采用组合方法进行链路预测的理论极限研究

吴翼腾[1]，于洪涛[1]，黄瑞阳[1]，李华巍[2]
（1. 信息工程大学，河南 郑州 450002；2. 92538 部队，辽宁 旅顺 116041）

摘 要：对链路预测组合方法是否存在理论极限以及如何抵近极限开展研究。从是否使用多维度信息或是否直接定义多维度信息之间关系的角度，将链路预测方法分为单机制方法和组合方法。采用简单函数列逼近可测函数的方法，得出链路预测组合方法的理论极限定理；提出使组合方法准确性达到理论上限的组合规则，并给出所提组合规则的几何解释和针对极限定理的仿真示例说明。极限定理揭示了组合方法的本质和组合方法相比单机制方法具有更高准确性及稳健性的原因。

关键词：复杂网络；链路预测；组合方法；理论极限
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2020125

Theoretical limit of link prediction using a combination method

WU Yiteng[1], YU Hongtao[1], HUANG Ruiyang[1], LI Huawei[2]
1. Information Engineering University, Zhengzhou 450002, China
2. Unit 92538 of the PLA, Lyushun 116041, China

Abstract: The problem that whether there a theoretical limit exists for link prediction combination methods and how to approximate was investigated. Link prediction methods were divided into single or combination methods, based on whether multidimension information was used, or whether the relation of multidimension information was defined directly. Limit theorems for link prediction by approximating a measurable function by a simple function sequence were provided. Combination rule and corresponding geometric interpretations and simulation examples for limit theorems were also provided. Limit theorems show why combination methods have higher accuracy and robustness than single methods.

Key words: complex network, link prediction, combination method, theoretical limit

1 引言

理论极限问题是各科学领域普遍关注的基本理论问题。*Science* 在建刊 100 年时曾提出的 125 个科学问题中，就包括什么是传统计算的极限[1]，机器学习的理论极限是多少等关于理论极限的问题。著名的香农极限定理是通信资源利用率的理论上限[2]；信号处理中的测不准原理说明了时间和频率分辨率不能同时趋于零，二者乘积存在确定的下界[3]；计算机领域也存在著名的兰道尔原理，得出计算机能耗的理论下限[4-5]。在复杂网络中，链路预测是理解网络演化规律的重要方法。近年来，由于各种链路预测方法的相继提出和预测准确度的不断提高，链路预测的理论极限问题日益受到学者的广泛关注。

大量实证研究表明，现实网络是处于确定性和不确定性之间的一种关系结构[6-7]（既具有规则性也具有随机性），致使在探究链路的可预测性上存在一定难度。对于网络的可预测性极限问题，相关学者以网络结构为研究对象，研究网络可预测性理论极限，这也是研究链路预测理论极限问题的最终目

收稿日期：2019−12−17；**修回日期**：2020−03−06
基金项目：国家自然科学基金资助项目（No.61601513）；郑州市协同创新重大专项基金资助项目（No.162/32410218）
Foundation Items: The National Natural Science Foundation of China (No.61601513), Major Collaborative Innovation Projects of Zhengzhou (No.162/32410218)

的。Lyu 等[8]基于结构稳定性的假设，提出一种结构一致性指标作为测度衡量网络的规则性和可预测性，并提出基于该理论的新的链路预测方法——结构微扰法（SPM, structural perturbation method），使链路预测理论极限问题得到推进。以网络拓扑为研究对象探讨网络的可预测理论极限固然意义重大，但仍然存在一些难题，能否以某一类链路预测方法为研究对象，研究这类方法预测的理论极限问题。

近年来，学者针对链路预测问题做了大量研究，在向链路预测理论极限逼近的实践过程中，产生了大量的链路预测方法。无论是基于节点相似性的链路预测方法、基于概率模型方法、基于机器学习的分类方法以及指标融合法，其基本假设都是节点间相似性越大，它们之间存在链接的可能性就越大，最终都归结为得出网络中节点对的相似性矩阵，即得到节点对间一维化的相似性得分。本文将链路预测方法分为两类，从是否使用多维度信息或是否直接定义多维度信息之间关系的角度，将链路预测方法分为单机制方法和组合方法。

目前，大多数组合方法都具有较高的预测准确性和适用于多种类型网络的稳健性，这种准确性和稳健性双重优势的原因是什么；组合方法是否存在预测的理论极限，若存在，其具有怎样的形式和意义。本文针对链路预测的组合方法展开研究，逐一对上述问题给出理论解释，提出并证明组合方法理论极限的充分必要条件，得出使组合方法达到理论上限的变换函数的完全集合，并对变换函数做出直观的几何解释，进一步揭示链路预测组合方法的本质。理论极限定理具有重要的实际应用价值，直接揭示了组合方法的最终目的，以理论极限定理为指导，在完全的变换函数集合中选取变换函数的合适形式，并依据具体网络数据做适当的简化，可以达到计算复杂度和预测准确性的折中[9]。

2　研究现状

在无权无向静态网络的链路预测研究中，按照经典的分类方式，现有链路预测方法主要分为基于节点相似性的链路预测算法、节点相似性的指标融合算法、基于机器学习的链路预测算法以及基于似然分析的链路预测算法。上述四类方法从是否使用多维度信息或是否直接定义多维度信息之间关系的角度可以分为链路预测的单机制方法和组合方法。

链路预测的单机制方法使用单一维度的网络信息，如共同邻居（CN, common neighbor）仅使用节点对的共同邻居数[10]，偏好连接（PA, preferential attachment）仅使用节点的度等[11]；或者直接、明确地定义多维度信息之间的关系，如资源分配（RA, resource allocation）[12]定义节点对的共同邻居节点度的倒数和为节点对的相似性。

链路预测的组合方法利用多维度的网络结构信息，但多维度网络信息的组合方式和物理意义并不明确，往往使用组合规则或通过数据的优化拟合[12-13]，不同于单机制方法直接给出多维度信息之间关系的明确定义方式。

2.1　链路预测的单机制方法

链路预测的单机制方法从网络演化的某一演化机理出发，直接构造节点对间的相似性测度，基于节点对的共同邻居数、路径数、节点度及其加权变换，综合网络中节点附近的局部结构信息或网络的全局信息，得出相似性评分，并根据评分大小顺序确定链路是否存在。单机制方法具有理论简洁、效率较高的优点。不加区分地考虑节点对的共同邻居，可得到经典的 CN[10]；Zhou 等[12]基于网络中资源传输过程的基本机制，提出共同邻居加权的 RA；Yao 等[14]提出局部加权路径；刘树新等[15-16]提出基于局部拓扑信息加权的相似性；Kumar 等[17]定义二级邻居节点的聚集系数用于链路预测；文献[18-20]考虑到网络的社区信息，利用社区信息对经典相似性加权，或仅在节点所属社区内计算经典相似性，提升链路预测准确度。

2.2　链路预测的组合方法

随着网络结构信息研究的不断深入，许多链路预测的单机制方法被相继提出。多维度的网络信息被充分挖掘；但单机制方法在某一类网络中表现较好，而其他类型的网络上表现一般，即在不同网络数据集上的算法稳健性不理想。为进一步提高单机制方法的准确性和稳健性，研究者从不同角度提出链路预测的组合方法。组合方法是将多种单机制方法或不同参数下的某种单机制方法通过变换函数得到综合指标的链路预测方法，主要分为组合规则法、网络模型法以及拟合学习法。

2.2.1　组合规则法

组合规则法对多种单机制方法按照规则策略（如多数投票、乘积规则、和式规则等）进行加权

组合，得到综合指标[21]。例如，选取CN、RA、PA等相似性指标的单机制方法，求其归一化得分并使用和式叠加规则，即可得到融合后的综合得分。但是这种组合方法容易得到中庸的结果，即融合得分的预测准确性介于各单机制方法预测准确性之间。其具体原因和各组合规则的理论解释将在第4节进行详细分析。

2.2.2 网络模型法

链路预测的网络模型法从宏观上对网络产生连边的机制进行建模，求得模型在各个参数下各节点对产生连边的概率，再对每种参数下的网络生成形式赋予权重，用各个参数下节点对产生连边概率的加权组合确定最终链路预测得分。典型的方法有随机分块模型[22-23]和层次结构模型[24]。网络模型法阐述了网络的生成演化机制，但最终链路预测得分的确定则根据组合规则，且组合函数是最简单的线性函数。该类方法的组合函数将在第4节进行详细分析。

2.2.3 拟合学习法

拟合学习法通过设置目标函数，根据训练集中正负例样本的实际数据分布对组合函数的非线性关系或组合系数反馈调节。典型方法有OWA（ordered weighted averaging）算子融合法[25]、模糊积分融合法[26]、基于逻辑回归的融合方法[13]、基于稀疏矩阵分解的融合方法[27]以及基于AdaBoost的融合方法[28]等。基于机器学习分类思想的链路预测方法将各种单机制相似性或其他特征输入分类器，对有、无连边这两类样本进行训练，按照分类器的输出进行判决，即将链路预测问题转化为机器学习的二分类问题[29-30]，因此朴素贝叶斯、逻辑回归、随机森林、支持向量机等多种分类器均可应用[31]，其本质上也属于多种单机制方法经分类器输出融合的拟合学习法。

近年来，深度学习技术在数据处理领域受到广泛关注，网络结构数据不易直接作为经典神经网络的输入，为解决深度学习方法适用于图数据的问题，研究者提出了网络表示学习方法和图神经网络。文献[32-34]对该类方法做了总结论述；文献[35-37]将网络的结构和属性信息向量拼接，通过表示学习方法得出带有结构和属性信息的节点向量表示。节点的向量表示同样需要通过拟合学习等方法得到节点对间的一维化得分用于链路预测。

3 评价指标和问题描述

3.1 链路预测算法的评价指标

3.1.1 AUC

为了评估算法的准确性，需要对网络的连边集合 E 进行训练集 E^{T} 和测试集 E^{P} 的划分，且满足 $E=E^{\mathrm{T}}\cup E^{\mathrm{P}}, E^{\mathrm{T}}\cap E^{\mathrm{P}}=\varnothing$。链路预测算法只允许运用 E^{T} 的信息进行预测。一般用AUC（area under the receiver operation characteristic curve）[39]准确度和精确度（Precision）衡量。

AUC不受有、无连边这两类样本非平衡性（即无连边的节点对远大于有连边的节点对数量）的影响。AUC可以理解为在测试集中随机选择一条边的分数值比随机选择一条不存在的边的分数值高的概率[39]。即每次从测试集中随机选择一条边，再从不存在的边中随机选择一条边，若前者高则加1分，若相等则加0.5分，这样独立比较 n 次。若有 n' 次测试集得分高，有 n'' 次二者相等，则AUC定义为

$$\mathrm{AUC}=\frac{n'+0.5n''}{n} \tag{1}$$

事实上，AUC定义为ROC（receiver operation characteristic）曲线下的面积[38]。ROC即为链路预测得分阈值变化时，不存在边的比例与测试边的比例关系曲线。根据文献[39]，AUC等价于在测试集中随机选择一条边的分数值比随机选择一条不存在的边的分数值高的概率，其形式化表述为，设随机变量 X 表示有连边的链路预测得分，X 服从概率密度 $f_X(x)$；随机变量 Y 表示不存在的连边链路预测得分，Y 服从概率密度 $g_Y(x)$，且 X 和 Y 相互独立。$\tilde{X}$ 是从 X 中抽取的简单随机样本，即 $\tilde{X}$ 与 X 独立同分布，$\tilde{X}$ 构成测试集得分，则AUC表示概率 $P\left(\tilde{X}>Y\right)=P(X>Y)$。

$$\begin{aligned}&P(X>Y)=\iint_{X>Y} f_X(x)g_Y(y)\mathrm{d}x\mathrm{d}y=\\&\frac{1}{2}\iint_{X>Y} f_X(x)g_Y(y)\mathrm{d}x\mathrm{d}y+\frac{1}{2}\left(1-\iint_{X\leqslant Y} f_X(x)g_Y(y)\mathrm{d}x\mathrm{d}y\right)=\\&\frac{1}{2}\iint \mathrm{sgn}(x-y)f_X(x)g_Y(y)\mathrm{d}x\mathrm{d}y+\frac{1}{2}=\\&\frac{1}{2}\mathbb{E}\left[\mathrm{sgn}\left(X-Y\right)+1\right]\end{aligned} \tag{2}$$

其中

$$\mathrm{sgn}(x)=\begin{cases}1, & x>0\\ 0, & x=0\\ -1, & x<0\end{cases} \tag{3}$$

根据式(2)所阐述 AUC 的定义和原理，可以得到式(1)中的 AUC 计算方法。

ROC 的横坐标为假正确率（FPR, false positive rate），表示得分大于给定阈值 μ 为不存在边的概率，$\mathrm{FPR}=\int_{\mu}^{+\infty} g_Y(x)\mathrm{d}x$；纵坐标为真正确率（TPR, true positive rate），表示得分大于给定阈值为 μ 测试边的概率，即 $\mathrm{TPR}=\int_{\mu}^{+\infty} f_X(x)\mathrm{d}x$ [40]。

3.1.2　精确度

精确度定义为前 L 个预测边中预测准确的比例。若前 L 个预测边中有 m 条边在测试集中[41]，则精确度为

$$\mathrm{Precision}=\frac{m}{L} \tag{4}$$

与 AUC 类似，将式(4)写成随机变量形式。Precision 定义为大于给定阈值 μ 预测正确的样本的比例与大于给定阈值 μ 的总样本的比例之比，即 Precision 是 2 个概率之比，如式(5)所示。

$$\mathrm{Precision}=\frac{P(\omega_1)\int_{\mu}^{+\infty} f_X(x)\mathrm{d}x}{P(\omega_1)\int_{\mu}^{+\infty} f_X(x)\mathrm{d}x+P(\omega_2)\int_{\mu}^{+\infty} g_Y(x)\mathrm{d}x}=\frac{P(\omega_1)\mathrm{TPR}}{P(\omega_1)\mathrm{TPR}+P(\omega_2)\mathrm{FPR}} \tag{5}$$

