中国科学技术大学

学士学位论文

基于概率图模型的药物敏感性预测

作者姓名: 王 航

学科专业: 自动化系

导师姓名: 李 骜 副教授

完成时间: 二〇一八年六月五日

University of Science and Technology of China

A dissertation for bachelor's degree

Probabilistic Graphical Model Based Drug Sensitivity Prediction

Author : Hang Wang

Speciality : Automation

Supervisor : Ao Li, Associate Professor

Finished time: June 5, 2018

致　谢

我在中国科大健康信息学实验室的一个半学期里，实验室成员的关怀和爱给予了我持续不断前进的力量，培养了我无畏开拓的执着意志，以及迈出走向科研人生道路的第一步。我想借本文完成之际，首先表达我对健康信息学实验的由衷的感谢，谢谢你们!

每念吾师恩，如沐春风，似海深。特别的，我想感谢我的导师李骜教授和王明会教授。他们在我最初开始课题时的迷茫中为我指引出正确的方向，在我犹豫不决时给予我前进的动力，在我本科期间最困难的时候给予我无尽的帮助和一如既往的信任、理解。我极其荣幸能够在本科期间接受到李老师和王老师的教诲，身为世范，为人师表。他们为我标榜出尽职尽责，关心学生、甘乳一生的模范，他们让我明白，我将要成为一名怎样的研究者，成为一名怎样的教师，成为一名怎样的学生，他们让我更加坚定自己的奋斗决心和永不言败的勇气。谆谆如父语，殷殷似友情，轻盈数行字，浓抹一生人。谢谢您们!

学贵得师，亦贵得友。在科大的四年里，我有幸结识了许许多多志同道合的朋友。科研方面，我想特别感谢，实验室的习佳宁师兄。师兄在繁忙的博士研究期间，仍然极其耐心的为我解答论文写作的疑惑，对我完成我的研究工作，提供了莫大的帮助和支持。师兄于我，亦师亦友，师兄总是耐心的帮助我解决生活的困惑，提供力所能及的帮助，天下快意之事莫若友，快友之事莫若谈！此外，我要感谢我的室友，陈铮，向琳华，刘昆麟同学，我们融洽的寝室氛围使我四年的大学生活丰富难忘，谢谢你们，我的室友！此外，在过去的三年里，我有幸能成为科大学生会的一员，在这个有爱的组织里，我想特别感谢，李书祺，郭文静，涂佳妮，梁航同学，我们一起为科大的校园文化建设添砖加瓦，谢谢你们!

爱子心无尽，归家喜及辰。见面怜清瘦，呼儿问苦辛。人出生，日初生，求学之路，愈走愈远，父母之爱，愈发强烈。转眼，我已不再是仍需哺喂的婴孩，二十一年的人生旅途，永远陪在我身边的，永远惦念我冷暖的，永远操心我前程的，是我深爱的，爸爸妈妈。过去二十一年的人生之路，我无比自信与骄傲；未来的长路漫漫，我将愈发坚强与豪迈。在父母脊背顶起的这篇苍穹之下，我将一如既往，不负重托，一往无前，成为父母的好儿子、母校之骄傲、国家之栋梁!

深恩永存，厚谊永志。红专并进一甲子，科教报国六十年。自四年前离家赴中国科大，我愈发坚定自己的人生理想，愈发激励起自己人生斗志。我将永远牢记我的老师们，我的朋友们，长风破浪会有时，直挂云帆济沧海!

学生：王航

二〇一八年六月

献给

我亲爱的爸爸妈妈

目　录

中文内容摘要

近些年来，研究者通过系统的分析已知的药物敏感性数据，来有效的加速现代肿瘤精准治疗药物的研发。与此同时，通过对癌细胞药物反应的筛选，研究者可以得到为实现高精度药物预测而所需的基因组和药理学数据。目前已有的工作主要利用基因组或者功能基因特征来实现对药物敏感性信息进行分类或者回归分析。在这项工作中，通过对传统商品推荐系统方法和模型的迁移和推广，我们创造性的提出双层加强协同主题回归模型 (DS-CTR) 来实现精准的药物敏感性预测。这一模型不仅利用了概率图模型以从药物基因组学数据学习得到药物和细胞系的特征，同时利用药物和细胞的相似性网络加强预测结果的相关性。在这项工作中，选择用癌症药物敏感性基因组项目公共数据集 (GDSC) 作为基准数据集，通过五折交叉验证试验，同时将所提出的方法与其余四种先进算法对比，实验结果表明，DS-CTR 模型显著提高了药物敏感性预测的受试者工作特征曲线 (ROC) 及曲线下面积 (AUC $>$ 0.9)。通过对最新药物敏感性文献的检索，此模型预测得到的未知的药物敏感性也得到了验证。该模型同时为在临床前试验和加速抗癌药物的研究提供了可能性。

关键词：推荐系统；协同回归；矩阵分解；优化；药物敏感性

Abstract

An effective way to facilitate the development of oncology precision medicine is the systematical analysis of the known drug sensitivities that have emerged in recent years. Meanwhile, the screening of drug response in cancer cell lines provides an estimable genomic and pharmacological data towards high accuracy prediction. To classify or regress the drug response, existing works primarily utilize genomic or functional genomic features to classify or regress the drug response. Here in this work, by the migration and extension of the conventional merchandise recommendation method, we introduce an innovation model on accurate drug sensitivity prediction by using dual-layer strengthened collaborative topic regression (DS-CTR), which incorporates not only the graphic model to jointly learn drugs and cell lines feature from pharmacogenomics data but also drug and cell line similarity network model to strengthen the correlation of the prediction results. Using Genomics of Drug Sensitivity in Cancer project (GDSC) as benchmark datasets, our 5-fold cross-validation experiment demonstrates that DS-CTR model significantly improves drug response prediction performance compared with four categories of state-of-the-art algorithms as for both Receiver Operator Curve (ROC) and the Area Under Receiver Operator Curve (AUC $>$ 0.9). By uncovering the unknown cell-drug associations with advanced literature evidences, our novel model DS-CTR is validated and supported. The model also provides the possibility to make the discovery of new anti-cancer therapeutics in the preclinical trials cheaper and faster.

