U$  and  $v \in V-U$ .
  - 2.2  $T \leftarrow T \cup \{(u, v)\}$ .
  - 2.3  $U \leftarrow U \cup \{v\}$ .

The proof of correctness of the MST property, the MST lemma and Prim's Algorithm can be found in literature<sup>[2,9]</sup>.

## 3 Graph theory models

**General abstraction and definition.** Suppose the sample data has  $m$  items of attributes. Then each sample individual may be regarded as an  $m$ -dimensional vector. The vertices of the graph are defined as all the vectors  $S_1, S_2, \dots, S_n$  in which  $n$  is the sample size. The weight value of an edge is the difference between two individuals. There are a variety of ways to compute the difference. The following are a few examples:

- ① The Euclidean distance formula:

$$d_{AB} = \sqrt{\sum_{i=1}^m |A_i - B_i|^2}$$

- ② The Minkowski distance formula:

$$d_{AB} = \left[ \sum_{i=1}^m |A_i - B_i|^p \right]^{1/p}$$

It is the generalized form of ①.

- ③ The exponential formula:

$$d_{AB} = \exp\left( \sum_{i=1}^m |A_i - B_i|^p \right)$$

Thus, a weighted undirected complete graph  $G=(V,E)$  is derived. It is the input of our models.

classification: ① The individuals within the same class should be as similar as possible. ② The class should include as many individuals as possible in order to leave following research with a larger sample size. ① and ② contradict with each other. To some extent the process of classification is to strike a balance.

To precisely delineate this for our models, some definitions are given. As to ① The difference of a subgraph is the cost of an MST of the subgraph. A  $\lambda$ -cut subgraph is a subgraph with difference less than a given constant  $\lambda$ . As to ② A  $\lambda$ -cut subgraph is maximal if any subgraph containing it is not a  $\lambda$ -cut subgraph. Thus, every maximal  $\lambda$ -cut subgraph is a class consisting of similar individuals with the consideration of ① and ②. We can adjust  $\lambda$  to control the fineness of classification.

But the above philosophy and principle may lead to excessive overlapping among classes. Another approach to classification is partition that is to divide the overall without any overlapping and missing.

The next are three different models all to find maximal  $\lambda$ -cut subgraphs or partitions based on which further medical research methodology may be applied.

**Model A.** The idea of this model is using brute-force search to find all maximal  $\lambda$ -cut subgraphs. Its algorithm is based on the MST property. An FIFO (First-In First-Out) structure—queue<sup>[9]</sup> is adopted as the frame of this model. Each node  $X=(\text{Diff}, U)$  in the queue (denoted as  $Q$ ) represents a subgraph. Diff is the difference of the subgraph and  $U$  is a set recording the vertices belong to it. In order to prevent duplication of computation efficiently during implementation, hash table<sup>[9]</sup> is applied to record the position whenever a new node is reproduced and put into the queue. Usually a hash table based on a modular arithmetic function and link lists would achieve an ideal effect.

#### Algorithm of Model A

1.  $Q \leftarrow \{(0, \{1\}), (0, \{2\}), \dots, (0, \{n\})\}$ .
2. While  $Q \neq \emptyset$  :
  - 2.1 Get a node  $X=(\text{Diff}, U)$  out of  $Q$ .
  - 2.2 For every vertex  $v \in V-U$ , let  $(u, v)$  be

2.3 If  $\text{Diff}+C(u, v) < \lambda$ , put a new node  $X'=(\text{Diff}+C(u, v), U \cup \{v\})$  into  $Q$ . Otherwise,  $X$  is a maximal  $\lambda$ -cut subgraphs and is to be output.

**Model B.** This model is to find a partition of the graph. It is directly based on Prim's Algorithm.

#### Algorithm of Model B

1.  $T \leftarrow$  an MST of  $G, U \leftarrow \{1, 2, \dots, n\}$ .
2. While  $U \neq \emptyset$ 
  - 2.1 Let  $u \in U, \text{Diff} \leftarrow 0, X \leftarrow \{u\}$ .
  - 2.2 Let  $(x, v) \in T$  be the lowest-cost edge such that  $x \in X$  and  $v \in U-X$ .
  - 2.3 If  $\text{Diff}+C(x, v) < \lambda, \text{Diff} \leftarrow \text{Diff}+C(x, v), X = X \cup \{v\}$ , return to step 2.2.
  - 2.4  $U = U - X$ .
  - 2.5  $X$  is a class to be output.