其中，$P(\omega_1)$ 是有连边节点对的先验概率，$P(\omega_2)$ 是无连边节点对的先验概率。

3.2　链路预测组合方法理论极限的问题描述

由于每种单机制方法或某种单机制方法的不同参数在不同网络上的优势不同，组合方法将多个单机制方法的链路预测得分输入一个融合函数，得到综合得分，使对于任意给定网络，综合指标都体现出优于单机制方法的预测准确性。图 1 为链路预测的组合方法示意。

链路预测组合方法的理论极限问题可以给出如下数学描述。设随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 表示 n 个结构相似性指标给出的有连边节点对的得分值，服从 $f(\boldsymbol{x})=f(x_1,x_2,\cdots,x_n)$ 的联合分布，随机向量 $\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^{\mathrm{T}}$ 表示 n 个结构相似性指标给出的无连边节点对的得分值，服从 $g(\boldsymbol{x})=g(x_1,x_2,\cdots,x_n)$ 的联合分布。求变换函数 $l(\boldsymbol{x})$，使综合得分 $X=l(\boldsymbol{X})$、$Y=l(\boldsymbol{Y})$ 的 AUC 值达到最大，即 $P(X>Y)$ 达到最大。

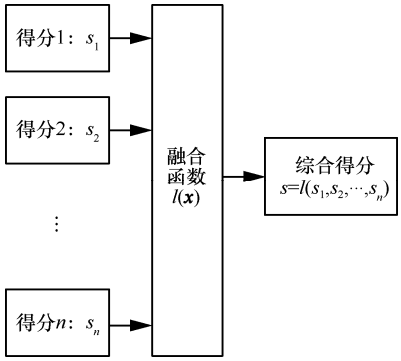


图 1　链路预测的组合方法示意

4　链路预测组合方法的理论极限定理

本文定理的证明均在附录中给出。为得出组合方法的理论极限定理，本文首先提出引理 1。

引理 1　$E\big(f(\boldsymbol{x})>\mu\big)$ 定义为集合 $\{\boldsymbol{x}\in\mathbb{R}^n: \mu\in\mathbb{R}, f(\boldsymbol{x})>\mu\}$。设 $f(\boldsymbol{x}), g(\boldsymbol{x})$ 为非负可测函数，$\mathbb{R}^n$ 上的非负递增的简单函数列 $\{\psi_\alpha(\boldsymbol{x})\}_{\alpha\geqslant 1}$ 和 $\{\varphi_\alpha(\boldsymbol{x})\}_{\alpha\geqslant 1}$ 满足 $\lim\limits_{\alpha\to\infty}\psi_\alpha(\boldsymbol{x})=f(\boldsymbol{x})$，$\lim\limits_{\alpha\to\infty}\varphi_\alpha(\boldsymbol{x})=g(\boldsymbol{x})$，则集合列 $E\left(\dfrac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})}>\mu\right)\big(\mu\in\mathbb{R}\big)$ 的极限存在，且

$$\lim_{\alpha\to\infty} E\left(\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})}>\mu\right)=E\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}>\mu\right) \tag{6}$$

定理 1　组合方法理论极限的充分条件。设随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 服从 $f(\boldsymbol{x})=f(x_1,x_2,\cdots,x_n)$ 的联合概率密度函数，随机向量 $\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^{\mathrm{T}}$ 服从 $g(\boldsymbol{x})=g(x_1,x_2,\cdots,x_n)$ 的联合概率密度函数，则变换函数 $l(\boldsymbol{x})=\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}$，$g(\boldsymbol{x})\neq 0$ 是使 AUC 达到最大的变换函数中的一种。

$\mathrm{AUC}=P(X>Y)=P\big(l(\boldsymbol{X})>l(\boldsymbol{Y})\big)=J\big(l(\boldsymbol{x})\big)$ 可以表示成泛函 $\mathrm{AUC}=J\big(l(\boldsymbol{x})\big)$ 关于宗量 $l(\boldsymbol{x})$（泛函中的函数变量称为宗量）的最大值问题[42]。该变分问题的表达式十分烦琐，不宜用变分法直接求解。根据 AUC 的等价定义，使 AUC 达到最大相当于 ROC 下的面积达到最大。若对任意 FPR，对应 ROC 上的每一点 TPR 的值最大，则 AUC 达到最大。

$$\mathrm{FPR}=\int_{\mu}^{\infty} g_Y(x)\mathrm{d}x=\int_{E(l(\boldsymbol{x})>\mu)} g(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{7}$$

$$\mathrm{TPR}=\int_{\mu}^{+\infty} f_X(x)\mathrm{d}x=\int_{E(l(\boldsymbol{x})>\mu)} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{8}$$

其中，$g_Y(x)$ 是 $Y=l(\boldsymbol{Y})$ 的概率密度函数，$E\left(l(\boldsymbol{x})>\mu\right)=\{\boldsymbol{x}\in\mathbb{R}^n:\mu\in\mathbb{R},l(\boldsymbol{x})>\mu\}$，$f_X(x)$ 是 $X=l(\boldsymbol{X})$ 的概率密度函数，$m\{\boldsymbol{x}:l(\boldsymbol{x})=C,\forall C\in\mathbb{R}\}=0$（$m$ 为集合的测度）。

问题转化为确定函数 $l_{\mathrm{FPR}}(\boldsymbol{x})$ 和 $\mu_{l,\mathrm{FPR}}$，使

$$\int_{E(l_{\mathrm{FPR}}(\boldsymbol{x})>\mu_{l,\mathrm{FPR}})} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\max_{l(\boldsymbol{x}),\mu_l}\int_{E(l(\boldsymbol{x})>\mu_l)} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{9}$$

即证明满足式(9)的 $l_{\mathrm{FPR}}(\boldsymbol{x})=\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}$。

该结论与奈曼皮尔逊准则[43-44]等价。附录 2)的证明中可以看出定理 1 的直观几何解释。

图 2 给出了定理 1 的几何解释和证明思路。设网络中有连边的节点对得分服从分布 $f(x)=0.64\exp\left(-2(x-1.8)^2\right)+0.32\exp\left(-8(x-2.8)^2\right)$，如图 2(a)的曲线 1 所示；无连边的节点对得分服从分布 $g(x)=1.56x\exp\left(-\dfrac{x}{0.8}\right)(x>0)$，如图 2(a)的曲线 2 所示；对 $f(x)$ 的错误估计 $\hat{f}(x)$ 如图 2(a)的曲线 3 所示。则存在简单函数列 $\psi_\alpha(x)\to f(x)$，$\varphi_\alpha(x)\to g(x)$。取简单函数 α=9，$\psi_9(x)=\sum_{i=1}^{9}a_i\chi_{H_i}(x)$ 逼近 $f(x)$；取 $\varphi_9(x)=\sum_{i=1}^{9}b_i\chi_{H_i}(x)$ 逼近 $g(x)$；取 $\hat{\psi}_9(x)=\sum_{i=1}^{9}\hat{a}_i\chi_{H_i}(x)$ 逼近 $f(x)$ 的错误估计 $\hat{f}(x)$，得到简单函数逼近密度函数的示意和对应的把简单函数当作密度函数的 ROC 曲线，如图 2(b)所示。

ROC 的横坐标表示对 $g(\boldsymbol{x})$ 的积分，纵坐标表示对 $f(\boldsymbol{x})$ 的积分。选择不同的变换函数 $l(\boldsymbol{x})$ 表示对 $f(\boldsymbol{x})$、$g(\boldsymbol{x})$ 的积分区域选取的顺序不同，但无论如何选取 $l(\boldsymbol{x})$，ROC 的横纵坐标都是对 $g(\boldsymbol{x})$ 和 $f(\boldsymbol{x})$ 的积分，证明中采用简单函数列逼近原概率密度函数的方法，使简单函数对应的 ROC 逼近原密度函数对应的 ROC。由于对于任意 α，$\int_{\mathbb{R}}\psi_\alpha(x)\mathrm{d}x<1$，$\int_{\mathbb{R}}\varphi_\alpha(x)\mathrm{d}x<1$，因此将 $\psi_\alpha(x),\varphi_\alpha(x)$ 乘以对应常数，使 $\tilde{\psi}_\alpha(x)=k_{\psi\alpha}\psi_\alpha(x)$，$\tilde{\varphi}_\alpha(x)=k_{\varphi\alpha}\varphi_\alpha(x)$，满足 $\int_{\mathbb{R}}\tilde{\psi}_\alpha(x)\mathrm{d}x=1$，$\int_{\mathbb{R}}\tilde{\varphi}_\alpha(x)\ \mathrm{d}x=1$ 后，再按照对应区域积分绘制简单函数列 $\tilde{\psi}_\alpha(x),\tilde{\varphi}_\alpha(x)$ 对应的 ROC。随着简单函数 α 取值的不断增大，对应的 ROC 逐渐逼近原密度函数对应的 ROC，如图 2(d)和图 2(f)所示。

容易看出，对于任意的 α，简单函数按照 $\dfrac{a_i}{b_i}$ 的降序沿对应的区域 H_i 积分得到的 ROC 下的面积最大。图 2(a)的简单函数列取 $\alpha=9$，其中曲线 3 选取的变换函数为 $\hat{l}(x)=\dfrac{\hat{l}(x)g(x)}{g(x)}=\dfrac{\hat{f}(x)}{g(x)}$，相当于按照 $\dfrac{\hat{a}_i}{b_i}$（而非 $\dfrac{a_i}{b_i}$）的降序沿对应区域 H_i 积分，所得的 AUC（即图 2(b)下侧的曲线下的面积）小于图 2(a)中曲线 1 的 AUC（图 2(b)上侧的曲线下的面积）。图 2(c)和图 2(e)是简单函数取 $\alpha=16$ 和 $\alpha=35$ 的示意图。因此，当 $\alpha\to\infty$ 时，按照 $\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}>\mu$（$\mu$ 从大到小连续变化）对应的 x 的集合进行积分，可使 AUC 达到最大，从而说明变换函数取 $l(x)=\dfrac{f(x)}{g(x)}$ 时的 AUC 最大；该结论容易推广到多维的情况，即对于 $\boldsymbol{x}\in\mathbb{R}^n$ 时取 $l(\boldsymbol{x})=\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}$ 同样成立。

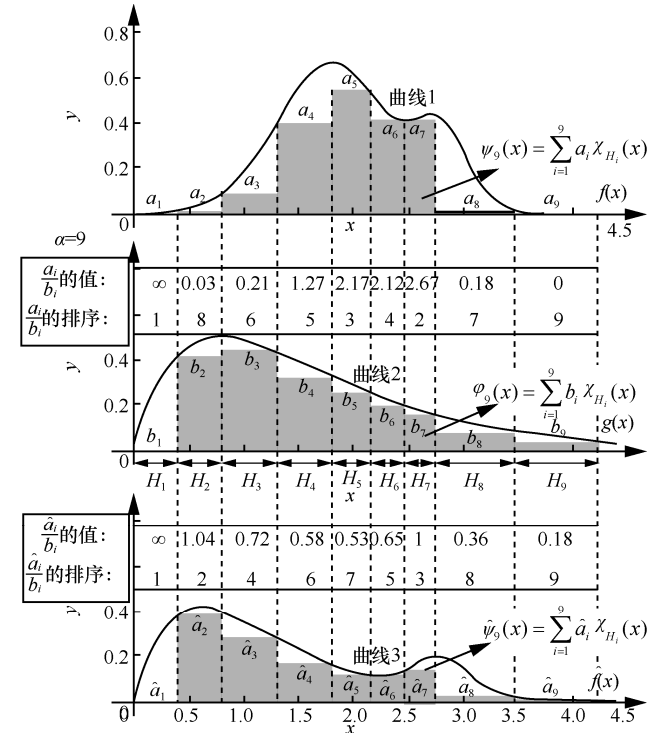
从定理 1 的结论可知，该充分条件同样适合于 Precision 指标的准确性理论极限。由式(5)可知，令 $k=\dfrac{\mathrm{TPR}}{\mathrm{FPR}}$ 表示 ROC 上任意一点到原点的割线的斜率，则 $\mathrm{Precision}=\dfrac{k}{k+\lambda}$，其中 $\lambda=\dfrac{P(\omega_2)}{P(\omega_1)}$，由定理 1 的证明过程，当变换函数取 $l(\boldsymbol{x})=\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}$ 时，对任意 FPR，TPR 达到最大，则割线斜率 k 达到最大，从而证明该变换函数可使阈值 μ 对应的 Precision 值达到最大。

引理 2 ROC 连续。

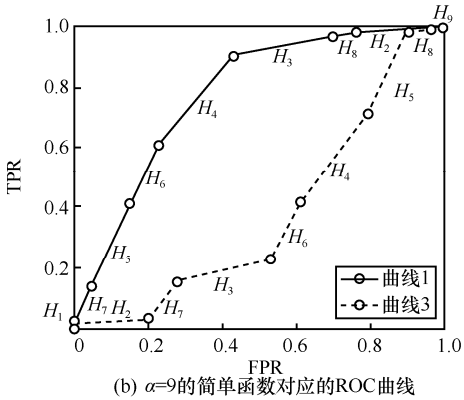
该引理的意义在于可以对任一点 $\mathrm{FPR}\in[0,1]$ 考察 ROC 和 AUC 的有关性质。

引理 3 设 $\boldsymbol{X}$ 的概率密度函数为 $f(\boldsymbol{x})$，$\boldsymbol{Y}$ 的概率密度函数为 $g(\boldsymbol{x})$。若不同的变换函数 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 对于任意的随机向量对 $(\boldsymbol{X},\boldsymbol{Y})$，可得到相同的 AUC，即对任意 $(\boldsymbol{X},\boldsymbol{Y})$，$P\left(l_1(\boldsymbol{X})>l_1(\boldsymbol{Y})\right)=P\left(l_2(\boldsymbol{X})>l_2(\boldsymbol{Y})\right)$，相同 AUC 值对应的 ROC 唯一。

