Computer Science and Application 计算机科学与应用, 2018, 8(8), 1239-1258
Published Online August 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.88135

Hans汉斯

Research on Social Network Link Prediction Algorithm Based on Multidimensional Similarity Attributes

Weijie Yang, Hecan Zhang, Lang Wu

School of Computer and Information Engineering, Beijing Technology and Business University, Beijing
Email: yangwj@th.btbu.edu.cn

Received: Aug. 5th, 2018; accepted: Aug. 20th, 2018; published: Aug. 28th, 2018

Abstract

Link prediction refers to searching for hidden links or predicting possible future links in social networks. It is important for analyzing social networks. This paper analyzes the current methods for social network link prediction, compares multidimensional similarity attributes, takes the link prediction problem as a classification problem, and realizes links prediction based on machine learning. The final experiment results verify that similarity attributes are effective for link prediction problem, and link prediction problem can be solved as a classification problem by machine learning algorithms.

Keywords

Link Prediction, Machine Learning, Similarity Attributes, Social Networks

基于多维相似度属性的社会网络链接预测算法研究

杨伟杰，张何灿，吴　朗

北京工商大学，计算机与信息工程学院，北京
Email: yangwj@th.btbu.edu.cn

收稿日期：2018年8月5日；录用日期：2018年8月20日；发布日期：2018年8月28日

摘　要

链接预测是寻找社会网络中隐藏的和未来可能出现的链接，它对于分析社会网络具有重要意义。本文在

文章引用: 杨伟杰, 张何灿, 吴朗. 基于多维相似度属性的社会网络链接预测算法研究[J]. 计算机科学与应用, 2018, 8(8): 1239-1258. DOI: 10.12677/csa.2018.88135

对现有社会网络链接预测研究的基础上，分析了社会网络链接预测算法中的多维相似度属性，并把链接预测问题转换为分类问题，尝试使用机器学习的方法解决社会网络链接预测问题，最终通过实验得到验证，相似度属性特征对链接预测具有较高影响力，链接预测问题可以转化为分类问题通过机器学习算法得到解决。

关键词

链接预测，机器学习，相似度属性，社会网络

1. 引言

复杂网络研究是众多科学领域的一个重要分支，大量的学者致力于研究网络的特征、演化过程、拓扑结构与功能之间的关系。其中，社会网络分析近年来成为复杂网络分析中一个新的非常重要的研究方向，尤其随着多种社交媒体的产生与发展繁盛，社会网络分析对于研究事件演化、信息推荐、舆情管控等都具有重要的作用。社会网络中的链接预测主要是通过网络的已知信息，预测没有链接的两个节点发生链接的可能性，这种链接关系可能是已经存在但尚未被发现的，也可能是目前不存在，但是在不久的将来非常可能发生的。社会网络实体之间的关系分析对事件发展具有极其重要的作用，因而链接预测在现实世界中具有非常广泛的应用性，比如社交网络中的好友推荐，电商中的商品推荐，社会安全中的实体识别等等。

链接预测作为社会网络分析的一个重要研究方向，是一个交叉学科问题，涉及社会学、系统学、图形学等等，逐渐发展成为国内外学者的研究热点。经典方法都是将社会网络看作是一个节点和关系的集合，网络中的每个节点对应一个实体，每条边对应着用户之间的一种关系，链接预测就是基于这些实体和关系的特征进行。为了实现快速的关系预测，如何综合社会网络中的多维特征是研究者们一直在探索的问题。本文提出的基于多维相似度属性的社会网络链接预测算法，就是要分析社会网络中节点和边的多维相似度属性，通过机器学习的方法，实现对社会网络有效的链接预测。

2. 相关工作

对社会网络链接预测的现有研究方法主要集中在基于概率模型、基于监督学习和基于相似度三类。

基于相似性的算法是链接预测中最直接有效的方法，但是也具有很多的挑战性。比如节点相似性的定义本身就存在异议，相似性指数的分类更是复杂。相似性通常是通过两个节点之间的共同特征来衡量，共同特征越多，相似性越高，则越有可能存在链接。然而通常情况下，能够直接得到的是网络拓扑结构，节点的属性是隐藏的，纯粹基于拓扑结构的相似性指数往往是比较浅显的，因而基于结构相似性的链接预测准确率高低，往往与结构相似性所提出的指标以及目标网络所具有的拓扑特性的匹配程度相关。Lv等人曾在文[1]中归纳了基于网络结构相似性指标的链接预测方法，并将这些指标归类为局部信息、路径、随机游走三大类，是对链接预测相似性指标比较系统全面的分类。现在对基于相似性的链接预测方法的分类有很多种方法，大致是基于两大类特征，即节点属性和路径信息。基于节点主要是基于公共邻居的相似度指标，比如 CN、Jaccard、Salton 等等[2]。基于路径的方法则是主要考虑最短路径、随机游走[3]

或者 page rank 这三类。

基于监督学习的链接预测主要是根据特征值对节点进行分类，这样链接预测就是一个典型的二分分类问题，但是这种方法的难点在于特征值的选取，通常做法是以图的拓扑结构来寻找特征值。例如 Kashima、Liben 等人的方法[4] [5]。但是如果社交网络规模较大时，特征值计算的复杂度往往也很高。后来，Doppa 等人研究发现，基于节点相似性，考虑拓扑结构特征，可以很好的提高链接预测的准确度[6]。

虽然很多研究表明，借助于节点属性的链接预测具有比较好的效果，但是这些属性往往是被看作独立的，于是有部分研究者考虑把节点属性进行有效结合，看是否对链接预测的准确性会有提高，这就产生了基于概率模型的方法，这类方法是系统的将节点和边属性进行结合，构造出一种联合概率分布，得到结构化的数据关系。Lise 和 Taskar 等人的研究证明，此种方法精确度较高[7] [8]。

目前，对链接预测的实现都是建立在对已知数据的分析之上，各种预测方法的目标都是相同的，但是分析问题的角度不同[9]。基于结构相似性的方法，主要考虑某个方向上的特征，如果所预测网络的特征不明显，预测结果也将不准确。但是这种方法的计算复杂度较低，具有比较好的实用基础。而基于概率模型的方法则考虑了结构信息和节点之间信息，以求达到较为完美的预测效果，这种方法的信息获取较为困难，算法复杂度也较高，在实际应用上有难度。因而本文中提出一种基于多维相似度特征的社会网络链接预测算法，以求选取能够有效融合的多种节点和路径相似性特征，采用机器学习算法，实现链接预测。

3. 相似度特征选择

基于相似性的链接预测，是根据所观察到的网络拓扑信息，构建相似性指标来进行链接预测，是目前链接预测领域的主流方法。为了选择合适的特征，避免特征越多效果越差的现象，本文从基于节点和路径的两大类属性中选择相似度特征进行链接预测。

3.1. 基于节点属性的特征选择

1) 公共邻居

$$S_{xy}^{\mathrm{CN}} = \left|\Gamma(x)\cap\Gamma(y)\right| \tag{1}$$

计算出两个节点 x，y 的公共邻居节点，并以此计算 x 和 y 节点的相似度。通常我们认为两个节点的邻居集合重复度很高时，即使两者之间没有直接连接的边，也可以认为 x 和 y 很可能存在某种关系，它们建立关系的可能性很大。

2) Salton 指标

$$S_{xy}^{\mathrm{Salton}} = \frac{\left|\Gamma(x)\cap\Gamma(y)\right|}{\sqrt{k_x\times k_y}} \tag{2}$$

这个指标又被称为余弦相似度，是在 CN 指标基础上加入了两个节点度的信息。

3) Sorensen 指标

$$S_{xy}^{\mathrm{Sorensen}} = \frac{2\left|\Gamma(x)\cap\Gamma(y)\right|}{k_x + k_y} \tag{3}$$

这个指标是在 Salton 指标基础上进行的改造，常常被用作生态社区数据。

4) adamic_adar 指标

$$S_{xy}^{\mathrm{AA}} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log\left|\Gamma(z)\right|} \tag{4}$$

这个指标通过赋予少数链接的邻居更多权重，来改善公共邻居的作用。思想是每个邻居节点在相似度计算中的贡献是不同的，度小的公共邻居节点的作用比度大的节点更重要。

5) RA(resource_allocation)指标

$$S_{xy}^{\mathrm{RA}}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\left|\Gamma(z)\right|} \tag{5}$$

这个指标是基于网络资源动态分配思想产生，通过计算两个节点间能够传输的资源数量来度量他们之间的相似性。

6) HPI 指标

$$S_{xy}^{\mathrm{HPI}}=\frac{\left|\Gamma(x)\cap\Gamma(y)\right|}{\min\left\{k_x,k_y\right\}} \tag{6}$$

根据定义，该指标的分母是由较小的节点度决定的，因而度大的节点更容易与其他节点有较高相似性，也就是说与枢纽节点相连接的链接会被分配较高的相似度分数。

7) HDI 指标

$$S_{xy}^{\mathrm{HDI}}=\frac{\left|\Gamma(x)\cap\Gamma(y)\right|}{\max\left\{k_x,k_y\right\}} \tag{7}$$

这是类似于 HPI 的相反效果的枢纽指标。

8) LHN 指标

$$S_{xy}^{\mathrm{LHN1}}=\frac{\left|\Gamma(x)\cap\Gamma(y)\right|}{k_x\times k_y} \tag{8}$$

这个指标给有许多共同的邻居的节点对分配高相似性。但是相比较与 CN 系数，该系数不会无限制的变大。

9) PA (Preferential Attachment)指标

$$S_{xy}^{\mathrm{PA}}=k_x\times k_y \tag{9}$$

这个指标被称作优先连接机制，它忽略了网络结构信息对相似度的影响，仅仅考虑了节点的度。一个节点的度越大，也就是与其他节点产生链接多，那么这个节点将来与未连接节点产生链接的可能性越大。

10) Jaccard 指标

$$S_{xy}^{\mathrm{Jaccard}}=\frac{\left|\Gamma(x)\cap\Gamma(y)\right|}{\left|\Gamma(x)\cup\Gamma(y)\right|} \tag{10}$$