Key Words: recommender system, graphical model, matrix decomposition, optimization, drug sensitivity

第一章　引言

第一节　研究背景与意义

一、精准医疗

精准医疗 (Precision Medicine) 指的是综合考虑患者本身的个体基因、生活习惯和环境差异，进而为患者提供相关的疫病预防以及疾病诊治的新兴方法[①]。这一概念最早于 2011 年被美国医学界正式提出，在美国政府发布的《向精准医学迈进》的报告中，提出了要对不同特征的疾病进行更加明确的区分，并且对每种病症对症下药。区别于传统的使用疾病原发灶未知和有关的细胞学特征来分类的方法，这份报告提出要根据掌握的生物信息学信息进而建立生物医学知识网，从而能够更容易的对疾病进行区分，这种思想也逐渐得到国内外学界的重视和认可。2015 年，美国提出“精准医学计划”，我国也在同年提出在相关领域加大投资的计划。实际上，我国东汉时期张仲景在《伤寒杂病论》一书中就提出了辩证论治的思想，即每一个病例都应当使用不同的客观的治疗方案。追根溯源，个体化医疗，转化医学在某种程度上都有精准医学的影子。

精准医疗的主要内容，也随着时间的推进而不断的得到补充和完善。早在 2006 年，美国展开了癌症基因组图集计划，在此计划中，研究者关注与分析癌症基因信息，并期望以此分析癌细胞的突变特征；在之后的 2011 年，《向精准医学迈进》报告发布后，更加注重疾病分类以及药物的精准运用。不同与美国版的精准医疗，注重基因测序，个性化以及肿瘤，我国对精准医疗的定义则是：充分结合现代科技手段和传统的医学方法，更加科学的了解人体机能和疾病的本质，同时实现个体医疗服务和社会健康效益两者的最大化。尤其是近些年来，随着研究者在人类基因组测序、改进的生物医药分析和技术以及为生物信息大数据分析等领域有了突飞猛进的发展，精准医疗得到了更多的关注和支持。尤其在癌症治疗方面，精准医疗有望帮助实现成人、儿童的癌症靶向药物的创新型临床实验，综合治疗方案以及克服耐药性等目标。同时，通过收集相当数量的志愿者基因信息，生物采样标本、生活习惯信息以及电子健康记录信息，研究者有望更加

[①]美国国立卫生研究院 www.nih.gov/precisionmedicine

深入药物基因组学的研究，为病人提供合适的正确的治疗药物，同时通过对生物信息的分析，确定治疗和预防的新目标，为多种疾病精确用药奠定科学基础。总的来说，我国的精准医疗计划包含是三个层次[①]，即基础层次方面的基因测序，中等层次的细胞免疫治疗以及最高层次的基因编辑。

我国在精准医疗方面虽然起步比美国晚，但是我国精准医疗发展十分迅猛，并具有一些先天的优势，主要在于以下几个方面[②]，不同于美国医疗资源的分散性，我国医疗资源相对集中，因此可以有效的实现医疗机构之间的信息共享，尤其在癌症治疗方面，我国几乎百分之七十的癌症患者都集中在全国最顶尖的三百家医院中。有效的数据共享可以很好的辅助精准医疗的研究，我国也可以通过较少的资源投入来建立起肿瘤精准治疗的大数据网络；此外，我国人口基数大，随着癌症发病率的提升，这一方面向我国的疾病预防和控制工作提出了极大地的挑战，另一方面，这也使能够积累较多的案例信息，从而建立其完善的疾病数据库，更好的知道我国乃至全球的癌症治疗。我国的精准医疗计划同样也面临着一些问题，首先，我国在临床试验与技术的合作不够密切。通常，精准治疗的三个步骤，即检测基因、分析数据以及最后的临床注释环节。尽管基因检测已经较为成熟，但我国在数据发分析等方面仍有着很大的技术滞后。在当前的国际大形势下，我国的精准医疗机遇与挑战并存，精准医疗作为一门新兴的科学，中国推动精准医学的发展，将为精准医学研究平台和社会医疗健康服务的建设具有极大的推动作用。

二、药物敏感性预测

药物敏感性（drug sensitivity）也称为药物不耐受性（drug intolerance），通常指在治疗剂量或者亚治疗剂量（subtherapeutic）下所表现出来的对药物副作用的不耐受性质。通常，一种抗生素如果在剂量很小的情况下也能抑制、杀死致病细菌，就称之为“敏感（sensitivity）”。与之相反的，称之为“耐受的（tolerating）”或者不敏感的。[③]。传统的药物敏感性实验指的是通过生物学临床实验，了解病

[①]《科技部关于发布国家重点研发计划精准医学研究等重点专项 2016 年度项目申报指南的》

[②]2018-03-13，经济参考报，《中国精准医疗确立“自己的方向”》

[③]区别于耐药性（drug resistance），指的是病原体因与药物多次接触之后，对药物的敏感性下降甚至小时，进而使得药物对原来的病原体疗效下降或者无效

原微生物对于各种抗生素的敏感性，之后根据实验结果指导临床用药以及治疗。大致方法如下：首先从患者的感染部位采集含有致病菌的标本，之后接种在适当的培养基上，之后将沾有一定量各种抗生素的纸片贴在培养基表面，培养合适时间之后观察结果。这一过程往往需要耗费大量的人力财力，尤其对于特殊疾病的药物敏感型预测，抗肿瘤新药的实际临床评价也具有非常高的失败率，从而使得用于临床实验的开支是不容忽视的。本文将主要关注于癌症细胞以及治疗药物的敏感性问题。

肿瘤指的是机体在某种致瘤因子的作用下，局部组织细胞增生从而形成的新生物。根据这些新生物的性质以及对机体的危害程度，肿瘤可以分为良性肿瘤和恶性肿瘤，在医学上，称上皮组织中的恶性肿瘤称之为癌。数据表明，无论是在城市还是在农村，恶性肿瘤都是威胁中国居民生命健康最大的一个因素，并且我国的癌症死亡率位居全球癌症死亡率 29 位，高于全球平均水平百分之 17[①]。肿瘤的发生过程可能包含多个基因、多个步骤。传统的细胞毒药物依旧是临床上主要的治疗手段，这种方法的目的是杀死肿瘤细胞从而发挥作用。由于肿瘤发病的机制极其复杂、高危因素难以控制导致肿瘤的预防工作难；此外，由于有效的筛查技术少、早诊技术水平较低从而导致肿瘤发现较晚；肿瘤的治疗效果差，治疗的副作用显著且精准性差等进一步导致了肿瘤的治疗难度大。这些都是我国目前肿瘤防控研究的难点。