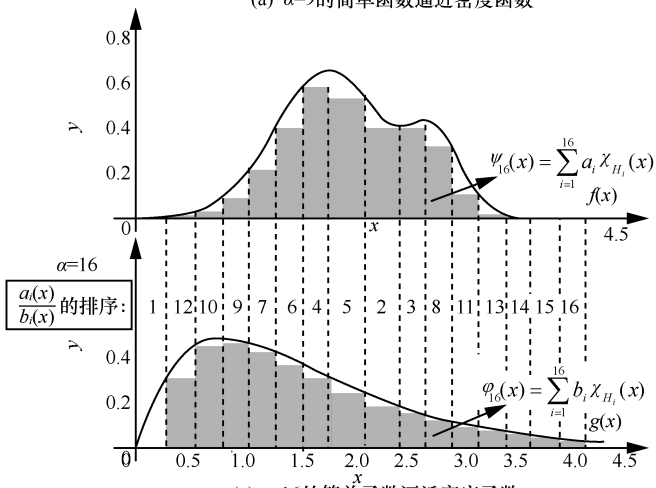
引理 3 的意义在于在普遍的意义上排除不同的 ROC 可以得到相同 AUC 的情况。虽然可以构造出随机向量对 $(\boldsymbol{X},\boldsymbol{Y})$ 经 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 变换后得到相同的



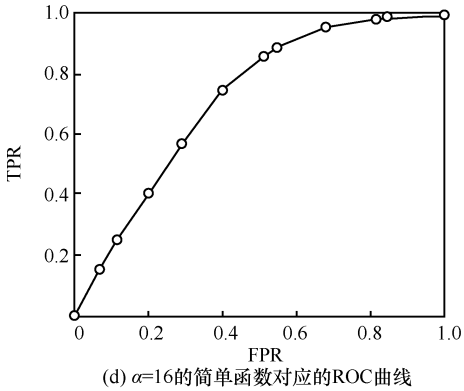
(a) α=9的简单函数逼近密度函数



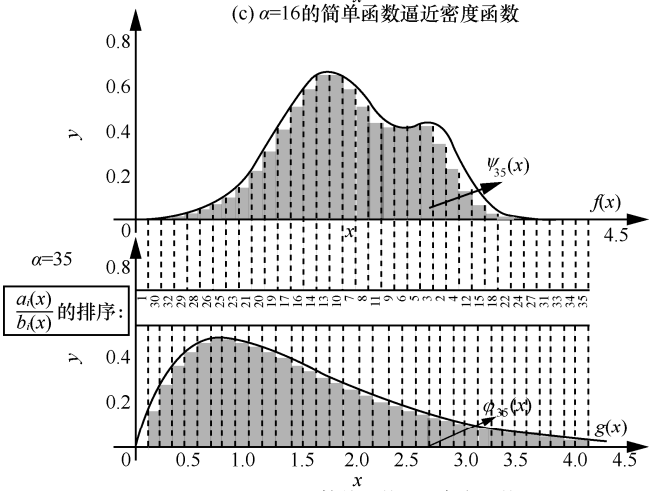
(b) α=9的简单函数对应的ROC曲线



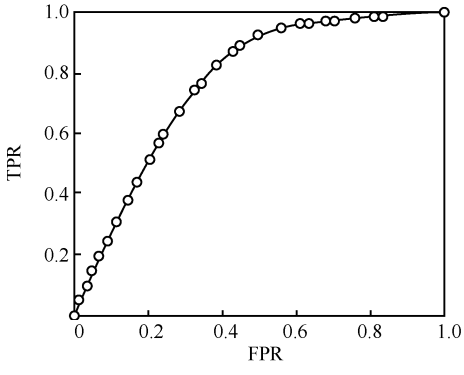
(c) α=16的简单函数逼近密度函数



(d) α=16的简单函数对应的ROC曲线



(e) α=35的简单函数逼近密度函数



(f) α=35的简单函数对应的ROC曲线

图 2　简单函数列逼近概率密度函数和对应的 ROC 曲线示意

AUC 而 ROC 不同（如图 3 所示），但是对于任意的随机向量对 $(\boldsymbol{X},\boldsymbol{Y})$ 不能普遍成立。根据定理 2，这相当于限制了 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 之间的关系，从而对于理论极限情况限制了变换函数的选取空间。

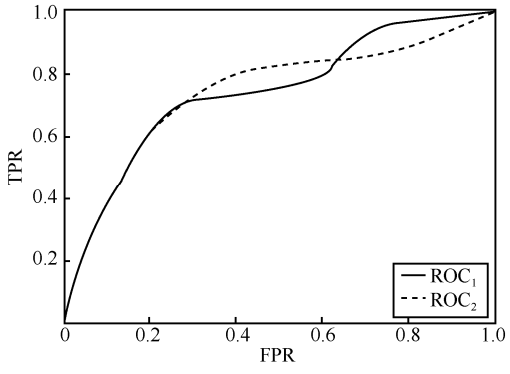


图 3 相同的 AUC 而 ROC 不同

定理 2 AUC 不变的条件。设 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 为 2 个变换函数，则下面 2 个条件等价。

1) 对任意的随机向量对 $(\boldsymbol{X},\boldsymbol{Y})$，$(\boldsymbol{X},\boldsymbol{Y})$ 经 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 变换后具有相同的 AUC，即 $P\left(l_1(\boldsymbol{X})>l_1(\boldsymbol{Y})\right)=P\left(l_2(\boldsymbol{X})>l_2(\boldsymbol{Y})\right)$。

2) 存在单调增函数 $r(x)$，使 $l_2(\boldsymbol{x})=r[l_1(\boldsymbol{x})]$，a.e. $\boldsymbol{x}\in\mathbb{R}^n$。

定理 3 组合方法理论极限的充分必要条件。设随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 服从 $f(\boldsymbol{x})=f(x_1,x_2,\cdots,x_n)$ 的联合概率密度函数，随机向量 $\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^{\mathrm{T}}$ 服从 $g(\boldsymbol{x})=g(x_1,x_2,\cdots,x_n)$ 的联合概率密度函数，且 $m\left\{\boldsymbol{x}:\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}=C,g(\boldsymbol{x})\neq 0,\forall C\in\mathbb{R}\right\}=0$（$m$ 表示集合的测度），则下面 2 个条件等价。

1) $l(\boldsymbol{x})$ 使 AUC 达到最大值。

2) 存在单调增函数 $r(x)$，使 $l(\boldsymbol{x})=r\left[\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right]$，$g(\boldsymbol{x})\neq 0$，a.e. $\boldsymbol{x}\in\mathbb{R}^n$。

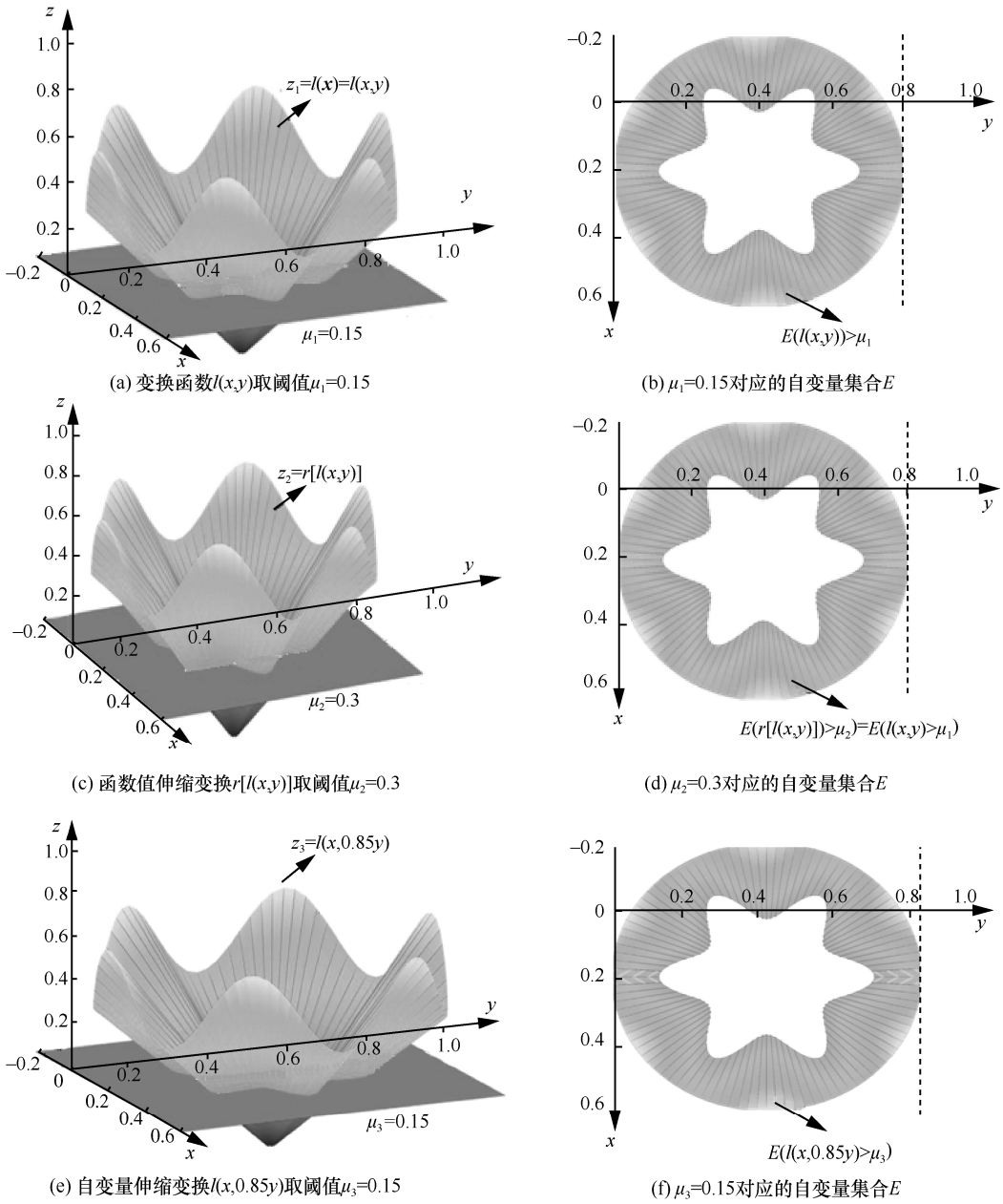
定理 3 表明，使链路预测准确性达到最大的组合函数是函数簇 $\boldsymbol{\Phi}=\left\{l(\boldsymbol{x}):l(\boldsymbol{x})=r\left[\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right],g(\boldsymbol{x})\neq 0\right\}$，$r(x)$ 是单调增函数。因此，组合方法的准确性一定大于或等于各个单机制方法的链路预测准确性，且在各类网络数据中理论上均可达到这一目的。这就是组合方法具有准确性和稳健性双重优势的原因。无论是后验概率的估计方法，还是链路预测得分的拟合方法，或是简单的组合规则法（如推论 3 所示），链路预测组合方法的本质是在估计有、无连边这两类样本的联合概率密度函数，目的是得到函数簇 $\boldsymbol{\Phi}$ 中的一种变换函数 $l(\boldsymbol{x})$。

定理 2 和定理 3 所述的内容可以用图 4 给出直观解释。$l(\boldsymbol{x})$ 为变换函数，随机向量 $(\boldsymbol{X},\boldsymbol{Y})$ 经变换函数变换后的 AUC 相当于按照变换函数设置单调减小的阈值，在变换函数值大于阈值所对应的自变量的集合上对 2 个随机向量概率密度函数进行积分，从而描绘出 ROC，并计算 ROC 的 AUC。若要求 2 个变换函数变换后的 AUC 值相等，相当于各个给定单调减小的阈值对应的自变量的集合可以顺次一一对应。为保证这一点，只能对变换函数 $l(\boldsymbol{x})$ 做函数值的伸缩变换，而不能对 $l(\boldsymbol{x})$ 进行自变量取值的伸缩变换。

设变换函数 $l(\boldsymbol{x})=l(x,y)$ 如图 4(a)所示，取阈值 μ_1=0.15，得到图 4(b)中 $l(x,y)>\mu_1$ 对应的自变量集合 $E\left(l(x,y)>\mu_1\right)$，图 4(b)为图 4(a)的俯视图（同理图 4(d)、4(f)为图 4(c)、4(e)的俯视图）；对图 4(a)的变换函数 $l(x,y)$ 值域实施单调增函数伸缩变换 $r(x)$，仍可得到一个与 $\mu_1=0.15$ 对应的阈值（该阈值不妨设为 $\mu_2=0.3$），使伸缩后 $r[l(x,y)]>\mu_2$ 对应的自变量集合保持不变，即 $E\left(r[l(x,y)]>\mu_2\right)=E\left(l(x,y)>\mu_1\right)$，如图 4(d)所示；然而无论对 $l(x,y)$ 的自变量做怎样的伸缩变换，对任意阈值 μ_1，自变量伸缩后的变换函数都无法找到相应阈值，得到与集合 $E\left(l(x,y)>\mu_1\right)$ 相等的集合。例如图 4(e)对自变量做伸缩变换 $l(x,0.85y)$，不存在对应的阈值 μ_3，使之满足 $E\left(l(x,0.85y)>\mu_3\right)=E\left(l(x,y)>\mu_1\right)$，如图 4(f)所示。

推论 1 设有连边节点对得分为随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$，且服从 $f(\boldsymbol{x})=f(x_1,x_2,\cdots,x_n)$ 的联合概率密度函数，无连边节点对得分为随机向量 $\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^{\mathrm{T}}$，且服从 $g(\boldsymbol{x})=g(x_1,x_2,\cdots,x_n)$ 的联合概率密度函数。有连边节点对的先验概率为 $P(\omega_1)$，无连边节点对的先验概率为 $P(\omega_2)$，且 $m\left\{\boldsymbol{x}:\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}=C,g(\boldsymbol{x})\neq 0,\forall C\in\mathbb{R}\right\}=0$（$m$ 表示集合的测度），则下面 2 个条件等价。

1) 对于任意 α，存在变换函数 $l(\boldsymbol{x})$ 的对应阈值 μ_l 满足 $\alpha=P(\omega_1)\int_{\mu_l}^{+\infty}f_X(x)\mathrm{d}x+P(\omega_2)\int_{\mu_l}^{+\infty}g_Y(x)\mathrm{d}x$，使变换函数 $l(\boldsymbol{x})$ 和阈值 μ_l 对应的 Precision 达到最大值。