Jaccard 方法的主要思想是两个节点拥有的共同邻居占它们所有邻居节点的比例越高，那么他们未来发生联系的可能性越大。

3.2. 基于路径的特征选择

相对于基于节点的相似度特征选择，基于路径的方法需要考虑网络的整体拓扑信息，因而有两个致命的缺点：第一，计算一个全局相似度指标是非常费时的，并且在网络规模巨大时，这种计算方案是行不通的；第二，全局的拓扑信息有时是不可获取的，特别是当使用一个分散的方式来实施算法时。因此，

如何选择既容易计算，并且链接预测准确率又高的相似度指标就显得尤为重要。

1) 最短距离

基于路径的方法中最简单的指标。如果两个节点之间的最短路径越短(除去直接连接的边)，则它们越容易产生作用，越可能连接。

2) Katz 方法

Katz 是一种计算节点声望的方法。它给予短路径更高的权重，然后计算全部的加权路径和，定义为

$$S_{xy} = \sum_{i=1}^{\omega} \beta^i \cdot \left| paths_{x,y}^{(i)} \right| \tag{11}$$

其中，$paths_{x,y}^{(i)}$ 是图中节点 x 和 y 之间所有长度为 L 的路径的集合，$\beta\ (0 \le \beta \le 1)$为衰减系数。$\beta$越小，则该方法越接近公共邻居算法，原因是路径越长，对和的贡献越小。

3) 本地距离

定义为

$$S = A^2 + \varepsilon A^3 \tag{12}$$

其中，ε 为参数，A 为邻接矩阵，如果节点 x 和 y 直接相连，则 $A_{xy} = 1$，否则 $A_{xy} = 0$。

4. 实验结果与分析

4.1. 实验数据集

在本实验中，数据集分为两大类：仿真数据集和真实数据集。其中，仿真数据集主要模拟了 erdos_renyi 模型，BA 模型，随机生长模型，森林火灾模型。真实数据集则是来自斯坦福官网的 SNAP 数据集，包括了 wiki-Vote，ca-GrQc，ca-HepTh，p2p-Gnutella08 这 4 个数据集。原始数据 txt 下载后如图 1 所示。

由于真实的数据集数据量很大，为了使算法性能达到最佳，对数据集进行如下预处理：

1) 数据集全部作为无向图处理；

2) 将数据集的真实节点从 1 开始重新编号；

3) 将真实网络删去 10%的边(连接边的两个节点的度至少为 1，否则有的指标会产生除零异常)，边数记为 n，记下每条边对应的节点对，作为测试集 label 为 1 的样本。剩下的节点对全部都未连接，作为训练集 label 为 0 的样本。

4.2. 评价指标

1) 正确率：正确预测的样本数占整个样本数的比率。

2) 识别准确度

$$\text{accuracy} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \tag{13}$$

TP (True Positives)：表示正确肯定的分类数。

TN (True Negatives)：表示正确否定的分类数。

FP (False Positives)：表示错误肯定的分类数。

FN (False Negatives)：表示错误否定的分类数。

3) 识别精确率

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{14}$$

```
# Directed graph (each unordered pair of nodes is saved once): p2p-Gnutella08.txt
# Directed Gnutella P2P network from August 8 2002
# Nodes: 6301 Edges: 20777
# FromNodeId    ToNodeId
0       1
0       2
0       3
0       4
0       5
0       6
0       7
0       8
0       9
0       10
3       703
3       826
3       1097
3       1287
3       1591
3       1895
```

Figure 1. Original text of the SNAP dataset
图 1. SNAP 数据集原始文本示意图

4) 反馈率(召回率)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{TN}} \times 100\% \tag{15}$$

5) 真正类率(灵敏度)

$$\text{True Positive Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{16}$$

6) 假正类率(特异性)

$$\text{False Positive Rate} = \frac{\text{FP}}{\text{FP} + \text{TN}} \times 100\% \tag{17}$$

7) ROC 曲线

ROC 曲线，是一个图形化说明二进制分类器系统的性能的工具，它的判别阈值是变化的。纵轴为灵敏度，横轴为特异性。曲线的绘制真正类率(TPR)创建对假正类率(FPR)在不同的阈值设置。该曲线下的积分面积能衡量模型的优劣，ROC 下的积分面积值越接近 1 则该模型效果越好。以图 2 为例，最左侧曲线下面积明显最大，因此其分类效果最好，右侧粗线次之。如果 ROC 曲线下面积在 1~0.85，则认为分类器表现优秀，在 0.85~0.7，认为表现良好，0.7~0.5 认为有待改善，小于 0.5，那可以认为分类器效果比随机猜测还要差。

4.2. 验证算法

本文将链接预测问题转化为一个分类问题来求解，使用的分类算法包括：最近邻节点算法(KNN)、朴素贝叶斯(NaiveBayes)、多层感知器神经网络(MLP)、支持向量机(SVM)、决策树(DT)，依次来验证所选择特征的有效性，以及基于机器学习算法在此问题中的可行性。

4.3. 实验结果

4.3.1. 仿真数据实验结果

1) Erdos_Renyi 模型

在如表 1 描述的 5000 节点的 Erdos_Renyi 模型上进行实验，结果如表 2、图 3 所示，实验效果没有随着拓扑结构变得稠密而有很大改进，如图 4 所示，MLP，SVM，DT 算法表现良好，NB 表现不稳定，KNN 算法表现不佳，此模型最优算法为 DT，所有算法正确率均能达到半数以上。

Table 1. ER model generation network
表 1. ER 模型生成网络说明

名称	类型	节点数	边数
Erdos_Renyi	随机图	5000	8743

Table 2. Experimental results of ER model
表 2. ER 模型实验结果

方法	正确率	ROC 曲线面积	召回率	精确率
KNN	0.6462	0.5809	1.0	0.6460
NB	0.6439	0.7180	1.0	0.6035
MLP	0.5974	0.7020	0.5332	0.7664
SVM	0.5301	0.7329	0.3593	0.8114
DT	0.6788	0.7386	0.7048	0.7719

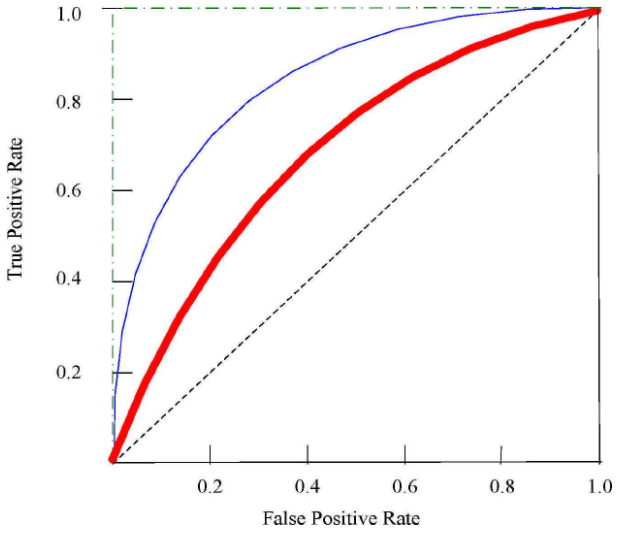


Figure 2. ROC curve example
图 2. ROC 曲线示例

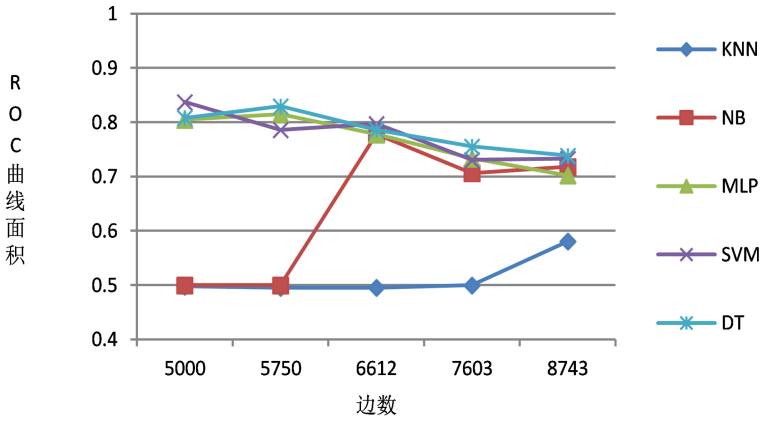
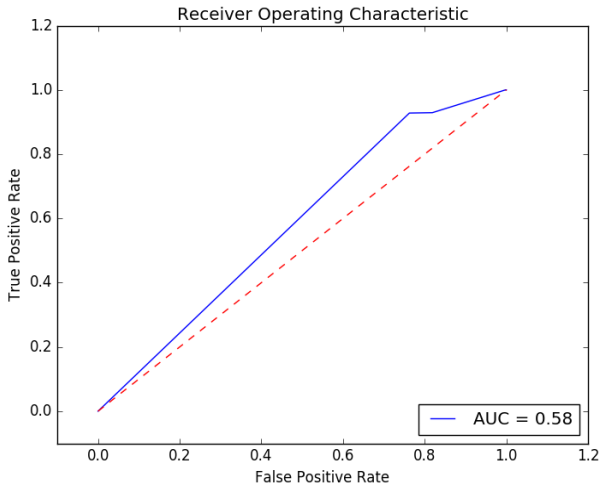
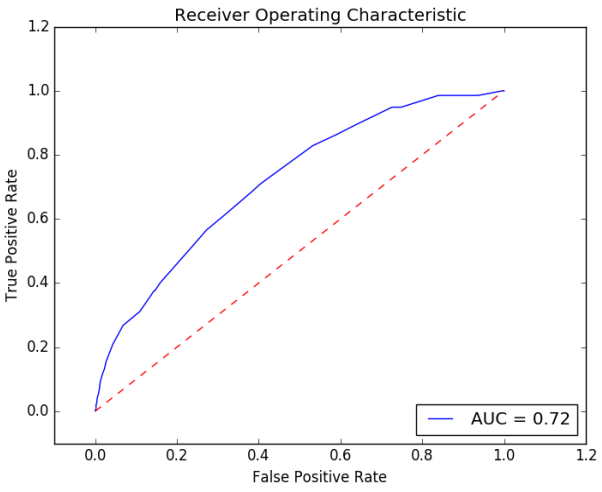