近年来，网络药理学、基因组学等学科以及相关的技术在药效评价和临床肿瘤分类中的应用，肿瘤的新信号通路以及其靶标的阐明，肿瘤的免疫疗法复兴，以及越来越多的靶向药物进入到世纪的临床评价，这些因素都使得抗肿瘤治疗向有效化、精准化等方向发展[1]。随着精准医疗的发展以及大规模药物基因组学数据库的建立，药物敏感性预测模型的研究吸引了研究者们极大的兴趣。充分利用测定的大量肿瘤样本的药物反应性形成的数据库，使得药物敏感性预测是个性化医疗的一个组成部分，从而最终实现针对患者量身定做治疗方案，而不是所有患者设计一种通用的治疗方案的诊疗目标。尤其是利用生物标志物信息进行肿瘤靶向治疗，使得治疗一些类型的肿瘤有了客观缓解，且 5 年生存率也有了较大程度的提高[2]。

药物敏感性预测算法的作用就是利用药物基因组学数据库、肿瘤样本的药

[①]2017-11-09,《中国肿瘤的现状和趋势》，中国科学院院士郝捷

物反应性形成的数据库等数据资源，对未知的或者未进行实验的药物名感性进行预测，从而达到减少药物研发开支，并实现精准医疗的目的。常用的算法模型包括弹性网，随机森林，核化贝叶斯多任务学习（kernelized Bayesian multi-task learning）等, 将在第二章进行具体的介绍。

第二节　推荐系统

一、推荐系统的定义

随着近些年来互联网的快速发展，人们经历了从信息匮乏转变为信息过载的时代变迁，信息消费者（客户）需要从过量的信息中筛选出自己感兴趣的信息，信息生产者需要使自己的的信息脱颖而出并最终得到信息消费者的关注。在这种情形下，推荐系统（recommender system）应运而生。推荐系统最早源于电子商务网站，例如亚马逊，用来帮助客户决定购买哪些产品，模拟在商场中的销售人员向客户的推荐商品过程。个性化推荐则是考量用户的兴趣和购买行为，从而向用户推荐其感兴趣的信息和商品。从物品角度出发，根据“长尾效应”，百分之八十的销售额将来自于百分之二十的热门品牌。在互联网商务网站上，由于货架成本低，商家往往能够比实体商店销售更多的商品。这种情形下，剩余百分之八十的非热门商品的销售额也将不容小觑。从客户角度出发，热门商品往往代表着主流消费，而非热门商品则象征着个性化的消费需求，推荐系统就旨在解决这一问题。推荐系统充分分析已有的数据资源，从而发现用户的个性化需求，进而将个性化商品推荐给相关的用户。推荐算法的本质是通过合适的方式将用户和商品联系在一起。正因如此，推荐系统已经被广泛的利用到多个领域，例如电影、音乐、社交、网络商务等。

二、推荐系统的分类和应用

根据推荐的结果，可以将推荐系统分为 TOP N 推荐和评分预测两种[3]。在评分类网站中，例如豆瓣、亚马逊，用户能够提供对商品显式的分数评价。利用用户-商品评分矩阵，来预测用户对于未知商品的评分或者预测新用户对某类商品的评分即为评分预测。研究者通常使用矩阵分解等方法完成预测任务[4]。推荐作为一个优化问题，需要确定损失函数（loss function），通常研究者选定的损

失函数就选择预测的评分与实际的评分的平均平方差的根（RMSE，Root Mean Square Root）或者平均绝对误差（MAE，Mean Square Error）。对于测试样本 T 中的用户 u 和商品 i，如果实际用户对商品评分为 r_{ui}，而推荐系统预测得到的评分为 $\hat{r}_{ui}$，则 RMSE 和 MAE 的定义分别为：

$$RMSE = \frac{\sqrt{\sum_{u,i \in T} (r_{ui} - \hat{r}_{ui})^2}}{|T|}$$

$$MSE = \frac{\sum_{u,i \in T} |r_{ui} - \hat{r}_{ui}|}{|T|}$$

另一方面，对于 TOP-N 推荐，即为用户提供一个推荐列表。一般用于不能获得显式评分的情形中。需要研究者通过将隐式反馈信息加以提取建模，最后预测得到用户可能感兴趣的商品列表。TOP-N 推荐常用的方法是贝叶斯个性化评分[5]。TOP-N 预测的效果通常使用准确率（precision）以及召回率（recall）来度量，设利用训练集得到的用户可能感兴趣的商品列表为 $R(u)$，利用测试集得到的用户可能感兴趣的商品列表为 $T(u)$，则准确率和召回率分别定义如下：

$$Precision = \frac{\sum_{u \in U} |R(u) \cup T(u)|}{|R(u)|}$$

$$Recall = \frac{\sum_{u \in U} |R(u) \cup T(u)|}{|T(u)|}$$

根据推荐过程用到的信息类型的不同，推荐算法又可以分为基于内容的推荐和基于协同过滤的推荐（CF，collaborative filtering）。基于内容的推荐思想是根据根据商品或者内容的元数据，发现其相关性，再结合用户曾经喜欢的商品（item），为用户推荐于这些商品类似的商品，这种方法仅仅考虑对象（商品或者内容）的本身性质，将对象按标签进行分类集合，当消费者消费一个集合中的商品时，则推荐系统倾向于向消费者推荐同类商品，这种方法还广泛应用在信息检索和信息过滤等方面。基于协同过滤的推荐算法又可以细分为基于用户（user-based）、基于项目（item-based）和基于模型（model-based）的协同过滤推荐。基于用户的协同过滤主要依赖于用户之间的相似性，其核心在于通过对用户行为的分析得到用户的相似性信息，最终向用户推荐与之相似的用户曾经购买过的商品。基于项目的协同过滤则主要是根据用户对商品的评价信息发现商品的相似性，然后根据用户偏好将相似商品予以推荐。基于模型的协同过滤是利用

用户的偏好训练得到偏好模型，再根据该模型进行推荐。与基于内容的推荐系统不同，协同过滤推荐算法能够更充分的利用用户的行为信息以及群体智慧[①]，并强调 1）兴趣类似的用户会对同样的商品感兴趣以及 2）用户会比较偏好与自己曾经购买过的商品类似的商品。关于协同过滤方法的实现原理，将在第三章进行系统的阐述。