(a) 变换函数$l(x,y)$取阈值μ_1=0.15　(b) μ_1=0.15对应的自变量集合E

(c) 函数值伸缩变换$r[l(x,y)]$取阈值μ_2=0.3　(d) μ_2=0.3对应的自变量集合E

(e) 自变量伸缩变换$l(x,0.85y)$取阈值μ_3=0.15　(f) μ_3=0.15对应的自变量集合E

图 4　定理 2 和定理 3 的示意

2) 存在单调增函数 $r(x)$，使 $l(\boldsymbol{x})=r\left[\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right]$，$g(\boldsymbol{x})\neq 0$，a.e. $\boldsymbol{x}\in\mathbb{R}^n$。

推论 2　理论极限得分与后验概率等价。设有连边节点对得分为随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$，且服从 $f(\boldsymbol{x})=f(x_1,x_2,\cdots,x_n)$ 的联合概率密度函数，无连边节点对得分为随机向量 $\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^{\mathrm{T}}$，且服从 $g(\boldsymbol{x})=g(x_1,x_2,\cdots,x_n)$ 的联合概率密度函数，有连边节点对的先验概率为 $P(\omega_1)$，无连边节点对的先验概率为 $P(\omega_2)$，则将样本有连边的后验概率作为综合得分可使 AUC 达到最大。

有连边节点对的后验概率为

$$P(\omega_1\mid\boldsymbol{x})=\frac{p(\boldsymbol{x}\mid\omega_1)P(\omega_1)}{\sum_{i=1}^{2}p(\boldsymbol{x}\mid\omega_i)P(\omega_i)}=\frac{f(\boldsymbol{x})P(\omega_1)}{f(\boldsymbol{x})P(\omega_1)+g(\boldsymbol{x})P(\omega_2)} \tag{10}$$

根据定理 3，设 $r(x)=\dfrac{x}{x+\lambda}$，$\lambda=\dfrac{P(\omega_2)}{P(\omega_1)}$，因为 $r'(x)=\dfrac{\lambda}{(x+\lambda)^2}>0$，$r(x)$ 为单调增函数。因此变换函数为

$$l(\boldsymbol{x})=r\left[\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right]=\frac{f(\boldsymbol{x})P(\omega_1)}{f(\boldsymbol{x})P(\omega_1)+g(\boldsymbol{x})P(\omega_2)} \tag{11}$$

式(11)与后验概率式(10)等价。

推论 2 表明，当变换函数 $l(\boldsymbol{x})$ 为式(11)时，融合后的链路预测得分即为样本 $\boldsymbol{x}$ 有连边的后验概率。根据定理 2，只要满足 $r(x)$ 是单调增函数的条件，即可保持 AUC 值不变，即 AUC 的值与 $r(x)$ 中 λ 的取值和样本的先验概率无关，从而 AUC 值不受到样本不平衡性的影响而改变。

推论 3 若定理 3 中增加各个维度相互独立的条件，则有

$$l(\boldsymbol{x})=\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}=\frac{f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n)}{g_{Y_1}(x_1)g_{Y_2}(x_2)\cdots g_{Y_n}(x_n)}=\frac{f_{X_1}(x_1)}{g_{Y_1}(x_1)}\frac{f_{X_2}(x_1)}{g_{Y_2}(x_1)}\cdots\frac{f_{X_n}(x_n)}{g_{Y_n}(x_n)}=s_1(x_1)s_2(x_2)\cdots s_n(x_n) \tag{12}$$

其中，$f_{X_i}(x),g_{Y_i}(x),i=1,2,\cdots,n$ 分别是随机向量 $\boldsymbol{X},\boldsymbol{Y}$ 的边缘概率密度函数，有

$$s_i(x)=\frac{f_{X_i}(x)}{g_{X_i}(x)},i=1,2,\cdots,n \tag{13}$$

根据推论 2，式(12)可通过单调增函数 $r(x)=\frac{x}{x+\lambda},\lambda=\frac{P(\omega_2)}{P(\omega_1)}$ 写成后验概率的形式

$$r\left[l(\boldsymbol{x})\right]=\frac{s_1(x_1)s_2(x_2)\cdots s_n(x_n)}{s_1(x_1)s_2(x_2)\cdots s_n(x_n)+\lambda} \tag{14}$$

容易看出，式(14)与朴素贝叶斯方法等价。

推论 3 表明，在各维度相互独立的条件下，组合方法理论极限得分相当于各个维度理论极限得分的乘积。在实际应用中可以适时做出独立的假设，根据该推论简化计算复杂度。基于此，可以得到一些简单有效的组合规则，从而对一些链路预测组合方法给出解释。

若在各维度相互独立的条件下，对各维度得分 $s_i(x)=\frac{f_{Xi}(x)}{g_{Xi}(x)},i=1,2,\cdots,n$，取单调增函数变换 $r(x)=\frac{x}{(x+\lambda)},\lambda=\frac{P(\omega_2)}{P(\omega_1)}$，则各个维度的后验概率之积（PPP, product of the posterior probabilitiy）可以表示为

$$l_{\mathrm{PPP}}(\boldsymbol{x})=P(\omega_1\mid x_1)P(\omega_1\mid x_2)\cdots P(\omega_1\mid x_n)=\frac{s_1(x_1)}{s_1(x_1)+\lambda}\frac{s_2(x_2)}{s_2(x_2)+\lambda}\cdots\frac{s_n(x_n)}{s_n(x_n)+\lambda} \tag{15}$$

由于 $\lambda=\frac{P(\omega_2)}{P(\omega_1)}\gg s_i(x),i=1,2,\cdots,n$，因此

$$l_{\mathrm{PPP}}(\boldsymbol{x})\approx\frac{1}{\lambda^n}s_1(x_1)s_2(x_2)\cdots s_n(x_n) \tag{16}$$

式(16)与式(12)等价，由此得到后验概率形式的乘积规则。由于在实际应用中，这 2 种乘积规则的条件大多不能严格满足，因此二者的组合效果依具体的数据特点而定，不存在严格的孰优孰劣。

设 $P(\omega_1\mid x_i)=P(\omega_1)(1+\delta_i),\delta_i\ll 1$，即设后验概率与先验概率接近，表示成先验概率的微小波动，则求和规则 1 PPS（posterior probability sum）为

$$l_{\mathrm{PPS}}(\boldsymbol{x})=\sum_{i=1}^{n}P(\omega_1\mid x_i)=\sum_{i=1}^{n}P(\omega_1)(1+\delta_i)=nP(\omega_1)+P(\omega_1)\sum_{i=1}^{n}\delta_i \tag{17}$$

对 PPP 做单调变换

$$\log\left[l_{\mathrm{PPP}}(\boldsymbol{x})\right]=\log\left[\prod_{i=1}^{n}P(\omega_1\mid x_i)\right]=\sum_{i=1}^{n}\log\left[P(\omega_1\mid x_i)\right]=\sum_{i=1}^{n}\log\left[P(\omega_1)(1+\delta_i)\right]\approx n\log\left[P(\omega_1)\right]+\sum_{i=1}^{n}\delta_i \tag{18}$$

一定条件下乘积规则可以转化为求和规则。类似地，可定义求和规则 2 OS（odd sum），记作 $l_{\mathrm{OS}}(\boldsymbol{x})=s_1(x_1)+s_2(x_2)+\cdots+s_n(x_n)$。由于条件不能严格满足，求和规则与乘积规则的效果也依具体的数据不同而不同。

进一步地，可以将求和规则中的各项赋予权值，得到加权形式的求和（WPPS, weighted posterior probability sum）规则，如式(19)所示。

$$l_{\mathrm{WPPS}}(\boldsymbol{x})=\sum_{i=1}^{n}w_iP(\omega_1\mid x_i) \tag{19}$$

需要说明的是，无论哪种组合规则，变换函数中各个单一维度的链路预测得分必须使用该维度下有、无连边链路的相对得分。各个维度的原始得分，如待融合的 CN、AA 等得分，均是该维度节点对的绝对得分。绝对得分与相对得分在单一维度的 AUC 评价或 Precision 评价中影响不大，但是对于组合方法，则有较大差别，因为绝对得分的尺度不

统一，无法合理地使用组合规则。若对其进行简单的归一化操作，如同时除以最大值等简单的归一化方法，得到的只不过是在新的打分区间下的绝对得分，没有体现有、无连边链路得分的相对性。因此必须通过式(13)的方式或其等价形式得到相对得分后再使用组合规则。

链路预测经典的随机分块模型和层次结构模型中链路预测得分的确定，与式(19)带有权值的求和规则具有相同的形式，即选取的变换函数 $l(\boldsymbol{x})$ 为线性函数，具体推导将在附录中给出。因此，这 2 种模型本质上也属于链路预测的组合方法，可以得到很好的预测效果；但是，理论上基于这 2 种方法一定存在一种更好的组合方式，可使预测准确性再推进一步。然而，这 2 种模型的计算复杂度已经相当高，若后面的加权组合再采用基于理论极限定理的方法，会带来更大的计算量，并且失去了原模型清晰直观的物理意义，不再具有良好的解释性。

推论 4　随机向量 $\boldsymbol{X},\boldsymbol{Y}$ 分别服从联合概率密度函数 $f(\boldsymbol{x}),g(\boldsymbol{x})$，若变换函数 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 满足 $l_2(\boldsymbol{x})=s[l_1(\boldsymbol{x})]$，其中 $s(x)$ 为单调减函数。设 $\boldsymbol{X},\boldsymbol{Y}$ 经 $l_1(\boldsymbol{x})$ 变换后的 AUC 为 AUC_1，经 $l_2(\boldsymbol{x})$ 变换后的 AUC 为 AUC_2，则 $\mathrm{AUC}_1+\mathrm{AUC}_2=1$，即 $P(l_1(\boldsymbol{X})>l_1(\boldsymbol{Y}))+P\big(l_2(\boldsymbol{X})>l_2(\boldsymbol{Y})\big)=1$。

定理 4　斜率定理。设随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 服从 $f(\boldsymbol{x})=f(x_1,x_2,\cdots,x_n)$ 的联合概率密度函数，随机向量 $\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^{\mathrm{T}}$ 服从 $g(\boldsymbol{x})=g(x_1,x_2,\cdots,x_n)$ 的联合概率密度函数。当变换函数取 $l(\boldsymbol{x})=\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})},g(\boldsymbol{x})\neq 0$ 时，ROC 在可导点 FPR 处的斜率是该点对应的阈值 μ_{FPR}。

对于变换函数 $l(\boldsymbol{x})$ 取其他函数的一般情况，ROC 的斜率为

$$k=\lim_{\mu_\alpha\to\mu_{\mathrm{FPR}}}\frac{\int_{E(l(\boldsymbol{x})\geqslant\mu_{\mathrm{FPR}})}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\int_{E(l(\boldsymbol{x})\geqslant\mu_\alpha)}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\int_{E(l(\boldsymbol{x})\geqslant\mu_{\mathrm{FPR}})}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\int_{E(l(\boldsymbol{x})\geqslant\mu_\alpha)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}=\lim_{\mu_\alpha\to\mu_{\mathrm{FPR}}}\frac{\int_{E(\mu_{\mathrm{FPR}}\leqslant l(\boldsymbol{x})<\mu_\alpha)}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\int_{E(\mu_{\mathrm{FPR}}\leqslant l(\boldsymbol{x})<\mu_\alpha)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}\tag{20}$$

对于一般情况，尽管 $\lim\limits_{\mu_\alpha\to\mu_{\mathrm{FPR}}}\int_{E(\mu_{\mathrm{FPR}}\leqslant l(\boldsymbol{x})<\mu_\alpha)}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\int_{E(l(\boldsymbol{x})=\mu_{\mathrm{FPR}})}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$，但仍不知道如何计算式(20)的极限或用什么方法来逼近这个极限。本文猜测该极限是 $f(\boldsymbol{x})$ 和 $g(\boldsymbol{x})$（$\boldsymbol{x}\in\mathbb{R}^n$）在 n 维空间 $E\big(l(\boldsymbol{x})=\mu_{\mathrm{FPR}}\big)$ 上的广义线积分之比。

定理 4 给出了理论极限情况下 ROC 斜率的物理意义，表明理论极限状态 ROC 是斜率单调减小的凸曲线，若给定 ROC 的斜率不满足单调减小，可判定该曲线未达到理论极限状态，方法的 AUC 还可进一步提升。

5　仿真示例

仿真示例模拟 4 种结构相似性指标对存在连边和不存在连边的节点对进行打分，给出 4 种结构相似性指标的联合概率密度函数，根据联合概率密度函数生成存在连边的样本观测值 10 000 个，不存在连边的样本观测值 100 000 个（每个样本观测值都是 4 维），并在 10 000 个存在连边的样本中随机选取 1 000 个样本作为测试集，其余 9 000 个样本作为训练集，1 000 个测试样本和 10 0000 个不存在连边的样本共同构成未知连边的训练样本。首先分别将各个单一维度相似性指标的样本观测值作为得分值，计算链路预测的 AUC、Precision；然后使用各类组合方法（包括绝对得分的求和、乘积规则，OS、PPS、PPP 等组合规则，朴素贝叶斯法，逻辑回归法等）得到综合得分，计算 AUC、Precision；最后根据组合方法的理论极限定理得到融合得分的理论值，计算 AUC、Precision，并将上述 AUC、Precision 列表对比分析。仿真中选择多元正态分布和任意构造的多元分布作为 4 种结构相似性指标的联合概率密度函数，用数据直观示例说明理论极限定理 1、定理 3 以及定理 2 和推论 2、推论 3 的含义。

5.1　多元正态分布

设随机向量 $\boldsymbol{X}$ 表示存在连边节点对的得分，服从 $f(\boldsymbol{x})$，$\boldsymbol{Y}$ 表示不存在连边节点对的得分，服从 $g(\boldsymbol{x})$。$f(\boldsymbol{x})$，$g(\boldsymbol{x})$ 是 4 元正态分布，即

$$f(\boldsymbol{x})=\frac{1}{(2\pi)^{\frac{p}{2}}\left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\big(\boldsymbol{x}-\boldsymbol{\mu}\big)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{x}-\boldsymbol{\mu}\big)\right\}$$