**Model C.** The mechanism of this model is to repeat for times each time finding a maximal  $\lambda$ -cut subgraph with the most vertices and then adjust the weight value of  $G$  in a certain way. A heuristic randomized approximate algorithm is applied as the kernel of the model. In detail, a heuristic function is computed according to the value of a monotonous transformation function taking the cost at the current step as input. Many classical functions can be our choices such as sigmoid functions, exponential functions and trigonometric functions within a given interval.

The means to modify  $G$  are flexible and this paper only offers a viable method. The modification of  $G$  also brings about a problem, that is, a maximal  $\lambda$ -cut subgraph obtained from current  $G$  may not be a maximal  $\lambda$ -cut subgraph for the original  $G$ . However, this does not affect the correctness of classification. Another point should be indicated is that, if the change to the weight value is monotonously increasing like in this paper, the resulted maximal  $\lambda$ -cut subgraph tends to diminish slightly each time and vice versa.

In practice, a great deal of optimal solutions exists. Consequently, an approximate algorithm produces a best solution in all likelihood. Even if the result is one or two vertices less than the optimal, it is still very acceptable.

A few parameters for this model must be pre-determined:

required

$T$ ——Number of times for computing an approximate solution

$O$ ——Overlapping coefficient indicating how much the resulting subgraphs overlap

$F(x)$ ——Transformation function

#### Algorithm of Model C

1. Repeat the following steps for  $N$  times:

1.1 Repeat the following steps for  $T$  times:

1.1.1 Choose a vertex  $u$  randomly as the source,  $T \leftarrow \emptyset$ ,  $U \leftarrow \{u\}$ ,  $\text{Diff} \leftarrow 0$ .

1.1.2 For every expandable vertex  $v$ , let  $(u, v)$  be the lowest cost edge such that  $u \in U$ .

1.1.3 SF is the sum of  $F(C(u, v))$  for every  $v$ .

1.1.4  $F(C(u, v))/\text{SF}$  is the probability for choosing  $v$  as the next vertex for extension.

1.1.5 Let  $v$  be the chosen vertex,  $(u, v)$  be the lowest-cost edge such that  $u \in U$ .

1.1.6  $T \leftarrow T \cup \{(u, v)\}$ ,  $U \leftarrow U \cup \{v\}$ ,  $\text{Diff} \leftarrow \text{Diff} + C(u, v)$ .

1.1.7 If  $\exists$  an expandable vertex  $v$ , return to step 1.1.2.

1.1.8 If  $U$  has more vertices than the current optimal solution  $X$ , update  $X$  with  $U$ .

1.2  $X$  is a maximal  $\lambda$ -cut subgraphs to be output.

1.3 For every edge  $(u, v)$  such that  $u \in X$  or  $v \in X$ ,  $C(u, v) \leftarrow C(u, v) + O$ .

(Note: A vertex  $v$  is expandable if  $v \in V - U$ ,  $\exists (u, v)$  such that  $u \in U, \text{Diff} + C(u, v) < \lambda$ .)

#### Discussion

Model B and C are both with polynomial algorithms. But model A is non-polynomial. When the number of resulting subgraphs is too large, it will be quit time-consuming and actually we are not able to analyze all of them in later research. So the remedy to model A is to cut off some nodes violently according to a specific rule in step 2.3 The significance of model A is to show a feasible means to find out all maximal  $\lambda$ -cut subgraphs.

Whether one model is superior to the other depends on practical circumstances. No assertion should be made before meticulous observation of data and several trials.

## 4 Applications of models

Sample data classification is a commonly used tool in medical research. In this Section, we proposed two application areas with the first explained more detailedly to reveal some essence of the difficulties confronted by researchers.

### 4.1 Research into the relationship between nutrition and disease

Literature [10] reviews research methodology in the relation between diet and disease. Medical research in this area can be divided into two categories: complex research that seeks to determine the mechanism and etiology of disease, and simple research that investigates directly the factors that cause or prevent disease. Great progress has been made since the 1970s in elucidating the relation between diet and disease such as the relation between selenium and cancer, carotenoids and cancer, vitamin E and coronary heart disease, sodium and hypertension, alcohol and stroke. In those instances, we see a dearth of evidence that complex research has been of significant practical value and efficiency whereas simple research has received well under half of the resources but has generated the large majority of effective valuable information.