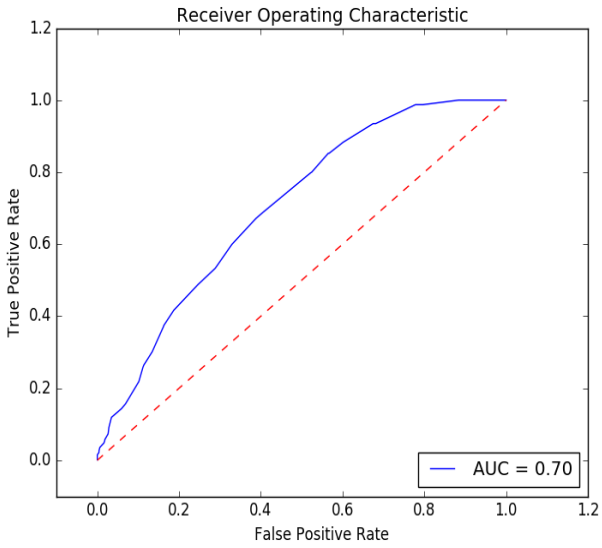
Figure 3. The relationship between the number of the connected edges and the recognition results
图 3. 连接边的数量与识别结果的关系



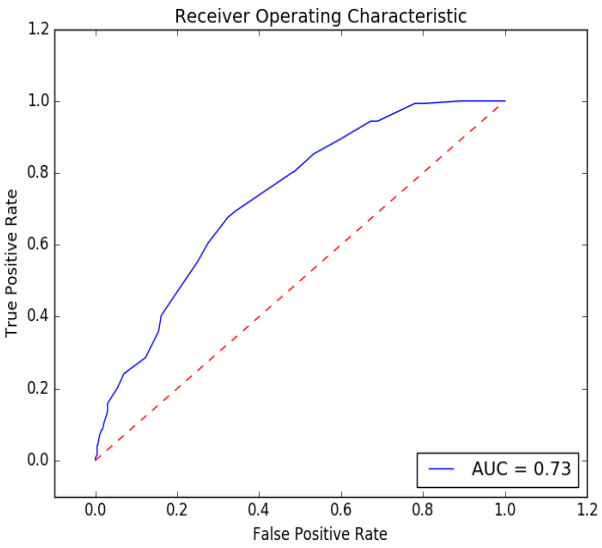
KNN



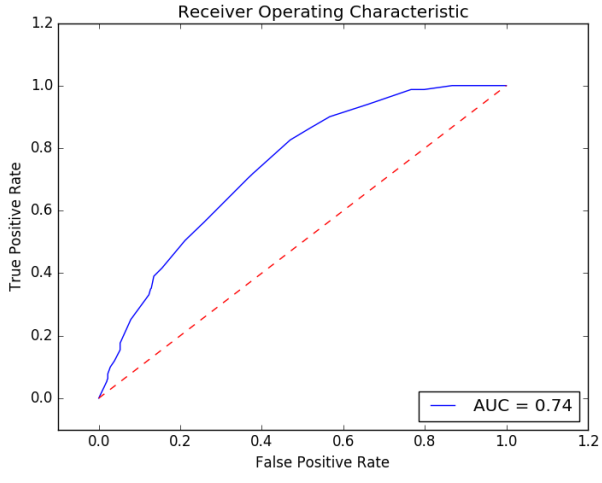
NB



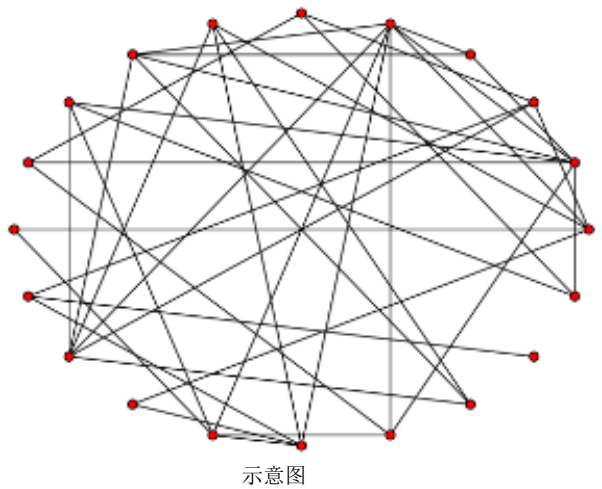
MLP



SVM



DT



示意图

Figure 4. ROC curve of ER model link prediction
图 4. ER 模型链接预测 ROC 曲线

2) BA 模型

BA 模型具有两个特性：1) 增长性，是指网络节点的数量不断增多；2) 优先连接机制，是指网络当中的新节点更倾向于和那些连接度较大的节点相连接，本文使用网络的说明如表 3 所示。

在如表 3 描述的 5000 节点的 BA 模型上进行实验，结果如表 4、图 5 所示，随着图越来越稠密，测试集正确率也随之有小幅提升，从图 6 中可以得出，SVM 和 MLP 都有优秀表现，MLP 表现要更好一些，KNN 在稀疏时表现一般，稠密时表现良好，DT 和 NB 则表现不稳定。

3) 随机生长模型

在每一个时间 t，从网络中随机添加节点，并形成一个新的完整的图，本文中使用的网络如表 5 中所述。

在 5000 初始节点的随机生长模型上，如表 6 和图 7 所示，随着图越来越稠密，测试集正确率也随之有小幅提升，SVM 表现最好。ROC 曲线面积，MLP 和 SVM 都表现优秀，难分伯仲，DT 表现良好，但是会随着更加稠密而略有下降，KNN 表现一般，NB 则表现不稳定，如图 8 所示。因此，该模型链接预测最好的方法是 MLP。

4) 森林生长模型

森林生长模型的图模型，在某一个时间，在图中添加新的顶点，本文中所使用网络如表 7 所述。

Table 3. BA model generation network
表 3. BA 模型生成网络说明

名称	类型	初始节点数	每个节点带来的边数
Barabasi-Albert	无标度网络	5000	3

Table 4. Experimental results of BA model
表 4. BA 模型实验结果

方法	正确率	ROC 曲线面积	召回率	精确率
KNN	0.7071	0.7132	0.4209	0.9844
NB	0.6191	0.9388	0.2395	0.9945
MLP	0.8600	0.9448	0.7645	0.9448
SVM	0.5967	0.9071	0.2001	0.9677
DT	0.4903	0.6989	0.8999	0.4947

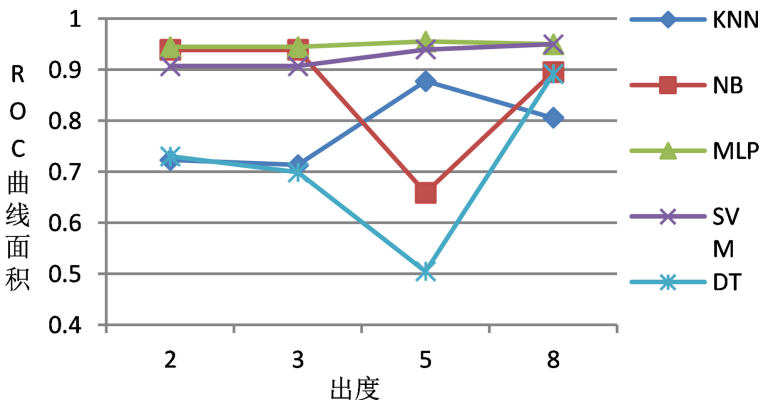
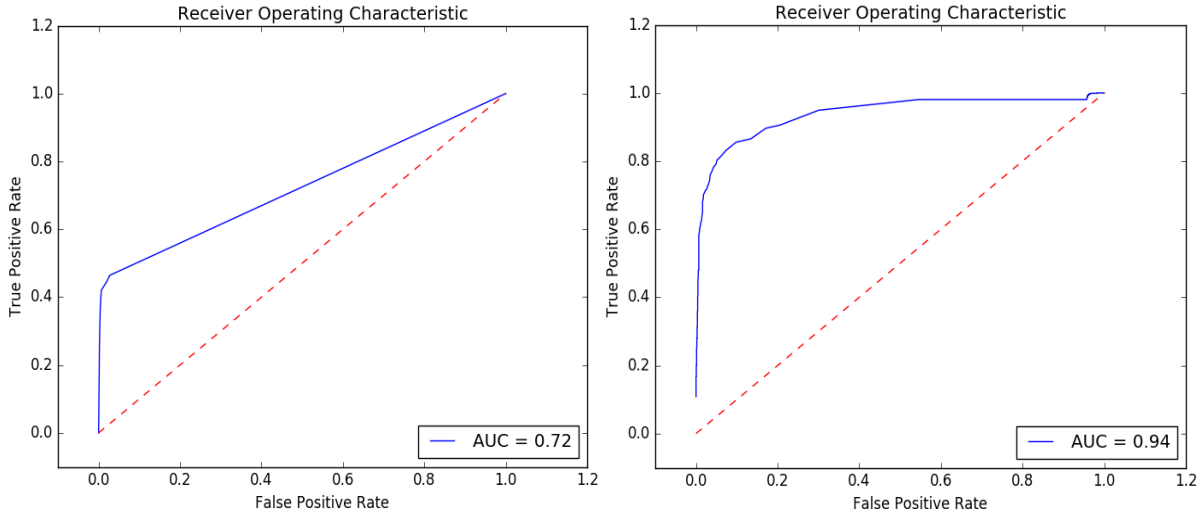
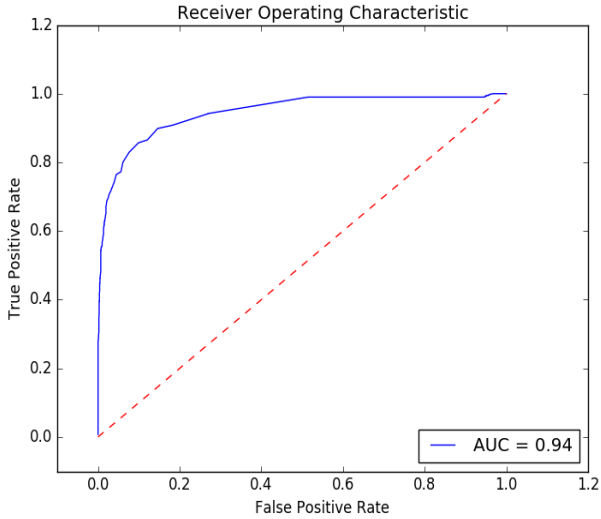