其中，$\mathrm{diag}\big(\boldsymbol{\Sigma}\big)\boldsymbol{1}=(\sigma_1^2,\sigma_2^2,\sigma_3^2,\sigma_4^2)^{\mathrm{T}}$，$\boldsymbol{\Sigma}_{ij}=\boldsymbol{r}_{ij}\sigma_i\sigma_j$。

设 4 种结构相似性指标在某网络中对存在连边的节点对的打分服从均值向量为 $\boldsymbol{\mu}_f$，协方差矩阵为 $\boldsymbol{\Sigma}_f$ 的 4 元正态分布，该 4 种指标对不存在连边的节点对的打分服从均值向量为 $\boldsymbol{\mu}_g$，协方差矩阵为

$\boldsymbol{\Sigma}_g$ 的 4 元正态分布。2 组仿真示例的参数集合分别为 $\Theta_{1f}=\{\boldsymbol{\mu}_{1f},\boldsymbol{\Sigma}_{1f}\}$ 和 $\Theta_{1g}=\{\boldsymbol{\mu}_{1g},\boldsymbol{\Sigma}_{1g}\}$ 以及 $\Theta_{2f}=\{\boldsymbol{\mu}_{2f},\boldsymbol{\Sigma}_{2f}\}$ 和 $\Theta_{2g}=\{\boldsymbol{\mu}_{2g},\boldsymbol{\Sigma}_{2g}\}$，其中 $\boldsymbol{\mu}_{1f}=(1,2,1.7,2.1)^{\mathrm{T}}$，$\boldsymbol{\mu}_{1g}=(1.3,2.5,2.1,2.8)^{\mathrm{T}}$，$\boldsymbol{\mu}_{2f}=(1,2,1.7,2.1)^{\mathrm{T}}$，$\boldsymbol{\mu}_{2g}=(1.5,3.5,2.8,3)^{\mathrm{T}}$。$\mathrm{diag}(\boldsymbol{\Sigma}_{1f})\mathbf{1}=(1.5^2,2.2^2,3^2,2.5^2)^{\mathrm{T}}$，$\mathrm{diag}\left(\boldsymbol{\Sigma}_{1g}\right)\mathbf{1}=\left(2^2,2.2^2,3^2,2.5^2\right)^{\mathrm{T}}$，$\mathrm{diag}(\boldsymbol{\Sigma}_{2f})\mathbf{1}=(1.5^2,2.2^2,3^2,2.5^2)^{\mathrm{T}}$，$\mathrm{diag}\left(\boldsymbol{\Sigma}_{2g}\right)\mathbf{1}=\left(2.5^2,3.5^2,4^2,2.5^2\right)^{\mathrm{T}}$。

$$\boldsymbol{r}_{1f}=\boldsymbol{r}_{1g}=\begin{bmatrix}1 & 0.8 & 0.76 & 0.56\\ 0.8 & 1 & 0.85 & 0.74\\ 0.76 & 0.85 & 1 & 0.93\\ 0.56 & 0.74 & 0.93 & 1\end{bmatrix}$$

$$\boldsymbol{r}_{2f}=\boldsymbol{r}_{2g}=\begin{bmatrix}1 & 0.62 & 0.45 & 0.34\\ 0.62 & 1 & 0.28 & 0.47\\ 0.45 & 0.28 & 1 & 0.65\\ 0.34 & 0.47 & 0.65 & 1\end{bmatrix}$$

各个单一维度（模拟单机制方法）、各典型组合方法、理论极限的 AUC、Precision（L=100）以及单调增（减）函数变换后的 AUC、Precision 如表 1 所示。

表 1　　多元正态分布的仿真结果

参数	$f:\Theta=\Theta_{1f}$ $g:\Theta=\Theta_{1g}$		$f:\Theta=\Theta_{2f}$ $g:\Theta=\Theta_{2g}$	
	AUC	Precision	AUC	Precision
维度 1	0.554	0.047	0.569	0.114
维度 2	0.566	0.015	0.660	0.140
维度 3	0.547	0.014	0.604	0.081
维度 4	0.585	0.027	0.622	0.038
绝对得分的求和规则	0.501	0.002	0.500	0.008
求和规则 1	0.612	0.047	0.767	0.185
求和规则 2	0.612	0.044	0.766	0.169
绝对得分的乘积规则	0.479	0.002	0.499	0.005
乘积规则	0.610	0.036	0.763	0.132
朴素贝叶斯	0.610	0.038	0.765	0.153
逻辑回归	0.668	0.020	0.676	0.051
理论极限	0.738	0.120	0.792	0.241
单调减函数变换	0.262	—	0.208	—
单调增函数变换	0.738	0.120	0.792	0.241

5.2　构建分布

设随机变量 X_1 服从参数为 λ、平移为 t 的指数分布，即

$$f(x)=\begin{cases}\lambda \mathrm{e}^{-\lambda(x-t)}, x\geqslant 0\\ 0 \qquad\qquad , x<0\end{cases}$$

随机变量 X_2 服从参数为 κ、平移为 τ 的指数分布。随机变量 X_3 服从参数为 α,β 的 Γ 分布。

$$f(x)=\begin{cases}\dfrac{x^{\alpha}\mathrm{e}^{-\frac{x}{\beta}}}{\beta^{\alpha+1}\Gamma(\alpha+1)}, x\geqslant 0\\ 0 \qquad\qquad\qquad , x<0\end{cases}$$

随机变量 X_4 服从均值为 μ、方差为 σ^2 的正态分布。

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(x-\mu)^2}{\sigma^2}\right]$$

设 X_1,X_2,X_3,X_4 相互独立，则随机变量 X_1,X_2,X_3,X_4 的联合概率密度函数为

$$f_1(x_1,x_2,x_3,x_4)=\begin{cases}\dfrac{\lambda\kappa}{\sqrt{2\pi}\sigma\beta^{\alpha+1}\Gamma(\alpha+1)}x_3^{\alpha}\cdot\\ \exp\left[-\lambda(x_1-t)-\kappa(x_2-\tau)-\dfrac{x_3}{\beta}-\dfrac{(x_4-\mu)^2}{\sigma^2}\right], x_1,x_2,x_3\geqslant 0\\ 0, \text{其他}\end{cases}$$

设 $\underset{4\times 4}{\boldsymbol{A}}$ 为可逆阵，$\boldsymbol{Y}=\left(Y_1,Y_2,Y_3,Y_4\right)^{\mathrm{T}}$，$\boldsymbol{X}=(X_1,X_2,X_3,X_4)^{\mathrm{T}}$，进行如式(21)所示的变换。

$$\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{X}=\left(u_1(X_1,\cdots,X_4),\cdots,u_4(X_1,\cdots,X_4)\right)^{\mathrm{T}} \tag{21}$$

则 $\boldsymbol{Y}=\left(Y_1,Y_2,Y_3,Y_4\right)^{\mathrm{T}}$ 的联合概率密度函数为

$$f(y_1,y_2,y_3,y_4)=\left|J\right|f\left[w_1(y_1,\cdots,y_4),\cdots,w_4(y_1,\cdots,y_4)\right] \tag{22}$$

其中，雅可比行列式 $J=\dfrac{\partial(x_1,x_2,x_3,x_4)}{\partial(y_1,y_2,y_3,y_4)}$，$\left|J\right|=\left|\det\left(\boldsymbol{A}^{-1}\right)\right|$，$w_1(y_1,\cdots,y_4),\cdots,w_4(y_1,\cdots,y_4)$ 为 $u_1(y_1,\cdots,y_4),\cdots,u_4(y_1,\cdots,y_4)$ 的反函数。设 $f(\boldsymbol{x})$ 和 $g(\boldsymbol{x})$ 的参数分别为

$$\Theta_f=\{(\lambda_f,t_f,\kappa_f,\tau_f,\alpha_f,\beta_f,\mu_f,\sigma_f)^{\mathrm{T}},\boldsymbol{A}_f\}$$

$$\Theta_g=\{(\lambda_g,t_g,\kappa_g,\tau_g,\alpha_g,\beta_g,\mu_g,\sigma_g)^{\mathrm{T}},\boldsymbol{A}_g\}$$

$$\Theta_{1f}=\{(\lambda_{1f},t_{1f},\kappa_{1f},\tau_{1f},\alpha_{1f},\beta_{1f},\mu_{1f},\sigma_{1f})^{\mathrm{T}},\boldsymbol{A}_{1f}\}=\{(2,0,2,2,1,2,1,2)^{\mathrm{T}},\boldsymbol{A}_1\}$$

$$\Theta_{1g}=\{(2,2,2,2.5,2,4,2,3)^{\mathrm{T}},\boldsymbol{A}_1\}$$

$$\Theta_{2f}=\{(0.5,0,1,1,1,1,2,4)^{\mathrm{T}},\boldsymbol{A}_2\}$$

$$\Theta_{2g}=\{(2,0.3,1,1.4,2,4,2.5,3)^{\mathrm{T}},\boldsymbol{A}_2\}$$

$$\boldsymbol{\Theta}_{3f}=\{(2,0,2,2,1,2,1,2)^{\mathrm{T}},\boldsymbol{A}_3\}$$

$$\boldsymbol{\Theta}_{3g}=\{(2,0.5,2,2.5,2,4,2,3)^{\mathrm{T}},\boldsymbol{A}_3\}$$

$$\boldsymbol{A}_1=\begin{bmatrix}0.86 & 0.32 & 0.23 & 0.81\\ 0.47 & 0.86 & 0.90 & 0\\ 0.91 & 0.47 & 0.23 & 0.60\\ 0.33 & 0.40 & 0.39 & 0.93\end{bmatrix}$$

$$\boldsymbol{A}_2=\begin{bmatrix}0.37 & 0.76 & 0.32 & 0.72\\ 0.31 & 0.09 & 0.99 & 0.67\\ 0.43 & 0.94 & 0.44 & 0.43\\ 0.76 & 0.29 & 0.76 & 0.44\end{bmatrix}$$

$$\boldsymbol{A}_3=\begin{bmatrix}0.9 & 0.1 & 0 & 0\\ 0.05 & 0.85 & 0.03 & 0.07\\ 0.1 & 0.03 & 0.75 & 0.21\\ 0.1 & 0.1 & 0.1 & 0.7\end{bmatrix}$$

仿真结果如表 2 所示。仿真示例模拟了各个单一维度相似性指标的链路预测的 AUC 和 Precision，并通过各类组合规则和组合方法对单机制方法进行组合。示例中，使用绝对得分的求和规则和乘积规则的链路预测准确性普遍较低，AUC 在 0.5 左右；而使用相对得分的组合规则相比绝对得分的准确度可大大提升。示例中所选用的组合方法，虽然在一些情况下效果不如单一维度的预测准确性，但是从理论上一定存在更佳的组合方式可获得大于（或等于）各个单一维度的准确性。仿真示例展示了各单一维度和组合方法距离组合方法理论极限的提升空间。仿真示例说明了变换函数的单调增函数变换不改变 AUC，单调减函数变换前后 AUC 之和为 1 的结论。

6　结束语

本文提出从是否使用多维度信息或是否直接定义多维度信息之间关系的角度，将链路预测方法分为单机制方法和组合方法；提出了链路预测组合方法的数学描述、链路预测组合方法理论极限的形式化表述。通过使用简单函数列逼近可测函数的方法（用离散逼近连续的思想）证明了使组合方法达到理论极限的充分条件，该充分条件即著名的奈曼–皮尔逊准则。根据这一证明过程可以看出该充分条件的几何解释。进而提出并证明了链路预测组合方法达到理论极限的充分必要条件，得到使链路预测准确性达到最大的组合函数的全体组成的函数簇，并对该充要条件做了几何解释。结论表明，组合方法的实质是有、无连边两类样本各个维度的联合概率密度函数的估计问题，组合方法预测准确性的理论极限是基于已知各维度及其关系的全部信息后所做预测的准确性。

基于极限定理，本文得出了理论极限得分与后验概率等价的推论，并对链路预测的各类组合方法给出理论解释，对它们之间的关系进行了梳理。

表 2　　**构建分布的仿真结果**

参数	$f:\boldsymbol{\Theta}=\boldsymbol{\Theta}_{1f}\ \ g:\boldsymbol{\Theta}=\boldsymbol{\Theta}_{1g}$		$f:\boldsymbol{\Theta}=\boldsymbol{\Theta}_{2f}\ \ g:\boldsymbol{\Theta}=\boldsymbol{\Theta}_{2g}$		$f:\boldsymbol{\Theta}=\boldsymbol{\Theta}_{3f}\ \ g:\boldsymbol{\Theta}=\boldsymbol{\Theta}_{3g}$	
	AUC	Precision	AUC	Precision	AUC	Precision
维度 1	0.769	0	0.575	0	0.813	0.033
维度 2	0.779	0.025	0.813	0.175	0.744	0.021
维度 3	0.709	0	0.671	0.042	0.909	0.048
维度 4	0.923	0.103	0.744	0.072	0.803	0.112
绝对得分的求和规则	0.529	0	0.503	0	0.514	0.001
求和规则 1	0.834	0.035	0.687	0.117	0.790	0.058
求和规则 2	0.831	0	0.688	0.104	0.798	0.106
绝对得分的乘积规则	0.531	0	0.508	0	0.576	0.002
乘积规则	0.836	0	0.710	0.068	0.837	0.134
朴素贝叶斯	0.838	0.017	0.710	0.108	0.830	0.100
逻辑回归	0.983	0.039	0.925	0.082	0.860	0.030
理论极限	0.999	0.482	0.991	0.461	0.972	0.185
单调减函数变换	0.001	—	0.009	—	0.028	—
单调增函数变换	0.999	0.482	0.991	0.461	0.972	0.185

本文给出了理论极限情况下 ROC 在可导点处斜率的物理意义，即对应点的两类概率密度函数之比的阈值。对于一般情况的 ROC 的斜率，本文尚未得出解析形式的数学表达式，但提出了关于解的形式的猜想。