第三节　研究现状及进展

近年来，研究人员从分子特征广泛研究了癌症的特征，并在这一领域取得了重大进展，如大规模高通量筛选成功的成功案例。这种研究分析了大量人类癌细胞系和药物的反应特性[6][7][8]，为研究者提供了更多可利用的信息。预测性生物标志物鉴定的问题也已经通过分析临床相关的人类细胞系和相应药物的药理学反应而被系统地解决，所述药物在一些研究项目例如癌细胞系百科全书（CCLE）[8] 和癌症基因组计划（CGP）[9]。同时，对于每个细胞系，基因突变状态和表达谱也包含在研究者的工作中[10]。因此，基于这些进展，建立高准确度的癌症细胞对药物反应的预测系统可以将肿瘤基因组信息系统性地转化为肿瘤生物学和治疗知识[11]，随着对癌症更深入的了解，药物敏感性预测系统也将更容易实现同时也更加可靠。

为了给精准医疗所指出的个性化癌症治疗提供更多的参考信息和指导，研究者已经开发了多种药物反应预测方案[12]。最近的研究进展表明，当预测某些类型细胞系中的药物反应时，基因组信息和细胞分子特征是必要和有用的[13]。因此，过去许多研究工作都是专注于基于体外药物敏感性和治疗前的基线基因表达水平来预测细胞对于抗癌药物的敏感性反应[9][14][15][16] 。基于之前提到的 CCLE 数据集，Geeleher 等人专注于乳腺癌细胞系，应用基线基因表达数据的稀疏回归模型展开预测[12]。Menden 等人也在使用药物的化学性质。以及细胞系的基因组特征来表示每种细胞系 - 药物对的特征来进行预测[16]。在 Ammad-ud-din 等人的工作中，除了细胞系的基因组特性之外，通过系统地利用已知的药物主要靶点（药物特性矩阵）作为先验知识，之后的预测结果最终通过核化贝叶斯矩阵

[①]集体智慧是指通过分析大量的人群的行为和数据，进而能够总结得到对于整个人群都具有统计意义上的结论。

分解[8][13] 得到。进一步，基于药物或细胞系的相似性网络可能潜在地提高药物反应预测质量的假设，研究人员已经完成了大量工作。例如，在 Zhang 等人的工作中，他们使用细胞相似性网络和药物相似性网络分别预测药物应答，即分两步进行预测，最终的预测结果是之前得到的两个网络结果的加权平均值[10]。同时，在 Berlow 等人的工作中，研究人员根据药物的抑制曲线以计算药物相似性，提出了基于综合功能和基因组表征的药物反应预测的新方法[17]。

如前一段所示，大多数药物敏感性方法仅依靠基因特征，例如突变基因和靶基因[12][14][16]，却忽视了癌细胞 - 药物反应筛选的两个相当重要的特征：一个是具有相似遗传特征的细胞系或样品可能对给定药物的反应非常相似，另一个线索是如果在考虑共享分子结构或其靶向时使用结构相似的药物，则可能出现类似的治疗效果[10]。相关研究表明，仅基于基因组信息的预测通常导致准确性和局限性较低[17][6][18]。更进一步，虽然在 Zhang 等人引入细胞相似性网络和药物相似性网络来进行预测，但细胞的分子结构特征和基因组特征却没有能够在模型中得以利用，这些特征信息已经在很多工作中被证明是对预测起积极作用的[13]。另一方面，还应当在此强调统计强度的问题，因为大多数药物反应研究的样本量很小，这使得预测任务非常具有挑战性，并可能使预测结果变得不可靠[19]。

为了克服上述问题和局限性，首先，有必要建立一个有效和充分的计算预测框架来组合可用的多视角数据源，提升预测表现，并从根本上提供更大限度的合理预测信息。做出此结论基于的假设是，当预测响应变量时，可能有用的信号/特征将由全部或部分数据源提供[19]。例如，两种药物之间的相似关系会使得对某种细胞系敏感性的预测值上调或者下调。的工作期望从不同的数据源推断出预测结论，这些计算方法也通常被称为多视角学习[19][20] 。

第四节　论文内容及结构

在本文中，提出了一种基于协同过滤（Collaborative Filtering）的新方法来预测与基因组信息于药物-细胞相似网络之间的相互依赖关系。在这项工作中，提出的双层加强主题回归模型（Dual layer Srengtheded-Collaborative Topic Regression），该模型包含两个层次来整合两类必要的信息：第一层是主题层，第二层是关系

层。第一层使用传统的协同过滤和概率主题建模[21][22]给出了药物和细胞的可解释特征结构。而第二层则反映了所谓的药物或细胞的相似关系，并利用相似性网络加强了推荐结果的相关性，从而提高了预测效果。与 Zhang 等人的工作不同，在本文中通过设计一个概率图模型，并直接优化一个损失函数来进行建模完成预测任务。通过这种方式，的模型将基因组信息和相似性信息考虑在内，这些信息已经被证明对于药物敏感性预测起积极作用[9,13–15]。最后，在这项研究中，选择了一个真实世界的数据集 GDSC[9] 来评估新提出的 DS-CTR 模型的优越性。结果表明，提出的模型在受试者工作特性曲线（ROC）和 ROC 曲线下面积（AUC）方面都优于另外四种最先进的药物反应预测系统。凭借最新的文献资料，DS-CTR 模型的预测结果通过揭示未知的细胞药物关系而得到进一步的验证。

本文的主要结构如下：

第二章将系统的介绍几种常见的药物敏感性预测方法以及原理，包括弹性网、随机森林两种传统的方法，这些方法的基本思想将在的模型中有所体现。同时详细介绍了贝叶斯方法和相似性网络等更先进的药物预测模型，这两种模型将在后续章节中的与提出的模型进行比较。

第三章介绍了图模型的基本概念以及基本原理，包括最基础的矩阵分解的图模型解释，以及通过使用图模型所展开的文本主题挖掘 LDA 模型，在这一章中还详细介绍了协同主题回归模型，该模型作为 DS-CTR 模型的子模型将予以详细介绍。之后介绍了新提出的 DS-CTR 模型的基本原理和学习过程，首先进行理论推导，之后介绍了为了解决该优化问题所提出的 EM 算法模型。

第四章将会详细介绍测试用数据集的特征，最终将 DS-CTR 模型与另外几种先进模型进行对比，并通过最近的文献资料予以验证。

最后，第五章将给出了本文的研究结论和将要开展的工作。

第二章　药物敏感性预测算法简介

实现精准医疗意味着医疗工作者将能够针对患者提供个性化的治疗方案。因此，能够将基因组序列等信息转化为预测患者如何对给定药物做出反应是至关重要的。这一章节总结了几中常用的药物敏感型推荐算法，这些算法利用基因组学数据以及药物敏感性特征信息来预测：1）某种新的肿瘤或者细胞系对于现有药物或其组合的敏感性；2）现有的肿瘤培养物对新药物的敏感性[23]。，首先在经典方法一节中简要介绍了随机森林、弹性网等预测方法。针对经典方法的局限性，之后介绍了相似性网络的方法，最后将介绍经典贝叶斯方法以及基于贝叶斯核的方法模型。