Figure 5. The relationship between the output and the recognition results
图 5. 出度与识别结果的关系

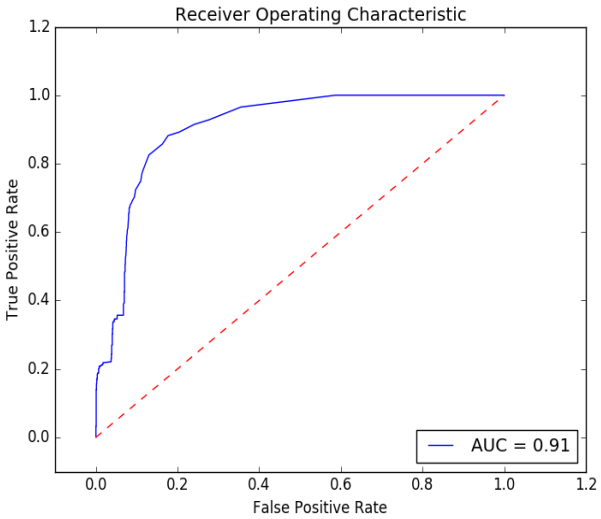


KNN

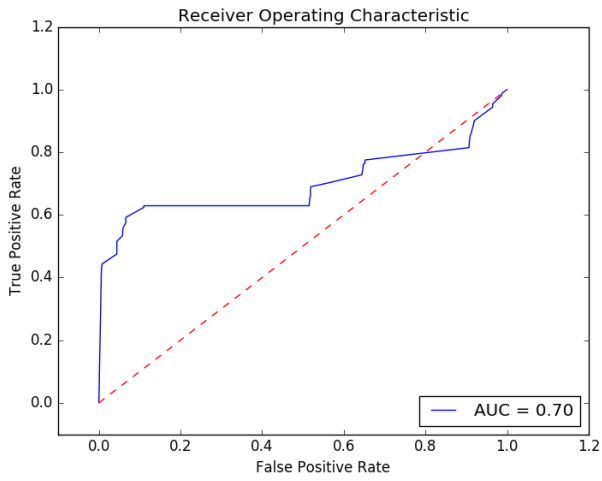
NB



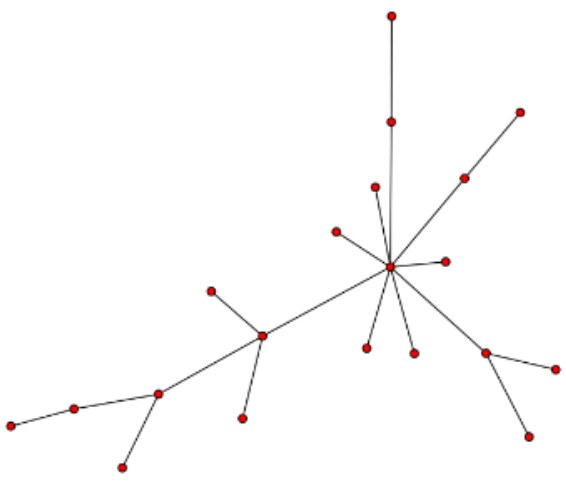
MLP



SVM



DT



示意图

Figure 6. ROC curve of BA model link prediction
图 6. BA 网络模型链接预测 ROC 曲线

如表 8 与图 9 所示，在 5000 初始节点的森林火灾模型中，随着图越来越稠密，测试集正确率整体变化不大，ROC 曲线面积，整体略有下降，其中 MLP 一直表现最好。各模型结果如图 10 中所示，MLP 和 SVM 都表现优秀，DT，KNN 表现有待提升，NB 表现不稳定。因此，该模型链接预测最好的方法是 MLP。

Table 5. Random growth model generation network
表 5. 随机生长模型生成图说明

名称	类型	初始节点数	每个节点带来的边数
Growing_Random	生长网络	5000	2

Table 6. Experimental results of random growth model
表 6. 随机生长模型实验结果

方法	正确率	ROC 曲线面积	召回率	精确率
KNN	0.7608	0.6424	0.9279	0.7957
NB	0.3595	0.5520	0.1842	0.9946
MLP	0.5724	0.9274	0.4605	0.9914
SVM	0.3576	0.9318	0.1722	0.9942
DT	0.6576	0.8695	0.5706	0.9777

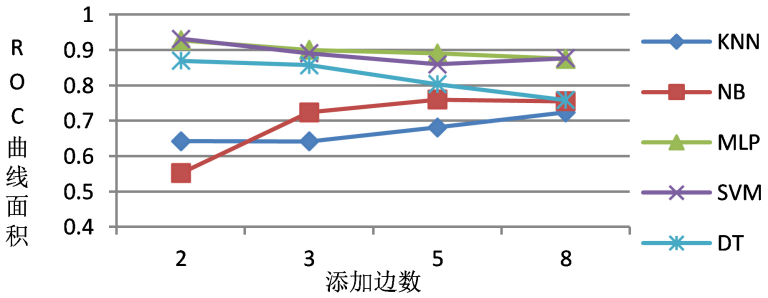
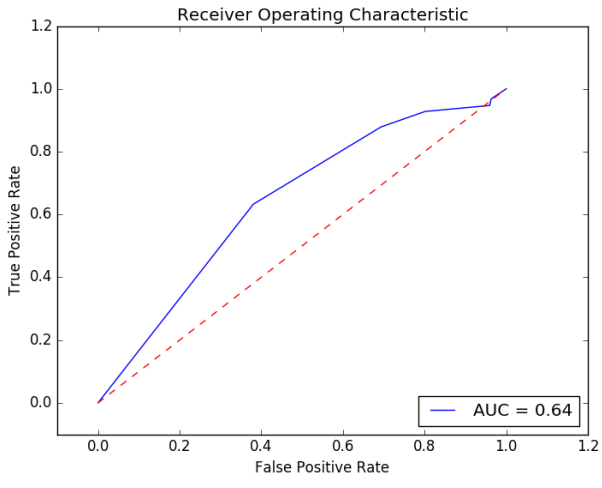
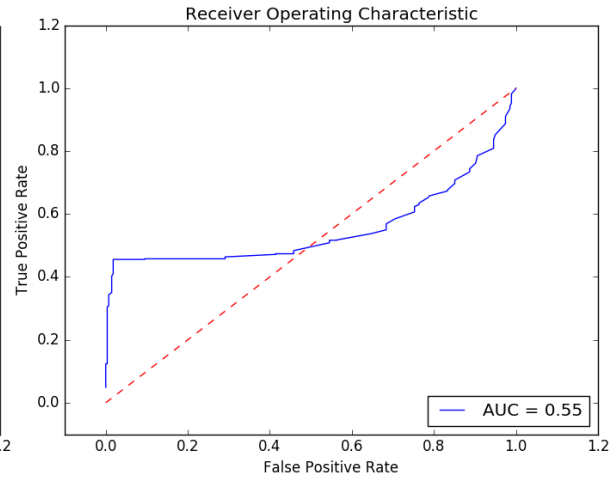


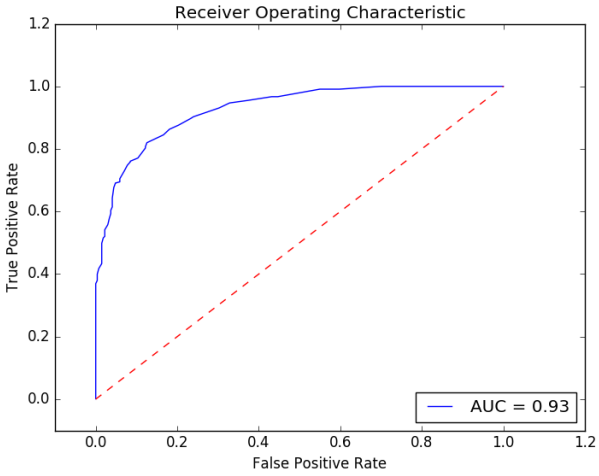
Figure 7. The relationship between the number of added edges and the recognition results
图 7. 添加边数与识别结果的关系



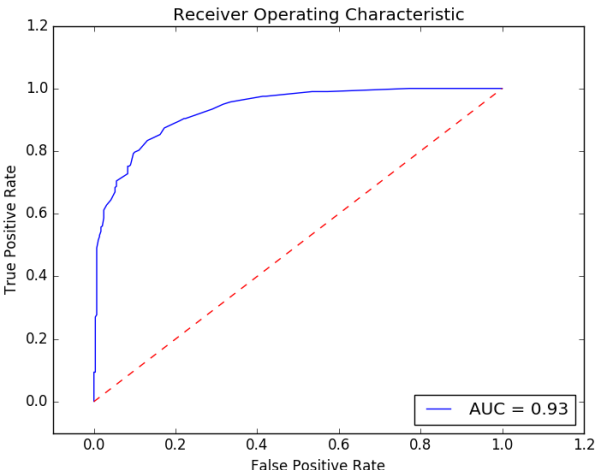
KNN



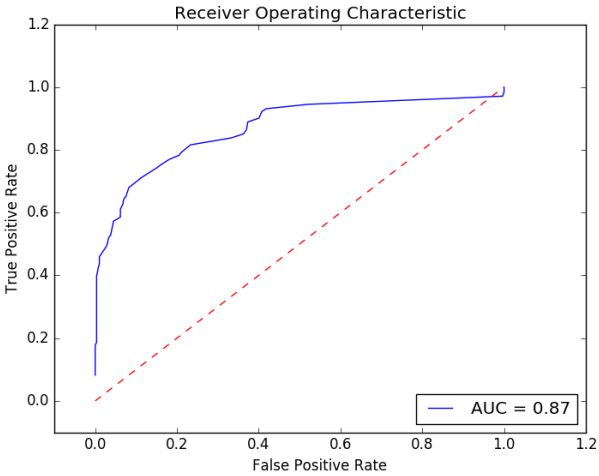
NB



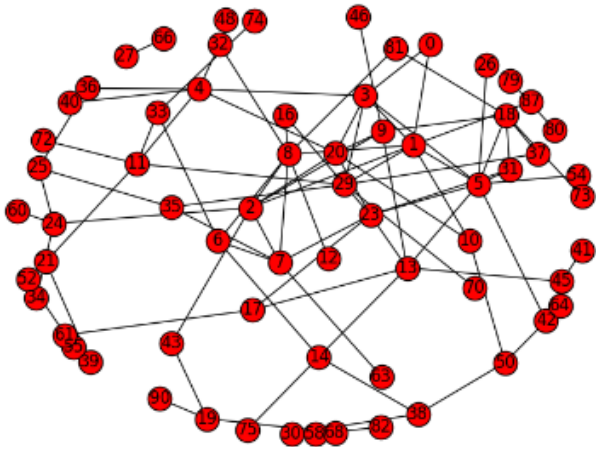
MLP



SVM



DT



示意图

Figure 8. ROC curve of random growth model link prediction
图 8. 随机生长网络模型链接预测 ROC 曲线