理论极限定理对链路预测的组合方法做出了理论解释；极限定理也可对设计链路预测方法提供一定的理论指导。

附录 引理 1~推论 4 的证明

1) 引理 1 的证明

记 $E_\alpha = E\left(\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} > \mu\right)$，首先证明

$$\overline{\lim_{\alpha\to\infty}} E_\alpha = \bigcap_{k=1}^{\infty}\bigcup_{\alpha=k}^{\infty} E\left(\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} > \mu\right) = E\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu\right)$$

对任意 $\boldsymbol{x} \in E\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu\right), \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu$，由于 $\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} = \lim_{\alpha\to\infty}\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})}$，因此存在 $M \in \mathbb{N}$，使 $\alpha > M$ 时 $\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} > \mu$，$\alpha > M$ 时 $\boldsymbol{x} \in E\left(\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} > \mu\right)$，$\boldsymbol{x} \in \overline{\lim_{\alpha\to\infty}} E_\alpha$。

对任意 $\boldsymbol{x} \in \overline{\lim_{\alpha\to\infty}} E_\alpha$，命题 $\boldsymbol{x} \in \overline{\lim_{\alpha\to\infty}} E_\alpha \Rightarrow \boldsymbol{x} \in E\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu\right)$ 等价于 $\boldsymbol{x} \notin E\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu\right) \Rightarrow \boldsymbol{x} \notin \overline{\lim_{\alpha\to\infty}} E_\alpha$，等价于

$$\boldsymbol{x} \in E\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} \leqslant \mu\right) \Rightarrow \boldsymbol{x} \in \left(\overline{\lim_{\alpha\to\infty}} E_\alpha\right)^C = \bigcup_{k=1}^{\infty}\bigcap_{\alpha=k}^{\infty} E\left(\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} \leqslant \mu\right)$$

由于 $\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} \leqslant \mu$，因此存在 $M \in \mathbb{N}$，使 $\alpha > M$ 时 $\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} \leqslant \mu$，$\alpha > M$ 时 $\boldsymbol{x} \in E\left(\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} \leqslant \mu\right)$，则

$$\boldsymbol{x} \in \bigcup_{k=1}^{\infty}\bigcap_{\alpha=k}^{\infty} E\left(\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} \leqslant \mu\right)$$

同理可证

$$\underline{\lim_{\alpha\to\infty}} E_\alpha = \underline{\lim_{\alpha\to\infty}} E\left(\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} > \mu\right) = \bigcup_{k=1}^{\infty}\bigcap_{\alpha=k}^{\infty} E\left(\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} > \mu\right) = E\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu\right)$$

从而

$$\overline{\lim_{\alpha\to\infty}} E_\alpha = \underline{\lim_{\alpha\to\infty}} E_\alpha = \lim_{\alpha\to\infty} E_\alpha = E\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu\right)$$

证毕。

2) 定理 1 的证明

因为 $f(\boldsymbol{x}), g(\boldsymbol{x})$ 是 $\mathbb{R}^n$ 上的非负可测函数，则存在 $\mathbb{R}^n$ 上的非负递增的简单函数列 $\{\psi_\alpha(\boldsymbol{x})\}_{\alpha\geqslant 1}, \{\varphi_\alpha(\boldsymbol{x})\}_{\alpha\geqslant 1}$，使 $\lim_{\alpha\to\infty}\psi_\alpha(\boldsymbol{x}) = f(\boldsymbol{x}), \lim_{\alpha\to\infty}\varphi_\alpha(\boldsymbol{x}) = g(\boldsymbol{x})$。设 $\psi_\alpha(\boldsymbol{x}) = \sum_{i=1}^{\alpha} a_i\chi_{H_i}(\boldsymbol{x})$，$\varphi_\alpha(\boldsymbol{x}) = \sum_{i=1}^{\alpha} b_i\chi_{H_i}(\boldsymbol{x})$，$a_i \in \mathbb{R}, b_i \in \mathbb{R}, i = 1,2,\cdots,\alpha, \alpha \in \mathbb{Z}^+$，满足 $H_i \cap H_j = \varnothing, i \neq j$，且 $\bigcup_{i=1}^{\alpha} H_i = \mathbb{R}^n$，其中

$$\chi_A(\boldsymbol{x}) = \begin{cases} 1, \boldsymbol{x} \in A \\ 0, \boldsymbol{x} \notin A \end{cases}$$

对于任意的可测函数 $l(\boldsymbol{x})$，不妨设 $l(\boldsymbol{x}) = \frac{\hat{f}(\boldsymbol{x})}{g(\boldsymbol{x})}, g(\boldsymbol{x}) \neq 0$，则存在 $\mathbb{R}^n$ 上的非负递增简单函数列 $\{\hat{\psi}_\alpha(\boldsymbol{x})\}_{\alpha\geqslant 1}$，使 $\lim_{\alpha\to\infty}\hat{\psi}_\alpha(\boldsymbol{x}) = \hat{f}(\boldsymbol{x})$。设 $\hat{\psi}_\alpha(\boldsymbol{x}) = \sum_{i=1}^{\alpha} c_i\chi_{H_i}(\boldsymbol{x})$，$c_i \in \mathbb{R}$，$i = 1,2,\cdots,\alpha$。

$$\int_{E(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu_{\frac{f}{g}})} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_{E(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu_{\frac{f}{g}})} \lim_{\alpha\to\infty}\psi_\alpha(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \overset{(a)}{=}$$

$$\lim_{\alpha\to\infty}\int_{E(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu_{\frac{f}{g}})} \psi_\alpha(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \overset{(b)}{=} \lim_{\alpha\to\infty}\int_{E(\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} > \mu_{\frac{f}{g}})} \psi_\alpha(\boldsymbol{x})\mathrm{d}\boldsymbol{x} =$$

$$\lim_{\alpha\to\infty}\int_{E(\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} > \mu_{\frac{f}{g}})} \sum_{i=1}^{\alpha} a_i\chi_{H_i}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \lim_{m\to\infty}\sum_{i=1}^{m} a_{p_i} m(H_{p_i}) =$$

$$\lim_{m\to\infty}\sum_{i=1}^{m}\frac{a_{p_i}}{b_{p_i}}\left(b_{p_i} m(H_{p_i})\right) = \lim_{m\to\infty}\sum_{i=1}^{m}\frac{a_{p_i}}{b_{p_i}}\sigma_{p_i}$$

其中，等号(a)根据 levi 定理；等号(b)根据引理 1。$\{a_{p_i}\}_{1\leqslant i\leqslant m}$，$\{b_{p_i}\}_{1\leqslant i\leqslant m}$ $(m, \alpha \in \mathbb{Z}^+, m \leqslant \alpha)$ 是简单函数列 $\{\psi_\alpha(\boldsymbol{x})\}_{\alpha\geqslant 1}, \{\varphi_\alpha(\boldsymbol{x})\}_{\alpha\geqslant 1}$ 中数列 $\{a_i\}_{1\leqslant i\leqslant\alpha}, \{b_i\}_{1\leqslant i\leqslant\alpha}$ 按照 $\left\{\frac{a_i}{b_i}\right\}_{1\leqslant i\leqslant\alpha}$ 由大到小排序后满足 $\frac{a_i}{b_i} > \mu_{\frac{f}{g}}$ 的前 m 项对应的子列，$\{H_{p_i}\}_{i\geqslant 1}$ 是集合列 $\{H_i\}_{i\geqslant 1}$ 对应的子列。

由于 $\int_{E(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu_{\frac{f}{g}})} g(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \lim_{\alpha\to\infty}\int_{E(\frac{\psi_\alpha(\boldsymbol{x})}{\varphi_\alpha(\boldsymbol{x})} > \mu_{\frac{f}{g}})} \sum_{i=1}^{\alpha} b_i\chi_{H_i}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \lim_{m\to\infty}\sum_{i=1}^{m}\sigma_{p_i}$，对任意给定 α，都存在 $\sum_{i=1}^{m}\sigma_{p_i} \geqslant \sum_{i=1}^{n}\sigma_{q_i}$，且 $\lim_{m\to\infty}\sum_{i=1}^{m}\sigma_{p_i} = \lim_{n\to\infty}\sum_{i=1}^{m}\sigma_{q_i} = \int_{E(\frac{\hat{f}(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu_{\frac{\hat{f}}{g}})} g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$。显然 $\sum_{i=1}^{m}\frac{a_{p_i}}{b_{p_i}}\sigma_{p_i} \geqslant \sum_{i=1}^{n}\frac{a_{q_i}}{b_{q_i}}\sigma_{q_i}$，因此

$$\int_{E(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu_{\frac{f}{g}})} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \lim_{m\to\infty}\sum_{i=1}^{m}\frac{a_{p_i}}{b_{p_i}}\sigma_{p_i} \overset{(c)}{\geqslant}$$

$$\lim_{n\to\infty}\sum_{i=1}^{n}\frac{a_{q_i}}{b_{q_i}}\sigma_{q_i} = \int_{E(\frac{\hat{f}(\boldsymbol{x})}{g(\boldsymbol{x})} > \mu_{\frac{\hat{f}}{g}})} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

其中，$\{a_{q_i}\}_{1\leqslant i\leqslant n}$，$\{b_{q_i}\}_{1\leqslant i\leqslant n}$ $(n, \alpha \in \mathbb{Z}^+, n \leqslant \alpha)$ 是简单函数列 $\{\psi_\alpha(\boldsymbol{x})\}_{\alpha\geqslant 1}, \{\varphi_\alpha(\boldsymbol{x})\}_{\alpha\geqslant 11}$ 中数列 $\{a_i\}_{1\leqslant i\leqslant\alpha}, \{b_i\}_{1\leqslant i\leqslant\alpha}$ 按照

$\left\{\dfrac{c_i}{b_i}\right\}_{1\leqslant i\leqslant\alpha}$ 由大到小排序后满足 $\dfrac{c_i}{b_i}>\mu_{\hat{f}/g}$，且满足 $\sum_{i=1}^{m}\sigma_{p_i}\geqslant\sum_{i=1}^{n}\sigma_{q_i}$ 的前 n 项对应的子列，$\{H_{q_i}\}_{i\geqslant1}$ 是集合列 $\{H_i\}_{i\geqslant1}$ 对应的子列。当且仅当 $\dfrac{a_{p_i}}{b_{p_i}}=\dfrac{a_{q_i}}{b_{q_i}}$ 时取等号。

综上，对于任意 FPR，对应 ROC 上的每一点 TPR 的值最大，则随机向量对 $(\boldsymbol{X},\boldsymbol{Y})$ 经变换函数 $l(\boldsymbol{x})=\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}$ 变换后使 AUC 达到最大。

若存在 $C\in\mathbb{R}$，使 $m\left\{\boldsymbol{x}:\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}=C,\ g(\boldsymbol{x})\neq0\right\}\neq0$，即 $\int_{E(l(\boldsymbol{x})>\mu)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\neq\int_{E(l(\boldsymbol{x})\geqslant\mu)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$，$\int_{E(l(\boldsymbol{x})=\mu)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\neq0$。设 $\mathrm{FPR}_{\mu^-}=\int_{E(l(\boldsymbol{x})>\mu)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$，$\mathrm{FPR}_{\mu}=\int_{E(l(\boldsymbol{x})\geqslant\mu)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$，不妨规定 ROC 在 $\mathrm{FPR}\in[\mathrm{FPR}_{\mu^-},\mathrm{FPR}_{\mu}]$ 时为点 $(\mathrm{FPR}_{\mu^-},\mathrm{TPR}_{\mu^-})$ 到点 $(\mathrm{FPR}_{\mu},\mathrm{TPR}_{\mu})$ 的直线。事实上一定存在 $l(\boldsymbol{x})$，满足 $m\{x:l(x)=C,\forall C\in\mathbb{R}\}=0$，可得到与上述规定相同的 ROC。为了物理意义的明确性，仍然取变换函数为 $l(\boldsymbol{x})=\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}$。

证毕。

3) 引理 2 的证明

由于 $m\{\boldsymbol{x}:l(\boldsymbol{x})=C,\forall C\in\mathbb{R}\}=0$，$\mathrm{FPR}_{\mu}=\int_{E(l(\boldsymbol{x})\geqslant\mu)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$，$\mathrm{TPR}_{\mu}=\int_{E(l(\boldsymbol{x})\geqslant\mu)}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$，不妨设 $\mu_\alpha>\mu$，因此有

$$\lim_{\mu_\alpha\to\mu}\int_{E(l(\boldsymbol{x})\geqslant\mu)}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\int_{E(l(\boldsymbol{x})\geqslant\mu_\alpha)}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=$$
$$\lim_{\mu_\alpha\to\mu}\int_{E(\mu\leqslant l(\boldsymbol{x})<\mu_\alpha)}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=$$
$$\int_{\lim\limits_{\mu_\alpha\to\mu}\bigcap_{\mu_\alpha>\mu}E(\mu\leqslant l(\boldsymbol{x})<\mu_\alpha)}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=$$
$$\int_{E(l(\boldsymbol{x})=\mu)}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=0$$

同理 $\lim_{\mu_\alpha\to\mu}\int_{E(l(\boldsymbol{x})\geqslant\mu)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\int_{E(l(\boldsymbol{x})\geqslant\mu_\alpha)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=0$。由于对任意 $\mathrm{FPR}_{\mu}\in[0,1]$，$\lim_{\mathrm{FPR}\to\mathrm{FPR}_{\mu}}\mathrm{TPR}=\mathrm{TPR}_{\mu}$，因此 ROC 连续。