第一节　经典方法

一、弹性网

在执行预测任务时，尤其在使用一些机器学习方法时，往往会过拟合或者欠拟合的情况。为了避免过拟合现象出现，研究者对线性回归（Linear Regression）进行了优化，于是产生了 Ridge、LASSO 还有 ElasticNet 回归。然而，LASSO 倾向于生成更稀疏的模型，但受数据集中样本数量的限制，而 Ridge 回归模型在找到相关特征集方面效果较好，但缺乏 LASSO 模型的稀疏性。在这种情形下，2005 年 Zou 等人提出弹性网模型[24]，弹性网（Elastic Net）是一种使用 L1 和 L2 先验作为正则化矩阵的线性回归模型，这种组合不仅可以用于很少的权重非零的稀疏模型之中，比如 LASSO, 同时也能保持 Ridge 的正则化属性。可以使用特定的参数来调节 L1 和 L2 的这种凸组合。

正因为弹性网的这些优良特性，Sokolov 等人在 2015 年提出了一种利用弹性网络（Generalized Elastic Net）方法来解决药物敏感性预测的问题上[25]。研究者提出通过使用模型正则化来整合有关特征的领域知识，例如基因相互作用信息。特别提出了将基因相互作用添加到大量的监督和无监督预测方法。

二、随机森林

随机森林（random forest）是一种常用的有监督机器学习算法，该算法基于决策树（decision tree），最早在 2001 年被 Breiman 提出[26]。决策树是一种用来对实例进行分类的树状结构，决策树是由若干的有向边和结点组成。其中结点有两类：叶节点和内部节点，叶节点用来表示一个类，内部节点则表示一个特征或者属性。在进行分类的时候，由根节点起始，测试实例的某一特征，之后根据测试结果，再将该实例分配到其子结点；此时，每一个子结点都将对应着这一特征的一个取值。如此继续向下递归，直至达到最终的叶结点，至此就将实例分配到了叶结点所表示的类中。随机森林算法，则是把单独的分类树组合成随机森林，即在数据（即行）的使用和变量（即列）的使用上分别进行随机化，生成许多的分类树，之后汇总这些分类树的结果，因此随机森林也是一种集成学习方法（Ensemble Learning）。随机森林算法的运算量在没有显著提高的条件下，提高了算法的预测精度，并能够有效的运行在大数据集上，同时在解决缺省问题时也具有很好的性能，因此备受研究者关注[26]。随机森林的实现过程如下：

Data: 训练集：$S=(x_i,y_i), i=1,2,\cdots,n, X,Y\in R^d\times R$；待测样本：$x_i\in R_d$

for $i=1,2,\cdots,N_{tree}$ **do**

  对原始训练集 S 进行 Boostrap 采样，生成训练集 S_i;

  使用 S_i 生成一棵不剪枝的树 h_i:;

  a. 从 d 个特征中随机选取 M_{try} 个特征

  b. 在每个节点上从 M_{try} 个特征依据 Gini 指标选取最优特征

  c. 分裂直到树长到最大;

end

Result:

(1). 树的集合：$h_i, i=1,2,\cdots,N_{tree}$；

(2). 对待测样本 x_i，决策树 h_i 的输出 $h_i(x_i)$

算法 2.1: 随机森林算法

Gregory Riddick 等人在 2011 年提出使用随机森林的方法来进行药物敏感性

预测[27]，在这项工作中，研究者使用两个数据源来构建预测系统：NCI-60[①]中所有细胞系对该药物的药物敏感性（IC_{50}）[②]和 NCI-60 中每个细胞系的基础基因表达（细胞在应用任何药物之前的“静息”生理状态）。使用 Random Forest 创建药物反应模型主要由三个步骤组成。首先，将特定药物的 IC_{50} 应答数据标准化为 [0,1] 区间。然后对随机森林模型在 NCI-60（16 644 探针组）的基础基因表达数据上进行训练。然后使用该模型产生的变量重要性（Variable Importance）来选择对药物反应具有高度预测性的探针组的较小子集（通常 100-500 个探针组）。在这一步中，基因表达标记（Signature）和 IC_{50} 应答之间拟合成为一个模型。去除了离群细胞系的 IC_{50} 值，就可以拟合得到基因表达特征的模型。该方法相较于弹性网的方法有所提高，但是在建立该模型时，未能考虑到基因之间的相互作用，以及其余的信息源，例如基因的突变特征和药物自身的结构特征等信息，因此该模型的预测能力仍然有较大的提升空间。

第二节　贝叶斯方法

一、贝叶斯决策论

贝叶斯决策论（Bayesian Decision theory）是主观贝叶斯派归纳理论当中非常重要的一个组成部分。尤其在解决分类问题时，贝叶斯决策论要求相关的概率都已知，在此条件下利用误判损失函数和这些已知的概率来执行分类任务，即完成对样本的类别标记[28]。首先假设的分类任务要求将样本分为 N 类, 即 $\mathcal{K} = c_1, c_2, \cdots, c_N$，此外定义 λ_{ij} 为将真实属于 c_i 的样本错误分类为 c_j 的损失（loss）例如：

$$\lambda_{ij} = \begin{cases} 1, & i \neq j \\ 0, & i = j \end{cases}$$

此时，在样本 **x** 分类（决策）至 c_i 所产生的条件风险（conditional risk），即期望损失 (expected loss) 为：

$$R(c_i|\mathbf{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j|\mathbf{x})$$

分类任务的核心是寻找合适的判定准则 $h : \mathcal{X} \mapsto \mathcal{K}$ 以最小化总体的条件风

[①]NCI DTP 网址：http://dtp.nci.nih.gov。

[②]IC_{50} (half maximal inhibitory concentration) 指的是被测量的拮抗剂的半抑制浓度。

险，即：$R(h) = \mathbb{E}_{\mathbf{x}}[R(h(\mathbf{x})|\mathbf{x})]$, 最终贝叶斯判定准则（Bayes Decision Rule）可以总结为选择使得条件风险 $R(c|\mathbf{x})$ 最小的映射 h^*，进而最小化总体风险 $R(h)$，即有 $h^*(\mathbf{x}) = \underset{c\in\mathcal{K}}{argmin} R(c|\mathbf{x} = \underset{c\in\mathcal{K}}{argmin}(1 - P(c|\mathbf{x}))$，等价于 $h^*(\mathbf{x}) = \underset{c\in\mathcal{K}}{argmax} P(c|\mathbf{x}))$（后验概率最大化与期望风险最小化等价），$h^*$ 被称为贝叶斯最优分类器（Bayes Optimal Classifier），总体风险 $R(h^*)$ 称为贝叶斯风险（Bayes Risk）[16]。