Table 7. Forest growth model generation network
表 7. 森林生长模型生成图说明

名称	类型	节点数	正向燃烧的概率	每一步选择的节点数
Forest_Fire	生长网络	5000	0.1	2

Table 8. Experimental results of Forest growth model
表 8. 森林生长模型实验结果

方法	正确率	ROC 曲线面积	召回率	精确率
KNN	0.7024	0.7876	0.6422	0.7301
NB	0.6482	0.7416	0.2982	0.9941
MLP	0.7218	0.8372	0.4735	0.9405
SVM	0.6326	0.8214	0.2680	0.9902
DT	0.7126	0.7910	0.5208	0.8451

4.3.2. SNAP 数据实验结果

在仿真模型取得不错的结果以后，针对斯坦福大学的 SNAP 实验数据，进行了以下实验。

1) CA-GrQc

在表 9 中描述的 CA-GrQc 数据集上进行测试，结果如表 10、图 11 所示，测试集正确率和 ROC 曲线下面积，五种方法表现都十分优秀，由于 MLP 的测试集正确率和 ROC 曲线下面积都为第二好，所以此数据集链接预测最优方法为 MLP。

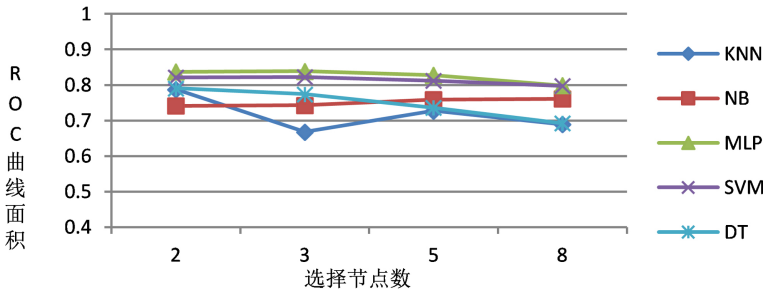
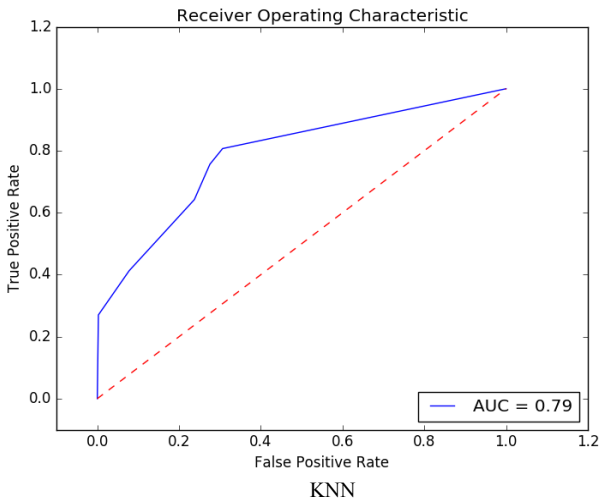
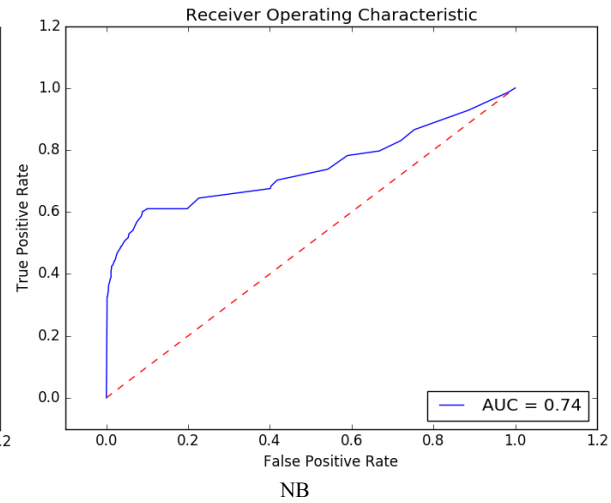


Figure 9. The relationship between the number of selected nodes and the recognition results

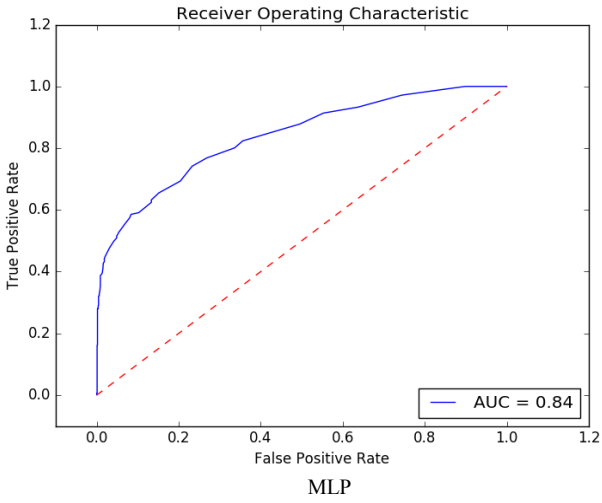
图 9. 选择节点数与识别结果的关系



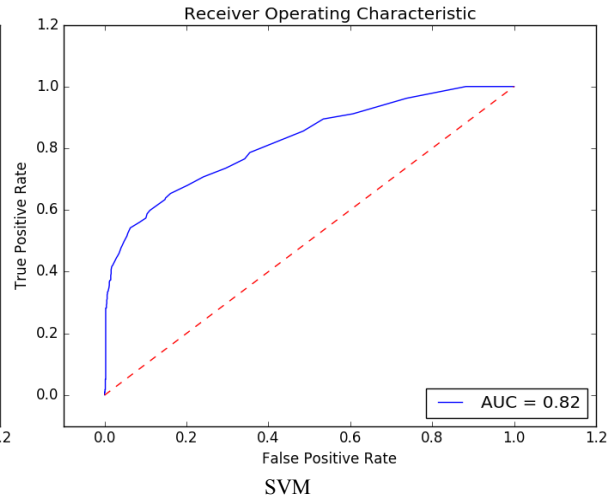
KNN



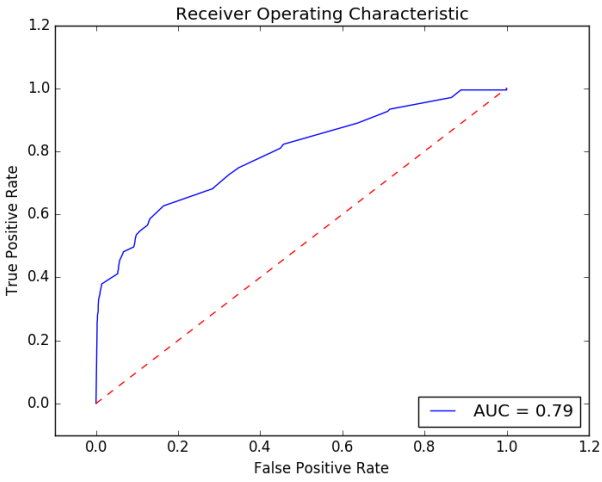
NB



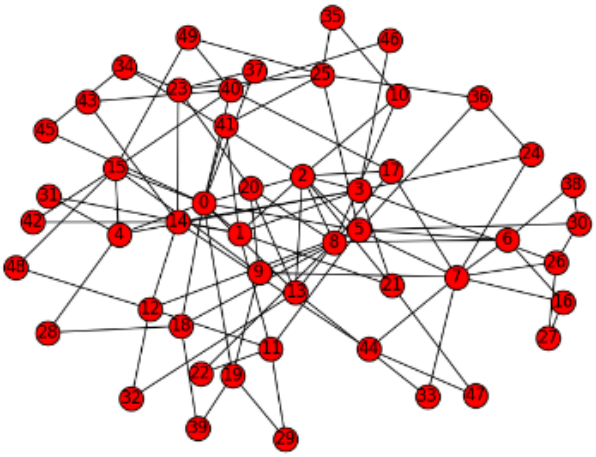
MLP



SVM



DT



示意图

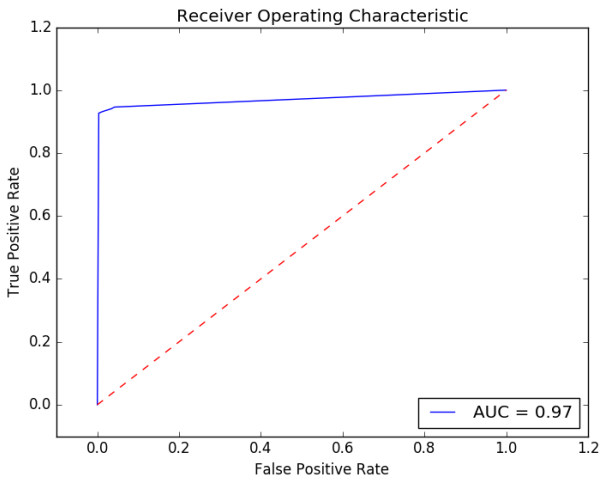
Figure 10. ROC curve of forest growth model link prediction
图 10. 森林生长模型链接预测 ROC 曲线