Though intervention trials are still the undisputed gold standard among different types of simple research, it is not humanitarian or realistic under many situations. The one that has proven of most value is cohort studies and it is gaining its importance in the recently decades. In cohort studies, after random sample, we often want to figure out the relation between a specific factor and disease. Before doing this, the interference from other factors should be eliminated as much as possible. Unlike case-control or intervention trials, we have no control to the interference factors, hence it leads to a compromise that is to classify the sample to gather individuals with similar interference together and assume they are the same. In these cases our models will work. This is an active method to interference elimination compared to random sample that we consider passive. Classification is not a substitution but a reinforcement of random sample: the former excludes significantly correlated interferences while the latter copes with unknown ones or those factors that are mild and diffi-

cult to measure.

To illustrate, we take the study of children obesity as an instance. Sothorn put forward constructive suggestions and systematic measures to prevent this epidemic among children<sup>[11]</sup>. A series of factors such as dietary intake, physical activity, obesity of parents, family environment, nutrient status during pregnancy and breast-feeding history, influence the weight of children. Suppose we probe to determine the impact of dietary intake. To employ our models, all factors must be quantified into one or several attributes. All attributes constitute a vector. Then any model from Section III may be adopted to accomplish classification. Because only interferences are to be eliminated, the attributes related to dietary intake should be exempt from this process.

#### 4.2 Classification in the human genome

DNA microarray technology has enabled the quantitative measurement of thousands of gene expression levels simultaneously. Through this technology, it is possible for molecular biologists to study the differential gene expression across a set of related assays. Being a high throughput technology, it allows whole genomes to be scanned, generating thousands of data points per microarray experiment. With the accomplishment of the Human Genome Project and complete genome sequence data becoming available at an increasing rate, the problem of gene classification with different criteria is becoming pressing.

The living cell is a complex system comprising multiple cellular pathways that performs different biological functions. Through genome-wide measurements of the mRNA expression levels across different experimental conditions, we can construct a global map of the functional associations of various genes based on their differential expression patterns. This process is called gene clustering that involves classifying genes into groups each encoding proteins required for a common function. Some related work: Mayor et al. presented a systematic study of the clustering of genes within the human genome based on homology inferred from both sequence and structural similarity<sup>[12]</sup>; Latent semantic analysis (LSA) was applied to microarray expression data for genome-wide functional classification of genes by Ng et al.<sup>[13]</sup>. If we abstract the map as a graph with edges defined using similarity met-

rics upon the gene expression data, our models are also applicable here.

## 5 Concluding remarks

In this paper, we describe a new method for classification with applications in nutrition and the human genome. Unlike traditional methods that usually divide the overall without overlapping and missing or what we call it partition in the sense of algebra, maximal  $\lambda$ -cut subgraphs are produced by model A and C. The latent hypothesis of partition is that the validity of information from every individual is equal. In fact, however, noises are inevitable and some individuals may be more representative than others. The rationality of non-partition classification methods is that it can filter some abnormal ones and give appropriately more weight for more representative ones.

A very mature and popular classification methodology is fuzzy cluster analysis with applications covering pattern recognition, image processing, rule generation and other scientific and engineering fields. As an alternative and complement, MST based-classifiers cluster things along a mainstream—the MST. With this entirely different clustering pattern, it is probable that they are superior to fuzzy clustering in some situations.

As we can see from Section 3, graph theory provides us with a flexible platform to create novel classifiers. This paper only provides a paradigm of MST based-models. Besides MST, various structures and algorithms from graph theory may be applied as a replacement or improvement of current classification methodologies.

In recently years, a few stirring attempts and progress has been made for the application of graph theory in medical which foretold its promising prospect: Michel Joubert et al. modelled both the information in patient databases and the queries to these databases with conceptual graphs and presented a prototype to exploit this model<sup>[14]</sup>; Allore and Schruben used event graphs to organize and extend scientific knowledge about diseases<sup>[15]</sup>. Nevertheless, further effort should be devoted to explore its potential and make graph theory a useful tool to assist and accelerate the process of modern medical analysis. (下转第9页)

对于 (9) 式，两边取 mod4 知它不成立.