证毕。

4) 引理 3 的证明

设 $l_1(\boldsymbol{x})=\dfrac{f_1(\boldsymbol{x})}{g(\boldsymbol{x})}$，$l_2(\boldsymbol{x})=\dfrac{f_2(\boldsymbol{x})}{g(\boldsymbol{x})}$，设随机向量 $(\boldsymbol{X},\boldsymbol{Y})$ 在给定的 $f(\boldsymbol{x})$，$g(\boldsymbol{x})$ 的分布下经过变换函数 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 变换具有相同的 AUC 且 ROC 不同，则 2 个 ROC 必然存在交点，不妨设其在 $\mathrm{FPR}=a,\mathrm{FPR}=c$ 之间不同，且存在一个交点为 $\mathrm{FPR}=b$。由于 ROC 不同，说明任意 $\mathrm{FPR}\in[a,b)\cup(b,c]$，$E\left(l_1(\boldsymbol{x})>\mu_{1\mathrm{FP}}\right)\neq E\left(l_2(\boldsymbol{x})>\mu_{2\mathrm{FP}}\right)$（$\mu_{1\mathrm{FPR}}$ 表示变换函数 $l_1(\boldsymbol{x})$ 在 FPR 处的阈值）。问题等价于若相同 AUC 值对应的 ROC 不唯一，则存在随机向量 $(\boldsymbol{X},\boldsymbol{Y})$ 使不同的变换函数 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 得到的 AUC 值不同。现构造 $\tilde{f}(\boldsymbol{x})$ 的取值使 $(\tilde{\boldsymbol{X}},\boldsymbol{Y})$ 在变换函数 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 下取得不同的 AUC 值。令

$$\tilde{f}(\boldsymbol{x})=\begin{cases}kf_1(\boldsymbol{x}),k\in\mathbb{R},\boldsymbol{x}\in E\left(\mu_{1c}\leqslant l_1(\boldsymbol{x})\leqslant\mu_{1a}\right)\\ f(\boldsymbol{x}),\text{其他}\end{cases}$$

且满足 $\int_{\mathbb{R}^n}\tilde{f}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=1$，不妨设 $m(\{\boldsymbol{x}:l_1(\boldsymbol{x})=C,g(\boldsymbol{x})\neq0,\forall C\in\mathbb{R}\}\cap E(\mu_{1c}\leqslant l_1(\boldsymbol{x})\leqslant\mu_{1a}))=0$（$m$ 表示集合的测度）。根据定理 1，$(\tilde{\boldsymbol{X}},\boldsymbol{Y})$ 在经变换函数 $l_1(\boldsymbol{x})$ 变换在 $[a,c]$ 上可得到理论极限的 ROC，因此 $(\tilde{\boldsymbol{X}},\boldsymbol{Y})$ 经变换函数 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 变换后的 ROC 不同，且 AUC 值不同。若此时仍保持 AUC 值相同，说明 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 在 $[a,c]$ 上均得到理论极限的 ROC，即对任意 $\mathrm{FPR}\in[a,b)\cup(b,c]$ 对应的 TPR 值相同，与 $E\left(l_1(\boldsymbol{x})>\mu_{1\mathrm{FPR}}\right)\neq E\left(l_2(\boldsymbol{x})>\mu_{2\mathrm{FPR}}\right)$ 矛盾。因此若不同的变换函数 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 对于任意的随机向量 $(\boldsymbol{X},\boldsymbol{Y})$，可得到相同的 AUC，则相同 AUC 值对应的 ROC 唯一。

证毕。

5) 定理 2 的证明

条件2) $\Rightarrow$ 条件1) 设 $f_{X_1}(x)$ 是 $X_1=l_1(\boldsymbol{X})$ 的概率密度函数，$g_{Y_1}(x)$ 是 $Y_1=l_1(\boldsymbol{Y})$ 的概率密度函数，$r(x)$ 为单调增函数，则 $X_2=r(X_1)$ 的概率密度函数为：$f_{X_2}(x)=f_{X_1}\left(h(x)\right)h'(x)$；$Y_2=r(Y_1)$ 的概率密度函数为 $g_{Y_2}(x)=g_{Y_1}\left(h(x)\right)h'(x)$。其中 $h(x)$ 是 $r(x)$ 的反函数。则

$$\mathrm{AUC}=P(X_2>Y_2)=\int_{-\infty}^{+\infty}f_{X_2}(x)\int_{-\infty}^{x}g_{Y_2}(y)\mathrm{d}y\mathrm{d}x=$$
$$\int_{-\infty}^{+\infty}f_{X_1}\left(h(x)\right)h'(x)\int_{-\infty}^{x}g_{Y_1}\left(h(y)\right)h'(y)\mathrm{d}y\mathrm{d}x=$$
$$\int_{-\infty}^{+\infty}f_{X_1}(x)\int_{-\infty}^{x}g_{Y_1}(y)\mathrm{d}y\mathrm{d}x=P(X_1>Y_1)$$

条件1) $\Rightarrow$ 条件2)。根据引理 3，若不同的变换函数 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 对于任意的随机向量对 $(\boldsymbol{X},\boldsymbol{Y})$ 可得到相同的 AUC，则相同 AUC 值对应的 ROC 唯一。即任意 FPR 对应的 TPR 值唯一，对任意随机向量对 $(\boldsymbol{X},\boldsymbol{Y})$ 成立。因此要求满足以下条件：① 对于任意 $\mu_{1\mathrm{FPR}}$ 存在 $\mu_{2\mathrm{FPR}}$ 满足 $E\left(l_1(\boldsymbol{x})>\mu_{1\mathrm{FPR}}\right)=E\left(l_2(\boldsymbol{x})>\mu_{2\mathrm{FPR}}\right)$ (a.e.)，对于 FPR 在 $[0,1]$ 上几乎处处成立；② 对于任意 $\mu^*_{1\mathrm{FPR}}>\mu_{1\mathrm{FPR}}$，满足 $E\left(l_1(\boldsymbol{x})>\mu^*_{1\mathrm{FPR}}\right)=E\left(l_2(\boldsymbol{x})>\mu^*_{2\mathrm{FPR}}\right)$ 对应的 $\mu^*_{2\mathrm{FPR}}$ 一定有 $\mu^*_{2\mathrm{FPR}}>\mu_{2\mathrm{FPR}}$。若存在 $y_1=l_1(\boldsymbol{x})$ 的一个测度不为 0 的集合满足 $l_2(\boldsymbol{x})\neq r[l_1(\boldsymbol{x})]$，即 $m\{y_1:l_2(\boldsymbol{x})\neq r[l_1(\boldsymbol{x})]\}\neq0$，记 $\sigma=\{y_1:l_2(\boldsymbol{x})\neq r[l_1(\boldsymbol{x})]\}$。若 $y_1\in\sigma$，$l_2(\boldsymbol{x}),l_1(\boldsymbol{x})$ 满足函数关系 $l_2(\boldsymbol{x})=s[l_1(\boldsymbol{x})]$ 但 $s(\boldsymbol{x})$ 不为增函数，则任意 $\mu_1\in\sigma$ 无法满足条件2)；若满足 $l_2(\boldsymbol{x})\neq r[l_1(\boldsymbol{x})]$ (a.e.)的 $y_1=l_1(\boldsymbol{x})$ 的取值在集合 σ 中时，$l_2(\boldsymbol{x}),l_1(\boldsymbol{x})$ 不具有函数关系，则无法满足条件 1)和条件 2)。因此得到了不同的 ROC，根据引理 3 的逆

否命题，存在随机向量对 $(\boldsymbol{X},\boldsymbol{Y})$ 变换后的 AUC 不同，因此条件1) $\Rightarrow$ 条件2) 成立。

证毕。

6) 定理 3 的证明

条件2) $\Rightarrow$ 条件1) 。根据定理 1，变换函数 $l_1(\boldsymbol{x})=\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})},g(\boldsymbol{x})\neq 0$ 是使 AUC 达到最大的变换函数中的一种，结合定理 2 易知 条件2) $\Rightarrow$ 条件1) 成立。

条件1) $\Rightarrow$ 条件2) 。若存在 $l_2(\boldsymbol{x})\neq r[l_1(\boldsymbol{x})]$，$r(x)$ 为单调增函数，使 $(\boldsymbol{X},\boldsymbol{Y})$ 经 $l_2(\boldsymbol{x})$ 变换后同样达到最大 AUC，那么有最大AUC对应的ROC相同。由 $m\left\{\boldsymbol{x}:\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}=C,g(\boldsymbol{x})\neq 0,\forall C\in\mathbb{R}\right\}=0$，知 $l_1(\boldsymbol{x}),l_2(\boldsymbol{x})$ 对应的 2 个相同的 ROC 上的任意相同点 (FPR,TPR)，对应阈值 $\mu_{1\mathrm{FPR}}$，$\mu_{2\mathrm{FPR}}$ 满足 $E\left(l_1(\boldsymbol{x})>\mu_{1\mathrm{FPR}}\right)=E\left(l_2(\boldsymbol{x})>\mu_{2\mathrm{FPR}}\right)$ (a.e.)，对于任意 $\mu^*_{1\mathrm{FPR}}>\mu_{1\mathrm{FPR}}$，满足 $E\left(l_1(\boldsymbol{x})>\mu^*_{1\mathrm{FPR}}\right)=E\left(l_2(\boldsymbol{x})>\mu^*_{2\mathrm{FPR}}\right)$ 对应的 $\mu^*_{2\mathrm{FPR}}$ 一定有 $\mu^*_{2\mathrm{FPR}}>\mu_{2\mathrm{FPR}}$。根据定理 2 中条件1) $\Rightarrow$ 条件2) 的证明过程，这个条件与 $l_2(\boldsymbol{x})\neq r[l_1(\boldsymbol{x})]$ 矛盾。因此 条件1) $\Rightarrow$ 条件2) 成立。

条件 $m\left\{\boldsymbol{x}:\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}=C,g(\boldsymbol{x})\neq 0,\forall C\in\mathbb{R}\right\}=0$ 是为了排除当 $\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}=C$ 为常数时，变换函数定义在集合 $\sigma=\left\{\boldsymbol{x}:\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}=C,g(\boldsymbol{x})\neq 0,\forall C\in\mathbb{R}\right\}\bigcap\mathbb{R}^n$ 上的函数值可以任意选取的问题。例如，构造 $\tilde{\boldsymbol{X}}$ 的概率密度为

$$\tilde{f}(\boldsymbol{x})=\begin{cases}f(\boldsymbol{x}),\boldsymbol{x}\in\mathbb{R}^n\setminus\sigma\\ kg(\boldsymbol{x}),\boldsymbol{x}\in\sigma\end{cases},k\in\mathbb{R},k<\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})},\int_{\mathbb{R}^n}\tilde{f}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=1$$

取变换函数

$$l(\boldsymbol{x})=\begin{cases}\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})},\boldsymbol{x}\in\mathbb{R}^n\setminus\sigma\\ l^*(\boldsymbol{x}),\boldsymbol{x}\in\sigma\end{cases}$$

则无论 $l(\boldsymbol{x})$ 中的 $l^*(\boldsymbol{x})$ 如何选取，只要满足 $l^*(\boldsymbol{x})<\min\left(\dfrac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right)$，则变换函数 $l(\boldsymbol{x})$ 都可使 $\tilde{f}(\boldsymbol{x}),g(\boldsymbol{x})$ 的 AUC 达到最大。特别地，当 $f(\boldsymbol{x})=g(\boldsymbol{x}),\boldsymbol{x}\in\mathbb{R}^n$ 时，由于 $\boldsymbol{X},\boldsymbol{Y}$ 的分布相同，无论变换函数 $l(\boldsymbol{x})$ 如何选取，变换后的 AUC 均为 0.5，同时 0.5 也是该种情况的理论极限。

证毕。

7) 推论 1 的证明

设 $k=\dfrac{\mathrm{TPR}}{\mathrm{FPR}}$ 是 ROC 上某点到原点间割线的斜率，则 $\mathrm{Precision}=\dfrac{k}{k+\lambda}$，$\lambda=\dfrac{P(\omega_2)}{P(\omega_1)}$。对任意 α，变换函数 $l(\boldsymbol{x})$ 使 Precision 最大等价于 k 达到最大。由于 $\alpha=P(\omega_1)\int_{\mu_l}^{+\infty}f_X(x)\mathrm{d}x+P(\omega_2)\int_{\mu_l}^{+\infty}g_Y(x)\mathrm{d}x$，$\mathrm{TPR}=\int_{\mu_l}^{+\infty}f_X(x)\mathrm{d}x$，$k$ 达到最大等价于对任意 α，TPR 达到最大，即对于任意 $\mathrm{FPR}\in[0,1]$ 对应的 TPR 达到最大，从而 $l(\boldsymbol{x})$ 使 Precision 达到最大。

注：推论 1 中对于 Precision 理论极限的等价条件必须强调 α 的任意性，从而决定了阈值 μ_l 的任意性，若省略此条件，条件 2) $\Rightarrow$ 条件 1)成立，而条件1) $\Rightarrow$ 条件 2)不成立。实际上，α 表示全部数据的比例，每一个 α 决定一个阈值 μ_l。

证毕。

8) 随机分块模型组合函数的推导

随机分块模型假设网络被分成若干个群，2 个节点产生链路的概率只取决于节点所在的群。设 $\boldsymbol{\Omega}$ 为所有可能分群方案集合，$p\in\boldsymbol{\Omega}$ 为一具体方案，$\boldsymbol{A}^0$ 为观察到的网络连边情况。随机分块模型最终链路预测的结果表达式为

$$s_{xy}=P\left(\boldsymbol{A}_{xy}=1\,|\,\boldsymbol{A}^0\right)=\frac{\sum_{\omega\in\Omega}\left(\dfrac{|E|_{\omega,xy}+1}{|U|_{\omega,xy}+2}\right)\exp\left[-H(\omega)\right]}{\sum_{\omega\in\Omega}\exp\left[-H(\omega)\right]}$$

$$H(\omega)=\sum_{\alpha\leqslant\beta}\left[\ln\left(|U|_{\omega,\alpha\beta}+1\right)+\ln\left(C^{|E|_{\omega,\alpha\beta}}_{|U|_{\omega,\alpha\beta}}\right)\right]$$