Table 9. CA-GrQc network
表 9. CA-GrQc 网络图说明

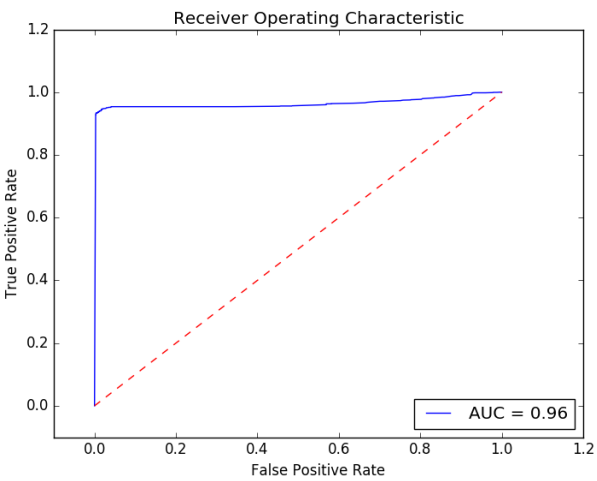
名称	节点数	边数	说明
CA-GrQc	5,242	14,496	Arxiv 普通相关性合作网络

Table 10. Experimental results of CA-GrQc network
表 10. CA-GrQc 网络实验结果

方法	正确率	ROC 曲线面积	召回率	精确率
KNN	0.9593	0.9698	0.9323	0.9854
NB	0.9641	0.9634	0.9316	0.9963
MLP	0.9585	0.9872	0.9206	0.9963
SVM	0.9185	0.9887	0.8391	0.9975
DT	0.9517	0.9467	0.9102	0.9925



KNN



NB

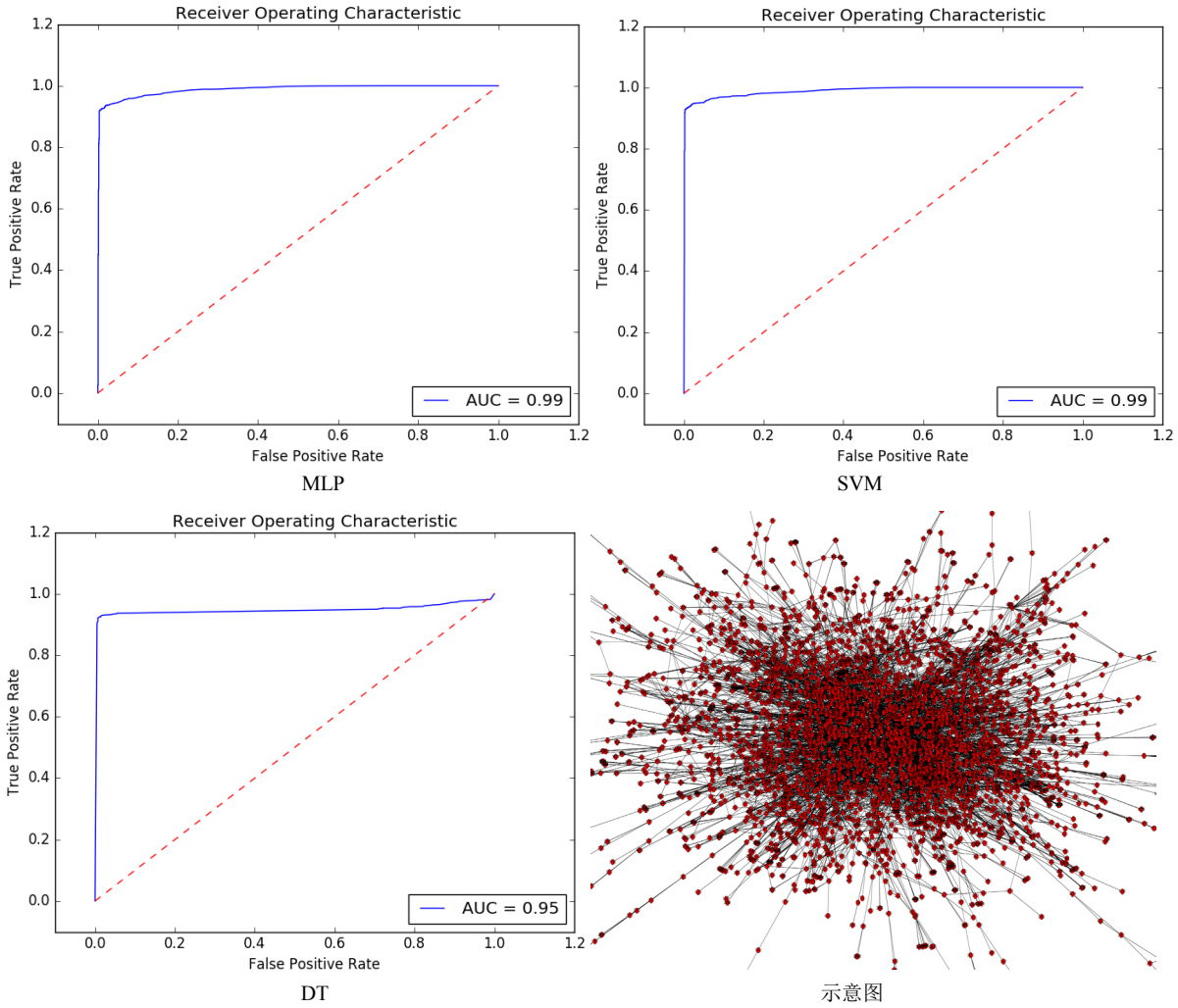


Figure 11. ROC curve of CA-GrQc network link prediction
图 11. CA-GrQc 数据集链接预测 ROC 曲线

2) CA-HepTh

在表 11 中描述的 CA-HepTh 数据集上进行测试，结果表 12，图 12 所示，测试集正确率 KNN 相对最低，为 0.86，其他四种方法表现都在 0.9 以上，五种方法都十分优秀。ROC 曲线下面积，五种方法表现也都十分优秀。由于 MLP 的测试集正确率和 ROC 曲线下面积都为最好，所以此数据集链接预测最优方法为 MLP。

3) Wiki-Vote

在表 13 描述的 Wiki-Vote 数据集上进行测试，结果如表 14、图 13 所示，发现测试集正确率 NB 相对最低，为 0.84，其他四种方法表现都在 0.9 左右，五种方法都十分优秀。ROC 曲线下面积，五种方法表现也都十分优秀。由于 MLP 的测试集正确率和 ROC 曲线下面积都为最好，所以此数据集链接预测最优方法为 MLP。

4) p2p-Gnutella08

在表 15 描述的 p2p-Gnutella08 数据集上进行测试，结果如表 16、图 14 所示，测试集正确率相比前三种有所下降，其中 MLP 和 DT 在 0.75 以上，表现良好，其余三种在 0.7 以内，有待提高。ROC 曲线

Table 11. CA-HepTh network
表 11. CA-HepTh 网络图说明

名称	节点数	边数	说明
CA-HepTh	9,877	25,998	Arxiv 高能物理合作网络

Table 12. Experimental results of CA-HepTh network
表 12. CA-HepTh 实验结果

方法	正确率	ROC 曲线面积	召回率	精确率
KNN	0.8693	0.9472	0.9608	0.8435
NB	0.9380	0.9488	0.8799	0.9956
MLP	0.9455	0.9860	0.8964	0.9940
SVM	0.9303	0.9805	0.8637	0.9964
DT	0.9438	0.9600	0.8971	0.9894

Table 13. Wiki-Vote network
表 13. Wiki-Vote 网络图说明

名称	节点数	边数	说明
Wiki-Vote	7,115	103,689	维基百科一对一投票网络

Table 14. Experimental results of Wiki-Vote network
表 14. Wiki-Vote 实验结果

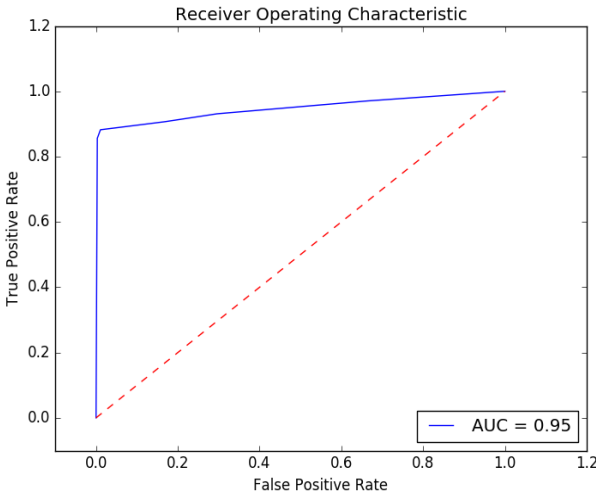
方法	正确率	ROC 曲线面积	召回率	精确率
KNN	0.9195	0.9581	0.9058	0.9313
NB	0.8477	0.9767	0.7085	0.9818
MLP	0.9125	0.9839	0.8526	0.9686
SVM	0.8974	0.9785	0.8398	0.9492
DT	0.8999	0.8827	0.8396	0.9546

Table 15. P2p-Gnutella08 network
表 15. p2p-Gnutella08 网络图说明

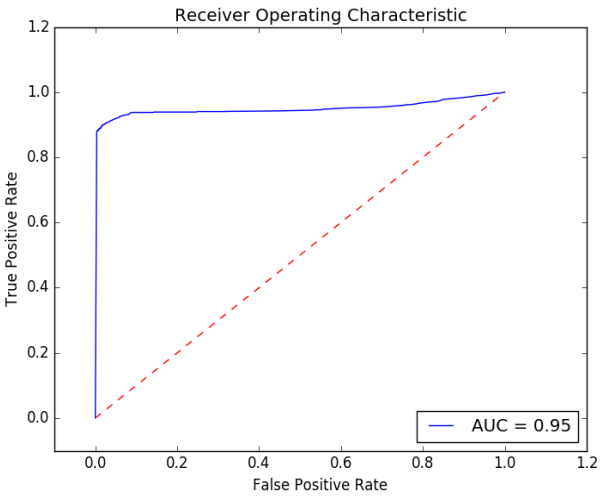
名称	节点数	边数	说明
p2p-Gnutella08	6,301	20,777	2002 年 8 月 8 日的 Gnutella 对等网络