为了获得后验概率 $P(c|\mathbf{x})$，通常可以通过以下两种方式得到：已知 $\mathbf{x}$，可以直接计算 $P(c|\mathbf{x})$ 来预测 c; 或者先求得联合概率分布 $P(\mathbf{x}, c)$, 之后再获得 $P(c|\mathbf{x})$, 即有 $P(c|\mathbf{x}) = \frac{P(\mathbf{x},c)}{P(\mathbf{x})}$，依据贝叶斯定理，$P(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)P(c)}{P(\mathbf{x})}$，其中 $P(c)$ 是类先验概率，表示样本当中的各类样本的比例，独立同分布且样本量足够大时，依据大数定理，先验 $P(c)$ 可以根据各种类别样本出现的频率来估计；$P(\mathbf{x}|c)$ 是确定的样本 $\mathbf{x}$ 相对于标记 c 的类条件概率，或者称之为似然（likelihood），似然的估计可以通过朴素贝叶斯分类器来完成[28]。

二、参数估计

极大似然估计（Maximum likelihood estimation, MLE）和最大后验概率估计，即 Maximum a posteriori estimation（MAP），是两种常用的参数估计方法。极大似然估计最早由高斯（C. F. Gauss）提出，对于类条件概率 $P(\mathbf{x}|c)$ 的估计，需要事先假定其具有确定的概率分布并只与参数 θ 有关，之后再根据训练样本对该概率分布中的相关参数进行估计，对于离散样本值 $\mathbf{x}$，设其分布率 $P(X = x) = p(x;\theta)$，其中 θ 是待估计的 k 维参数，其形式已经确定（例如正态分布等形式），$X_1, X_2, ..., X_n$ 是从总体中抽出的样本，则它们的联合概率分布为：

$$\prod_{i=1}^{n} p(x_i;\theta)$$

设 $x_1, x_2, ..., x_n$ 是对应的样本值，则事件：由样本 $X_1, X_2, ..., X_n$ 观察得到 $x_1, x_2, ..., x_n$ 的概率为：

$$L(\theta) = L(x_1, x_2, ..., x_n;\theta) = \prod_{i=1}^{n} p(x_i;\theta)$$

这一概率由 θ 唯一确定，则在 θ 的所有可取范围内使得上述似然函数达到最大的参数 θ^* 即为 θ 的估计值，使得：

$$L(x_1, x_2, ..., x_n;\theta^*) = \underset{\theta}{argmax} L(x_1, x_2, ..., x_n;\theta)$$

通常为了避免连乘时出现下溢，通常会使用对数似然函数（log-likelihood），即在上式两边取对数，最终通过求导等方法得到参数的估计值。

不同于频率学派，在贝叶斯学派中，常使用 MAP 的方法进行参数估计。这种方法在极大似然估计的基础上同时考虑了被估计量的先验分布因此也被称为极大似然估计的规则化（rugularization）。在似然估计的基础上，假设模型参数 θ 具有一定的先验分布 $p(\theta)$，即参数 θ 是一个超参数（hyperparameter）。给定观测值（样本），最大化后验概率，使得该参数的概率最大，即：

$$\hat{\theta} = \underset{\theta}{argmax} p(\theta|\mathbf{x})$$

进一步根据贝叶斯公式，展开上式得：

$$\theta^* = \underset{\theta}{argmax} \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})} = \underset{\theta}{argmax} p(\mathbf{x}|\theta)p(\theta)$$

对比两种参数估计方法，最大似然估计的目的是求取满足似然函数 $P(\mathbf{x}|\theta)$ 最大的 θ^*，最大后验概率则是估计满足 $P(\mathbf{x}|\theta)P(\theta)$ 最大的 θ^*，即：

$$\theta_{MAP}(\mathbf{x}) = \underset{\theta}{argmax} p(\mathbf{x}|\theta)p(\theta)$$

三、EM 算法

现实情况中，训练样本中往往包含隐式（latent）变量，即为未观测或者未知的变量。为了能够实现参数估计的目的，Dempster 等人于 1977 年提出期望最大化算法（Expectation-Maximization）[29]。这一算法是一种利用坐标下降（coordinate decent）的迭代优化算法，每次迭代中包含两个部分，第一步称为期望（E）步，第二步叫做极大化（M）步。该算法的基本思想是首先根据已观测得到的变量估计出模型参数，之后根据估计出来的参数值估计隐式变量，再将这部分变量于已观测变量整合再对模型参数进行估计，反复迭代，直到收敛或者达到最大迭代次数。令已经观测得到的变量集合记为 $\mathbf{X}$，未猜测得到的变量集合为 $\mathbf{Z}$，模型的参数为 Θ，对 Θ 做极大似然估计（用对数表示），即：

$$LL(\Theta|\mathbf{X}, \mathbf{Z}) = lnP(\mathbf{X},\ \mathbf{Z}|\Theta)$$

通过计算 $\mathbf{Z}$ 的期望，解决原问题中由于隐变量而无法求解的情况，即最大化已经观测得到的变量的边际似然（marginal likelihood）：

$$LL(\Theta|\mathbf{X}) = lnP(\mathbf{X}|\Theta) = ln\sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z}|\Theta)$$

设 Θ 的初始值为 Θ^0，迭代到第 t 步时的参数值为 Θ^t，隐变量集合记为 $\mathbf{Z}^t$，则 EM 算法可以表示为：

- **E** 步：根据 Θ^{t-1} 推断得到 $\mathbf{Z}^{t-1}$
- **M** 步：根据 $\mathbf{Z}^{t-1}$ 和 $\mathbf{X}$ 极大似然估计参数得到 Θ^t