对于 (10) 式，由于它可分解为

$$(y_2^2+1)(y_2^2-1)=py_1^4$$

这里  $y_3>0, y_4>0, y_1=y_3y_4$ ，故有以下两式

$$\begin{cases} y_2^2+1=2py_3^4 \\ y_2^2-1=2y_4^4 \end{cases} \quad (11)$$

或

$$\begin{cases} y_2^2-1=2py_3^4 \\ y_2^2+1=2y_4^4 \end{cases} \quad (12)$$

对于 (11)，由于  $y_2^2-1=2y_4^4$  无正整数解 [2]，故它不成立.

对于 (12) 式，由于  $y_2^2+1=2y_4^4$  仅正整数解 [2]

$(y_2, y_4) = (1, 1), (239, 13)$ ，将它代入 (12) 式的第一式知它也不成立.

③  $a=0, b=0$  时，(3) 式给出

$$py_1^4 - y_2^4 = 2 \quad (13)$$

或

$$py_1^4 - y_2^4 = -2 \quad (14)$$

对于 (13) 式，两边取 mod16，由于  $p \not\equiv 3 \pmod{16}$ ，故不成立.

对于 (14) 式，两边取 mod16，由于  $p \not\equiv 15 \pmod{16}$ ，故它也不成立.

总之，可见定理成立.

参考文献:

[1] 柯 召, 孙 琦. 数论讲义[M]. 北京:高等教育出版社,1987 218~244.  
 [2] 曹珍富. 丢番图方程引论[M]. 哈尔滨:哈尔滨工业大学出版社,1989 273, 26, 289.

(上接第 18 页)

References:

[1] R L Graham and P Hell. On the history of the minimum spanning tree problem[J]. Annals of the History of Computing, 1985,7(1):43~57.  
 [2] R C Prim. Shortest connection networks and some generalizations[J]. Bell Systems Technology Journal, 1957,36:1389~1401.  
 [3] J B Kruskal. On the shortest spanning tree of a graph and the traveling salesman problem [J]. Proceedings of the American Mathematical Society, 1956,7:48~50.  
 [4] R K Ahuja, J B Orlin and D Sharma. A composite very large-scale neighborhood structure for the capacitated minimum spanning tree problem[J]. Operations Research Letters, 2003,31(3):185~194.  
 [5] F J Rohlf. A probabilistic minimum spanning tree algorithm[J]. Information Processing Letters, 1978,7(1):44~48.  
 [6] L Caccetta and S P Hill. A branch and cut method for the degree-constrained minimum spanning tree problem[J]. Networks, 2001,37(2):74~83.  
 [7] Y J Chu and T H Liu. On the shortest arborescence of a directed graph[J]. Scientia Sinica, 1965,14:1396~1400.  
 [8] J Edmonds. Optimum branchings[J]. Journal of Research National Bureau of Standards.1967,71B:233~240.  
 [9] C A Shaffer. Practical introduction to data structures and algorithm analysis, 2nd edition[M]. Upper Saddle River: Prentice Hall, 2000.  
 [10] N J Temple. Nutrition and disease: challenges of research design[J]. Nutrition, 2002,18(4):343~347.  
 [11] M S Sothorn. Obesity prevention in children: physical activity and nutrition[J]. Nutrition, 2004,20:704~708.  
 [12] L R Mayor, K P Fleming, Arne Muller, D J Balding and M J Sternberg. Clustering of protein domains in the human genome[J]. Journal of Molecular Biology,2004,340(5):991~1004.  
 [13] S K Ng, Z Zhu and Y S Ong. Whole-genome functional classification of genes by latent semantic analysis on microarray data[A]. Proc. 2nd Asia-Pacific Bioinformatics Conference (APBC2004)[C], Dunedin, New Zealand.  
 [14] Michel Joubert, Marius Fieschi, J J Robert and Ali Tafazzoli. Users conceptual views on medical information databases [J]. International Journal of Bio-Medical Computing, 1994,37(2):93~104.  
 [15] H G Allore and L W Schruben. Disease management research using event graphs[J]. Computers and Biomedical Research, 2000,33(4):245~259.