其中，$|E|_{\omega,xy}$ 表示在方案 ω 中节点对 x,y 所在群（群内或群间）的实际连边数，$|U|_{\omega,xy}$ 表示节点所在群（群内或群间）的最大可能连边数。$|E|_{\omega,\alpha\beta}$ 表示群 α,β 间的实际连边数（允许 $\alpha=\beta$），$|U|_{\omega,\alpha\beta}$ 表示群 α,β 间的最大可能连边数。设

$$k_\omega=\exp\left[-H(\omega)\right]$$

$$p_{\omega,xy}=\frac{|E|_{\omega,xy}+1}{|U|_{\omega,xy}+2}$$

则有

$$s_{xy}=P\left(\boldsymbol{A}_{xy}=1\,|\,\boldsymbol{A}^0\right)=\frac{\sum_{\omega\in\Omega}p_{\omega,xy}k_\omega}{\sum_{\omega\in\Omega}k_\omega}=\sum_{\omega\in\Omega}\frac{k_\omega}{K}p_{\omega,xy}\tag{23}$$

其中，$K=\sum_{\omega\in\Omega}k_\omega$。

证毕。

9) 层次结构模型组合函数的推导

层次结构模型将网络结构用族谱树的形式表示，设 $\boldsymbol{\Omega}$ 表示网络中所有族谱树的集合，$\omega\in\boldsymbol{\Omega}$。网络的 N 个节点称为叶子节点，通过 N–1 个非叶子节点将它们联系起来。在每个族谱树 ω 中每个非叶子节点 r 将被赋予一个概率值 $p_{\omega,r}$。一对节点 x,y 产生链路的概率 $p_{\omega,xy}$ 由它们之间最近共同祖先的非叶子节点的概率值决定。当节点 r 为 x,y 在 ω

中的最近共同祖先时， $p_{\omega,xy}=p_{\omega,r}$ 。概率值 $p_{\omega,r}$ 的极大似然解与直观观察解相同。

$$p_{\omega,r}=\frac{|E|_{\omega,r}}{|U|_{\omega,r}}$$

其中， $|E|_{\omega,r}$ 表示非叶子节点 r 连接的 2 片叶子中节点间的实际节点连边数， $|U|_{\omega,r}$ 表示最大可能的连边数。层次结构模型将遍历所有可能的族谱树，最终链路预测评分值是在各族谱树中对应非叶子节点的概率值 k_ω 对该族谱树似然值 $k_{p_{\omega,xy}}$ 的加权平均。

$$s_{xy}=\sum_{\omega\in\Omega}k_\omega p_{\omega,xy} \tag{24}$$

式(23)与式(24)的组合函数相同。层次结构模型中若一个节点从属于某一叶子分支，其本身也属于上一级非叶子节点所属的叶子分支，即可以表示网络的层次结构特性。但是，节点不可从属于同级非叶子节点所属的其他叶子分支，即无法体现网络的重叠性。

证毕。

10) 推论 4 的证明

设 $X_1=l_1(\boldsymbol{X}),Y_1=l_1(\boldsymbol{Y});X_2=l_2(\boldsymbol{X}),Y_2=l_2(\boldsymbol{Y})$ 。则 $X_2=s(X_1),Y_2=s(Y_1)$ 。设 $s(x)$ 的反函数为 $t(x)$ 。由 $t'(x)=\dfrac{1}{s'(x)}$ 知， $s(x)$ 与 $t(x)$ 具有相同的单调性，同为减函数， $|t'(x)|=-t'(x)$ 。

$$\begin{aligned}
&\mathrm{AUC}_2=P(X_2>Y_2)=\int_{-\infty}^{+\infty}f_{X_2}(x)\int_{-\infty}^{x}g_{Y_2}(y)\mathrm{d}y\mathrm{d}x=\\
&\int_{-\infty}^{+\infty}-f_{X_1}\left(t(x)\right)t'(x)\int_{-\infty}^{x}-g_{Y_1}\left(t(y)\right)t'(y)\mathrm{d}y\mathrm{d}x=\\
&\int_{-\infty}^{+\infty}-f_{X_1}\left(t(x)\right)\left(1-\int_{-\infty}^{t(x)}g_{Y_1}(y)\mathrm{d}y\right)\mathrm{d}t(x)=\\
&\int_{-\infty}^{+\infty}f_{X_1}(x)\left(1-\int_{-\infty}^{x}g_{Y_1}(y)\mathrm{d}y\right)\mathrm{d}x=1-P(X_1>Y_1)
\end{aligned}$$

证毕。

11) 定理 4 的证明

由于

$$k=\lim_{\mu_\alpha\to\mu_{\mathrm{FPR}}}\frac{\int_{E\left(\frac{f(x)}{g(x)}\geqslant\mu_{\mathrm{FPR}}\right)}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\int_{E\left(\frac{f(x)}{g(x)}\geqslant\mu_\alpha\right)}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\int_{E\left(\frac{f(x)}{g(x)}\geqslant\mu_{\mathrm{FPR}}\right)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\int_{E\left(\frac{f(x)}{g(x)}\geqslant\mu_\alpha\right)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}=$$

$$\lim_{\mu_\alpha\to\mu_{\mathrm{FPR}}}\frac{\int_{E\left(\mu_{\mathrm{FPR}}\leqslant\frac{f(x)}{g(x)}<\mu_\alpha\right)}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\int_{E\left(\mu_{\mathrm{FPR}}\leqslant\frac{f(x)}{g(x)}<\mu_\alpha\right)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}=$$

$$\frac{\lim\limits_{\mu_\alpha\to\mu_{\mathrm{FPR}}}\int_{E\left(\mu_{\mathrm{FPR}}\leqslant\frac{f(x)}{g(x)}<\mu_\alpha\right)}\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\lim\limits_{\mu_\alpha\to\mu_{\mathrm{FPR}}}\int_{E\left(\mu_{\mathrm{FPR}}\leqslant\frac{f(x)}{g(x)}<\mu_\alpha\right)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}$$

因此

$$\lim_{\mu_\alpha\to\mu_{\mathrm{FPR}}}\mu_{\mathrm{FPR}}\leqslant k=\frac{\lim\limits_{\mu_\alpha\to\mu_{\mathrm{FPR}}}\int_{E\left(\mu_{\mathrm{FPR}}\leqslant\frac{f(x)}{g(x)}<\mu_\alpha\right)}\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\lim\limits_{\mu_\alpha\to\mu_{\mathrm{FPR}}}\int_{E\left(\mu_{\mathrm{FPR}}\leqslant\frac{f(x)}{g(x)}<\mu_\alpha\right)}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}\leqslant\lim_{\mu_\alpha\to\mu_{\mathrm{FPR}}}\mu_\alpha$$

根据夹逼定理可得， $k=\mu_{\mathrm{FPR}}$ 。

证毕。

参考文献：

[1] SEIFE C. What are the limits of conventional computing[J]. Science, 2005, 309(5731):96.

[2] SHANNON C E. A mathematical theory of communication[J]. ACM SIGMOBILE Mobile Computing and Communications Review, 2001, 5(1): 3-55.

[3] DONOHO D L, STARK P B. Uncertainty principles and signal recovery[J]. SIAM Journal on Applied Mathematics, 1989, 49(3): 906-931.

[4] LANDAUER R. Irreversibility and heat generation in the computing process[J]. IBM Journal of Research and Development, 1961, 5(3): 183-191.

[5] BÉRUT A, ARAKELYAN A, PETROSYAN A, et al. Experimental verification of Landauer's principle linking information and thermodynamics[J]. Nature, 2012, 483(7388): 187-189.

[6] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science & Technology, 2007, 58(7):1019-1031.

[7] ALON U. Network motifs: theory and experimental approaches[J]. Nature Reviews Genetics, 2007, 8(6): 450-461.

[8] LYU L, PAN L, ZHOU T, et al. Toward link predictability of complex networks[J]. Proceedings of the National Academy of Sciences, 2015, 112(8): 2325-2330.

[9] WU Y T, YU H T, HUANG R Y, et al. A fusion link prediction method based on limit theorem[J]. Applied Sciences, 2017, 8(1):32.

[10] FRANÇOIS L, WHITE H C. Structural equivalence of individuals in social networks[J]. Social Networks, 1977, 1(1):67-98.

[11] BARABASI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439):509-512.

[12] ZHOU T, LYU L, ZHANG Y C. Predicting missing links via local information[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2009, 71(4): 623-630.

[13] MA C, BAO Z K, ZHANG H F. Improving link prediction in complex networks by adaptively exploiting multiple structural features of networks[J]. Physics Letters A, 2017, 381(39): 3369-3376.

[14] YAO Y, ZHANG R, YANG F, et al. Link prediction based on local weighted paths for complex networks[J]. International Journal of Modern Physics C, 2017, 28(4): 1-23.

[15] LIU S X, JI X S, LIU C X, et al. Similarity indices based on link weight assignment for link prediction of unweighted complex networks[J]. International Journal of Modern Physics B, 2016:1650254.

[16] 刘树新, 季新生, 刘彩霞, 等. 局部拓扑信息耦合促进网络演化[J]. 电子与信息学报, 2016, 38(9):2180-2187.
LIU S X, JI X S, LIU C X, et al. Information coupling of local topology promoting the network evolution[J]. Journal of Electronics & Information Technology. 2016, 38(9):2180-2187.

[17] KUMAR A, SINGH S S, SINGH K, et al. Level-2 node clustering

coefficient-based link prediction[J]. Applied Intelligence, 2019(1): 1-18.

[18] BISWAS A, BISWAS B. Community-based link prediction[J]. Multimedia Tools and Applications, 2017, 76(18):18619-18639.

[19] DE B C, POWER E A, LARREMORE D B, et al. Community detection, link prediction, and layer interdependence in multilayer networks[J]. Physical Review E, 2017, 95(4-1):042317.

[20] MAN G, LING C, BIN L, et al. A link prediction algorithm based on low-rank matrix completion[J]. Applied Intelligence, 2018, 48: 4531-4550.

[21] 韦布, 科普西. 统计模式识别: 第三版[M]. 王萍, 译. 北京:电子工业出版社, 2015.
WEBB A R, COPSEY K D. Statistical pattern recognition[M]. 3rd ed. WANG P, transl. Beijing: Publishing House of Electronics Industry, 2015.

[22] HOLLAND P W, LASKEY K B, LEINHARDT S. Stochastic blockmodels: first steps[J]. Social Networks, 1983, 5(2): 109-137.

[23] GUIMERÀ R, SALES-PARDO M. Missing and spurious interactions and the reconstruction of complex networks[J]. Proceedings of the National Academy of Sciences, 2009, 106(52): 22073-22078.

[24] CLAUSET A, MOORE C, NEWMAN M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7191):98-101.

[25] HE Y, LIU J N K, HU Y, et al. OWA operator based link prediction ensemble for social network[J]. Expert Systems with Applications, 2015, 42(1): 21-50.

[26] YU H T, WANG S H, MA Q Q. Link prediction algorithm based on the Choquet fuzzy integral[J]. Intelligent Data Analysis, 2016, 20(4): 809-824.

[27] 刘冶, 朱蔚恒, 潘炎, 等. 基于低秩和稀疏矩阵分解的多源融合链接预测算法[J]. 计算机研究与发展, 2015, 52(2):423-436.
LIU Y, ZHU W H, PAN Y, et al. Multiple sources fusion for link prediction via low-rank and sparse matrix decomposition[J]. Journal of Computer Research and Development. 2015, 52(2):423-436.

[28] 吴祖峰, 梁棋, 刘峤, 等. 基于 AdaBoost 的链路预测优化算法[J]. 通信学报, 2014(3):116-123.
WU Z F, LIANG Q, LIU Q, et al. Modified link prediction algorithm based on AdaBoost[J]. Journal on Communications. 2014(3):116-123.

[29] WANG P, XU B W, WU Y R, et al. Link prediction in social networks: the state-of-the-art[J]. Science China Information Sciences, 2015, 58(1):1-38.

[30] SARUKKAI R R. Link prediction and path analysis using Markov chains[J]. Computer Networks, 2000, 33(1):377-386.

[31] 詹姆斯. 统计学习导论[M]. 王星, 译. 北京: 机械工业出版社, 2015.
JAMES G. An introduction to statistical learning[M]. WANG X, transl. Beijing: China Machine Press, 2015.

[32] WANG Y, YAO Y, TONG H, et al. A brief review of network embedding[J]. Big Data Mining and Analytics, 2019, 2(1):35-47.

[33] ZHOU J, CUI G, ZHANG Z, et al. Graph neural networks: a review of methods and applications [J]. arXiv Preprint, arXiv:1812.08434, 2018.

[34] WU Z, PAN S, CHEN F, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020:1-21.

[35] LIAO L, HE X, ZHANG H, et al. Attributed social network embedding[J]. IEEE Transactions on Knowledge & Data Engineering, 2018, 30: 2257-2270.

[36] CHEN T, SUN Y. Task-guided and path-augmented heterogeneous network embedding for author identification[C]//10th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2017: 295-304.

[37] WANG Z, CHEN C, LI W. Predictive network representation learning for link prediction[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2017: 969-972.

[38] BRADLEY A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern Recognition, 1997, 30(7): 1145-1159.

[39] HANLEY J A, MCNEIL B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology, 1982, 143(1): 29-36.

[40] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861-874.

[41] LYU L, ZHOU T. Link prediction in complex networks: a survey[J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(6): 1150-1170.

[42] 豪格, 克莱格. 数理统计导论[M]. 王忠玉, 卜长江, 译. 北京: 机械工业出版社, 2015.
HOGG R V, CRAIG A T. Introduction to mathematical statistics [M]. WANG Z Y, BU C J, transl. Beijing: China Machine Press, 2015.

[43] NEYMAN J, PEARSON E S. The testing of statistical hypotheses in relation to probabilities a priori[C]//Mathematical Proceedings of the Cambridge Philosophical Society. Cambridge: Cambridge University Press, 1933.

[44] MAYO D G, SPANOS A. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction[J]. The British Journal for the Philosophy of Science, 2006, 57(2): 323-357.

[作者简介]

吴翼腾（1992– ），男，山东乐陵人，信息工程大学博士生，主要研究方向为信息内容安全、对抗机器学习。

于洪涛（1970– ），男，辽宁丹东人，博士，信息工程大学研究员、博士生导师，主要研究方向为大数据与人工智能。

黄瑞阳（1986– ），男，福建漳州人，博士，信息工程大学副研究员，主要研究方向为知识图谱与文本挖掘。

李华巍（1983– ），男，吉林省吉林市人，92538 部队工程师，主要研究方向为大数据。