Zhou 等人还提出，类似的如果在 **E** 步计算隐变量集合 $\mathbf{Z}$ 的概率分布，则 EM 算法可以总结如下 [16]：

- **E** 步：根据 Θ^{t-1} 推断得到隐变量集合 $\mathbf{Z}$ 的分布 $P(\mathbf{Z}|\ \mathbf{X}, \Theta^{t-1})$，之后计算关于 $\mathbf{Z}$ 的对数期望 $\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Theta^{t-1}} LL(\Theta|\ \mathbf{X}, \mathbf{Z})$
- **M** 步：根据极大似然估计的原理，计算得到 $\Theta^t = \underset{\Theta}{argmax}\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Theta^{t-1}} LL(\Theta|\ \mathbf{X}, \mathbf{Z})$

EM 算法的思想在很多算法模型中均有体现，例如 K-means 算法，GMM（高斯混合模型）算法。Wu 等人指出，EM 算法可以保证收敛到一个稳定点，但不能保证收敛到全局极大值点，因此它是一个局部最优[①]算法[30–31]。然而 EM 算法对于初始值较为敏感，参数 Θ 的选取可能会影响算法最终达到全局最优解。

第三节　相似性网络方法

在最新的两个相关项目，癌症细胞系百科全书)[6] 和癌症基因组计划[8] 通过分析约 1000 个临床相关的人类细胞系及其对于 149 种癌症药物的药理学特征，系统地解决了预测性的生物标志物识别的问题。这两项研究还包括每个细胞系的基因表达谱和突变状态，并应用弹性网络模型选择预测药物反应的表达和突变特征。基于相同的数据集，Geeleher 等人应用另一种稀疏回归模型 Ridge，用基线基因表达数据预测乳腺癌细胞系的药物反应[10]。Menden 等人细胞系的整合基因组特征（突变，拷贝数和微卫星[②]不稳定性）与药物的化学特性来表示每一细胞系-药物作用对，并使用神经网络预测 CGP 数据集中的药物反应[12]。尽管这些方法对于某些药物的敏感性预测取得了比较理想的结果，但他们都没有考

[①]局部最优，指的是所求得的解仅仅在有限范围或者区域内最优，或解决该问题的方法只有在设定一定的范围或限制条件下达到内最优。

[②]微卫星指的是遍布于人类基因组中的短串联重复序列。其不稳定与肿瘤的发生密切相关。

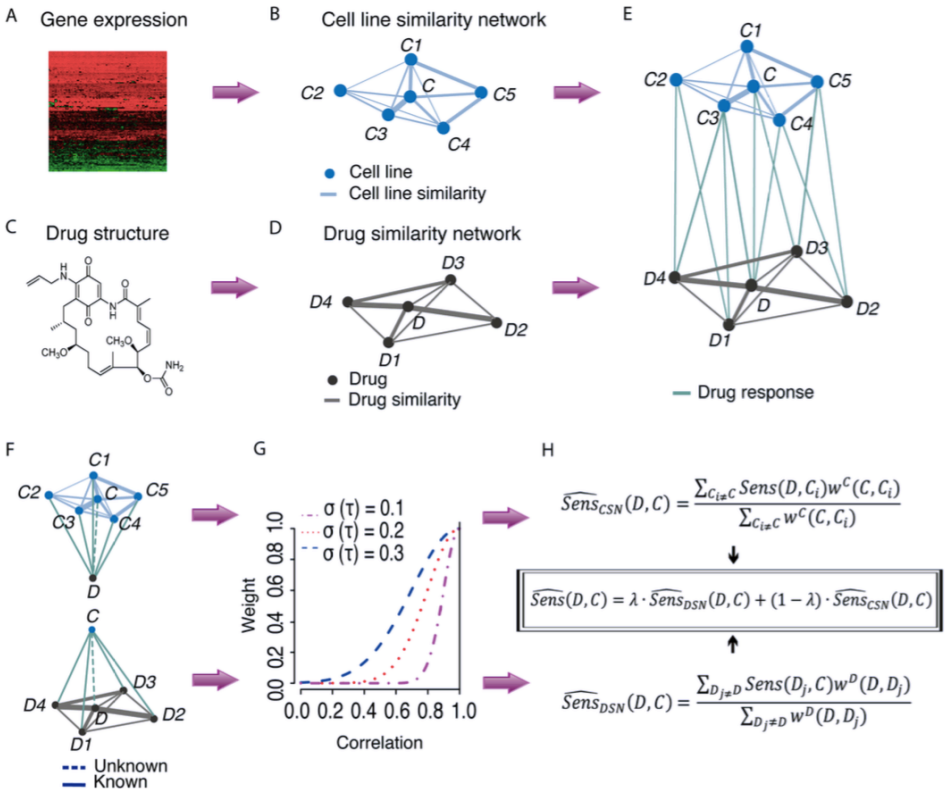

图 2.1　相似性网络计算框架示意图[9]。（A-E）整合双层细胞系 - 药物网络模型。（F-H）基于来自细胞系 - 药物网络的信息的药物敏感性预测。（F）用于预测细胞系 C 对药物 D 的响应的两个子网络。（G）对于每个子网络，使用加权算术平均值来估计基于其相邻细胞系或药物的未知药物 - 细胞系敏感性。（H）使用加权模型合并来自两个单独子网络（CSN、DSN）的预测结果以确定最终预测。

虑癌细胞药物应答筛选的两个重要特征，包括：1）基因特征相似的细胞系或样品可能对于给定药物的反应也非常相似; 和 2）由于共有的分子结构或靶向模式，结构相关的药物可能具有相似的治疗效果。

Zhang 等人在研究中验证了以上两个特征，做出整合细胞系和药物之间的相似性可能会改善药物反应预测的假设，并最终提出了药物相似性网络（Drug Similarity Network，DSN）和细胞系相似性网络（Cell Similarity Network，CSN）两个概念[9]。研究者们构建出一个双层细胞系-药物网络，并基于基因表达谱建模了细胞系之间的相似性，基于一维和二维化学结构建模了药物之间的相似性。在这项工作中，作者利用细胞系相似性网络（CSN）或药物相似性网络（DSN）的一层或两层的加权模型来预测给定细胞系对药物的敏感性。研究者使用 CCLE 和 CGP 作为基准数据集，并评估了模型的预测能力，发现双层整合细胞系-药物

网络模型显着优于单独使用 CSN 或 DSN 层的模型，以及弹性网模型。在实际的测试中，作者应用双层网络模型来填补 CGP 数据集中所有遗漏的药物反应值（活性区域和 IC_{50}），并且发现模型预测得到的 MEK1/2 抑制剂在 CGP 研究中的响应与其他可用药物的响应值具有非常相似的分布[9]。