Table 16. Experimental results of Wiki-Vote network
表 16. Wiki-Vote 实验结果

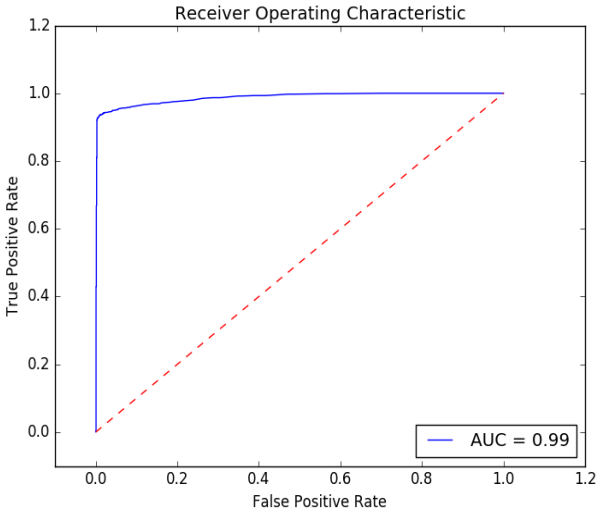
方法	正确率	ROC 曲线面积	召回率	精确率
KNN	0.6978	0.7625	0.8874	0.6434
NB	0.6295	0.7352	0.2666	0.9719
MLP	0.7647	0.9116	0.6367	0.8558
SVM	0.6076	0.9070	0.2262	0.9533
DT	0.7769	0.8455	0.7141	0.8167



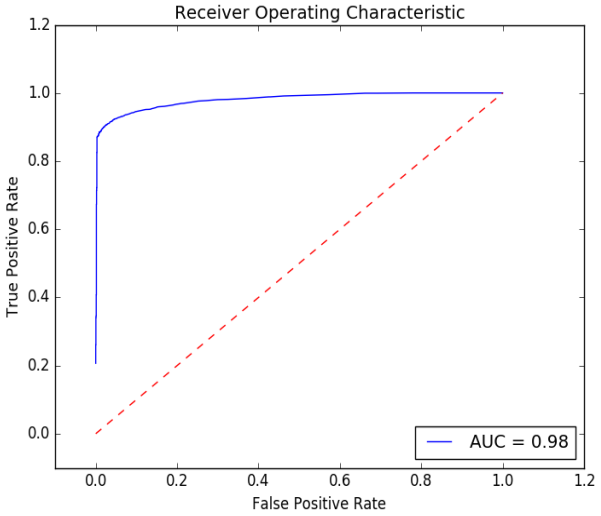
KNN



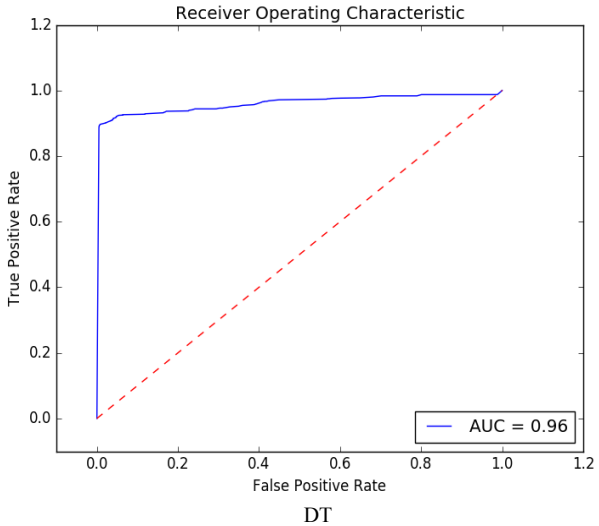
NB



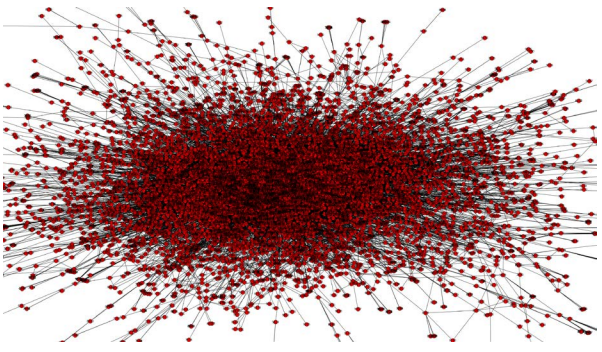
MLP



SVM

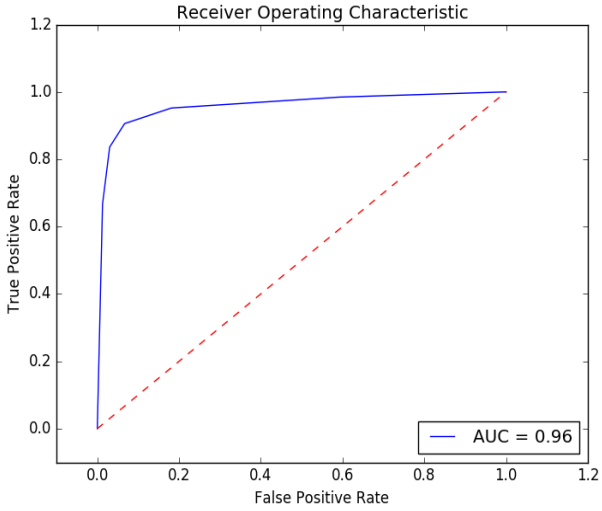


DT

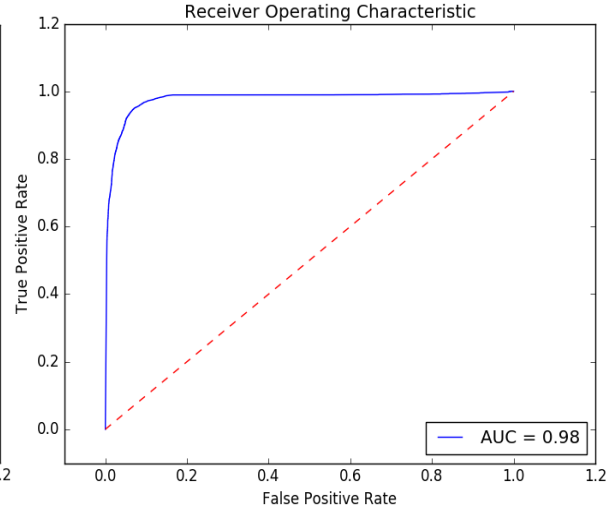


示意图

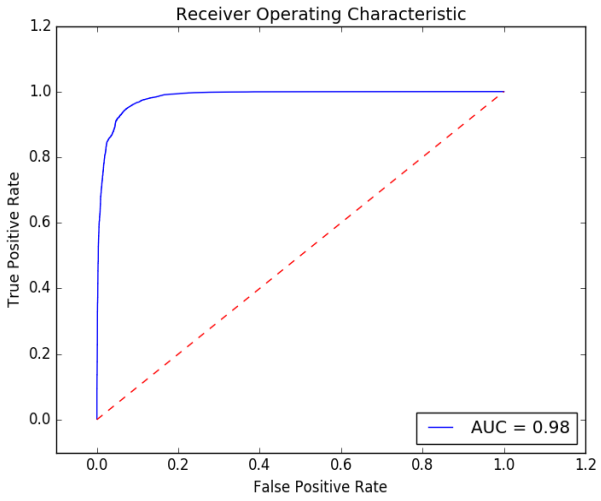
Figure 12. ROC curve of CA-HepTh network link prediction
图 12. CA-HepTh 数据集链接预测 ROC 曲线



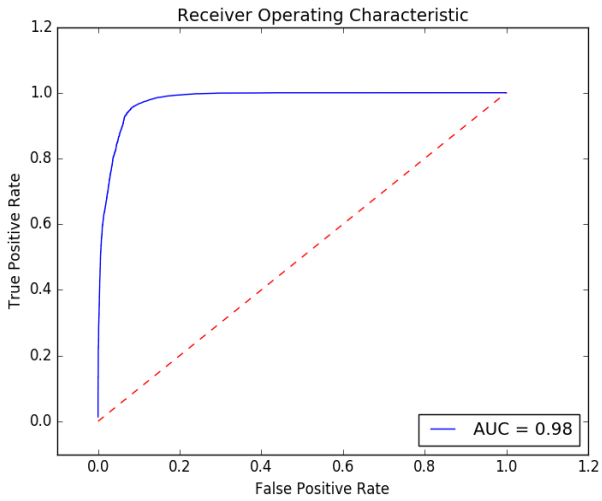
KNN



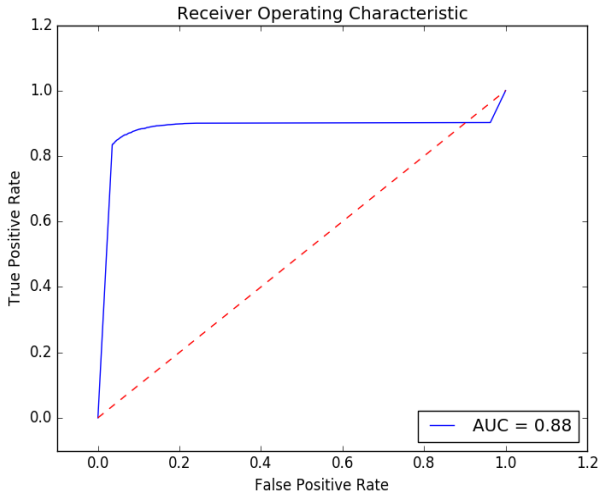
NB



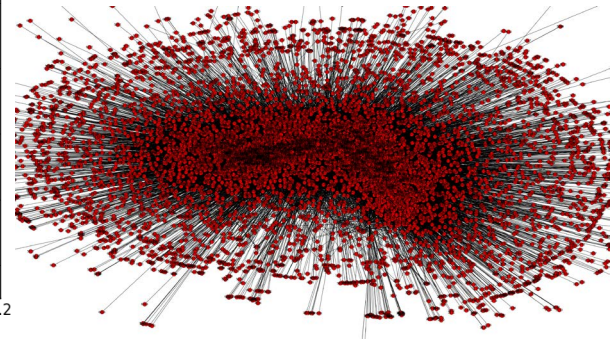
MLP



SVM

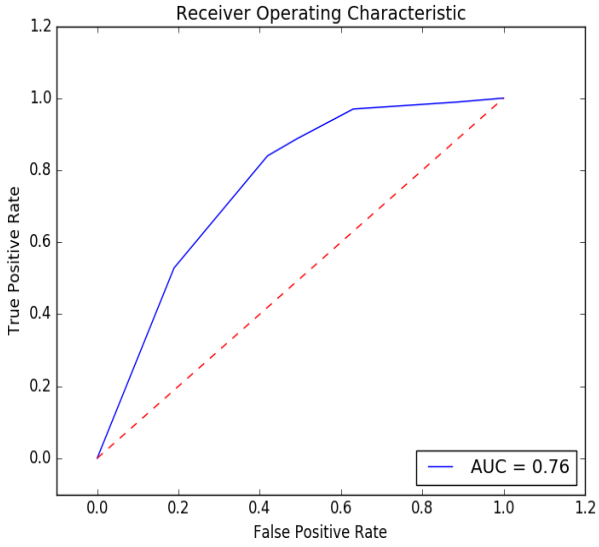


DT

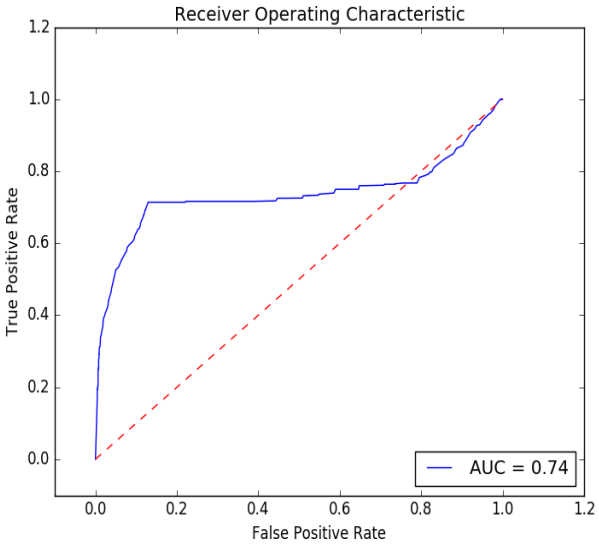


示意图

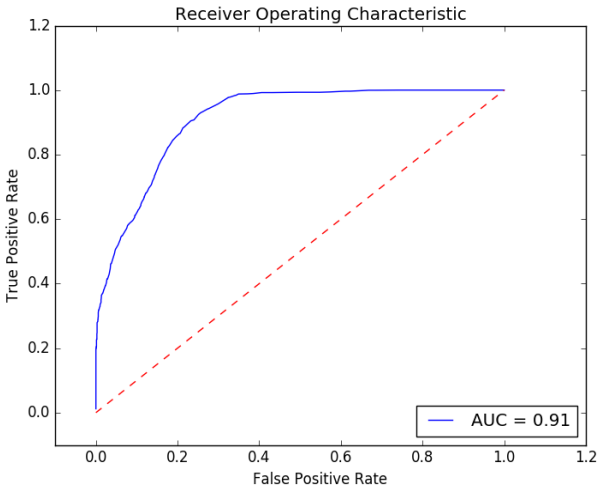
Figure 13. ROC curve of Wiki-Vote network link prediction
图 13. Wiki-Vote 数据集链接预测 ROC 曲线



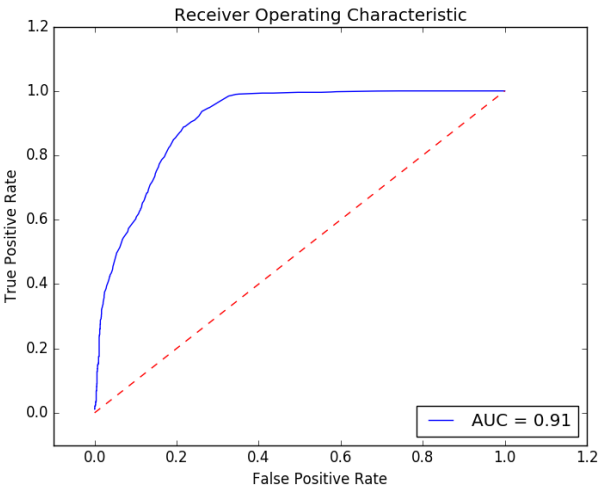
KNN



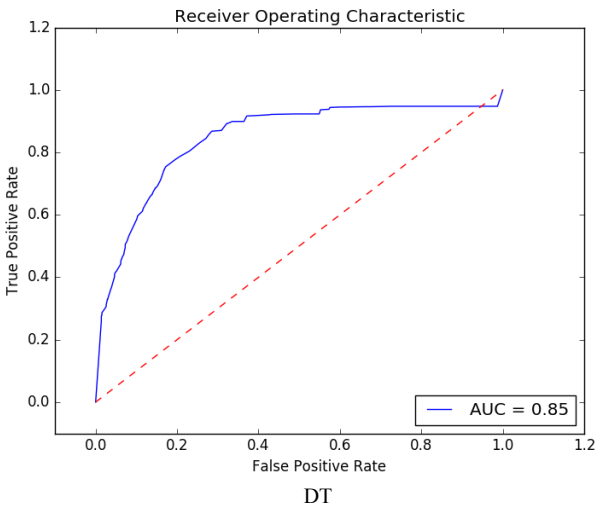
NB



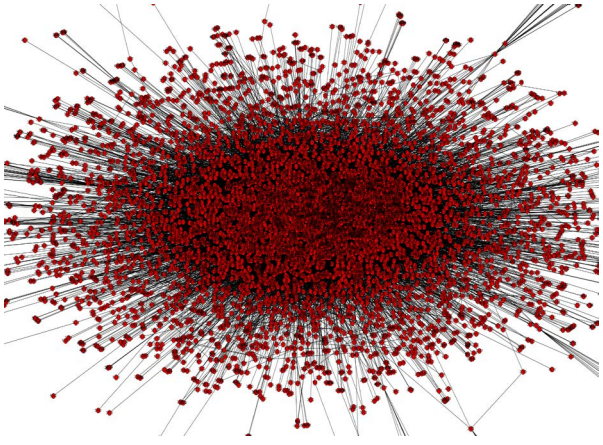
MLP



SVM



DT



示意图

Figure 14. ROC curve of CA-HepTh network link prediction
图 14. CA-HepTh 数据集链接预测 ROC 曲线

下面积，KNN 和 NB 在 0.75 左右，表现良好，DT 表现优秀，MLP 和 SVM 都在 0.91，表现优秀。由于 MLP 的测试集正确率和 ROC 曲线下面积都为最好，所以此数据集链接预测最优方法为 MLP。

5. 结束语

本文首先概括分析了现有社会网络链接预测方法，尝试挖掘多维相似性特征，并使用机器学习的方法进行链接预测。通过大量实验验证了此方法的可行性。基于多维相似度属性的方法是现在常用的链接预测方法，但是相似度属性的选择是一个难点，选择不好则会出现维度越多反而效果更差的现象，如何选择相辅相成的特征，尝试进行主成分分析可能是下一步的一个工作重点，另外，如何将机器学习的方法达到实用，与并行、大数据等技术进行融合，并且考虑真实社交网络的动态变化，是链接预测实际应用的一个关键。

基金项目

北京市自然科学基金(4172016,4152054)；北京市教委科研计划一般项目(KM201710011006)。

参考文献

[1] Lü, L.Y., Jin, C.H. and Zhou, T. (2009) Similarity Index Based on Local Paths for Link Prediction of Complex Networks. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, **80**, Article ID: 046122. https://doi.org/10.1103/PhysRevE.80.046122

[2] Lü, L.Y. (2010) Link Prediction in Complex Networks. *Journal of University of Electronic Science and Technology of China*, **39**, 651-661.

[3] Yin, Z.J., Gupta, M., Weninger, T., *et al.* (2010) A Unified Framework for Link Recommendation Using Random Walks. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Odense, 9-11 August 2010, 152-159. https://doi.org/10.1109/ASONAM.2010.27

[4] Kashima, H. and Abe, N. (2006) A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction. *Transactions of the Japanese Society for Artificial Intelligence*, **22**, 340-349. https://doi.org/10.1109/ICDM.2006.8

[5] Liben-Nowell, D. and Kleinberg, J. (2007) The Link-Prediction Problem for Social Networks. *Journal of the American Society for Information Science and Technology*, **58**, 1019- 1031. https://doi.org/10.1002/asi.20591

[6] Doppa, J.R., Jun, Y., Tadepalli, P., *et al.* (2010) Chance-Constrained Programs for Link Prediction. *European Conference on Machine Learning & Knowledge Discovery in Databases*, **6321**, 344-360. https://doi.org/10.1007/978-3-642-15880-3_28

[7] Getoor, L., Friedman, N., Koller, D., *et al.* (2002) Learning Probabilistic Models of Link Structure. *The Journal of Machine Learning Research*, **3**, 679-707.

[8] Taskar, B., Wong, M.F., Abbeel, P., *et al.* (2003) Link Prediction in Relational Data. *Neural Information Processing Systems*, 659-666.

[9] 赵姝, 刘晓曼, 段震, 等. 社交关系挖掘研究综述[J]. 计算机学报, 2017, 40(3): 535-555.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org