该模型（Fig. 2.1）主要整合了三种类型的数据：1）每种细胞系的基因表达谱；2）每种药物的一维和二维化学结构特性；和 3）每种细胞系的药物反应。网络的顶层，称为细胞系相似性网络，这一层预测细胞系 C 对与药物 D 的敏感性，再为与细胞 C 具有相似基因表达特征的细胞的药物应答信息进行加权。该模型计算了细胞 C 与其他细胞系之间基因表达的相关性，对于更相似的细胞赋值更高的权重（见 Fig. 2.1A 和 Fig. 2.1B）。底层，称为药物相似性网络（Fig. 2.1C 和 Fig. 2.1D），用来预测细胞 C 对于药物 D 的敏感性，同样对与药物 D 具有相似化学结构的药物敏感性信息进行加权，同时，在这一层根据在 PubChem[①] 系统中的一维、二维的结构特征来计算每一对药物之间的相似性[32]。这两层网络通过细胞系的药物反应数据进行连接 (Fig. 2.1E)，这些数据在 CCLE 和 CGP 研究中被表示为活动区域。值得注意的是，这个网络并不是一个完整的二分图（Fig. 2.1F），因为这些研究中一些细胞系的药物反应数据缺失，的目标就是预测未知的药物敏感性数据。

该模型共包含三个参数，分别用来决定不同细胞系 ω^C 的权重、不同药物 ω^D 的权重以及两个网络（CSN，DSN）的权重 λ。其中 ω^C 的定义式为：$\omega^C(C, C_i) = \mathrm{e}^{\frac{-[1-\rho^C(C,C_i)]^2}{2\delta^2}}$，其中 δ 决定当细胞表达相似性降低时的衰退率（decay rate），其范围在 [0, 1] 区间，增量为 0.001；$\rho^C(C, C_i)$ 指的是细胞系 C_i 与 C 之间的相关性。同样的 ω^D 的定义式为：$\omega^D(C, D_i) = \mathrm{e}^{\frac{-[1-\rho^D(D,D_i)]^2}{2\tau^2}}$，其中 τ 决定当药物相似性降低时的衰退率（decay rate），其范围在 [0, 1] 区间，增量为 0.01，根据实验结果确定 δ（进而确定 ω^C）和 τ（进而确定 ω^D）之后，计算平衡 CSN 和 DSN 网络的权重系数 λ 以使平方误差和最小。在这项工作中，研究者定义细胞系 C 和药物 D 之间的敏感性为：

$$\widehat{Sense}_{CSN(D,C)} = \frac{\sum\limits_{C_i \neq C} Sense(D, C_i)\omega(C, C_i)}{\sum\limits_{C_i \neq C} \omega(C, C_i)}$$

其中 $Sens(D, C_i)$ 指的是细胞系 C_i 对于新药物 D 的敏感性数据（预测值），C_i

[①] Wang 等人于 2009 年提出的一个用于分析小分子生物活性的公共信息系统

指的是在 CSN 网络中与细胞系 C 相关联的细胞，根据模型假设，同样细胞系 C 相关的细胞系与其他细胞系相比，对于 $Sens(D, C_i)$ 的贡献更大。同样类似的可以定义预测药物对于新的细胞系的作用表达式：

$$\widehat{Sense}_{DSN(D,C)} = \frac{\sum\limits_{D_j \neq D} Sense(D_j, C)\omega(D, D_j)}{\sum\limits_{D_j \neq D} \omega(D, D_j)}$$

为了充分利用 CSN 和 DSC 网络，进一步研究者将两种网络通过权重系数 λ 联系起来，就有：

$$\widehat{Sense}_{DSN(D,C)} = \lambda \cdot \widehat{Sense}_{DSN(D,C)} + (1-\lambda)\widehat{Sense}_{CSN(D,C)}$$

上式的权重系数 λ 可以通过留意交叉验证来进行优化，但 λ 值趋近于 1 时，集成模型将由 DSN 控制，相反，当 λ 趋于 0 时，由 CSN 控制，当 $0 < \lambda < 1$ 时, 则是两个单独模型互相补充。

第三章　双层协同主题滤波模型的构建

第一节　图模型的概念

概率图模型是在人工智能领域的一个主要的研究方向，这种方法将图论与概率论结合在一起，最早由 MIT 的 Michael. I. Jordan 提出[33]。为了描述复杂系统，往往利用图的形式帮助研究者分析系统，并用图来突出变量之间的概率关系，即利用图的形式来描述与模型有关系的变量的概率分布，或者实体之间的约束关系。概率图模型由图结构组成，包含节点和边。其中，每一个节点表征一个随机变量，而边表示这些随机变量之间的约束关系。图的模式可以分为两大类：有向图也叫做贝叶斯网络和无向图，即马尔科夫网络。在本文中，主要关注贝叶斯网络。贝叶斯网络的结构是无环的，因此它也被称为有向无环图网络。

在与概率图模型类似的概率模型中，利用已经观测得到的变量来推测没有观测得到的（未知的）变量的分布即推断。设已经观测得到的变量集合为 O，期望预测得到的变量集合为 Q，剩余其他的变量的集合为 R，联合分布 $P(Q,R,O)$ 表征生成式模型（Generative Model），条件分布 $P(Q,R|O)$ 表征的是判别式模型（Discriminative Model），推断即由判别式模型或者生成式模型得到条件概率分布 $P(Q|O)$ 的过程。由于变量之间有会存在着极为复杂的联系，因此用前面章节讲的样本训练来估计模型参数的方法往往会有过高的计算复杂度，例如，如果用 x_i 来表示变量 X_i 的其中一个取值（在其可能取值范围内），此时这 K 个变量的联合概率分布可以表示为 K 个条件概率的积：$p(x_1,x_2,\cdots,x_k) = P(X_1 = x_1,,X_2 = x_2,\cdots,X_K = x_k) = \prod_{k=1}^{K} p(x_k|x_1,\cdots,x_{k-1})$ 特别的，对于多元变量，为了简化计算，利用局部马尔可夫性质（**定理 3.1**），可以将其联合分布予以简化。概率图模型也正因为其简介、直观的特点受到越来越多研究者的青睐。

定理 3.1　（**局部马尔可夫性质**）　贝叶斯网络拥有局部马尔可夫性质，指的是在给定其父节点的情况下，每个随机变量都条件独立于其非后代的节点。

概率图模型的参数估计，指的是对具体分布的参数进行估计，同样主要依赖于在第二章讲述的两种方法，极大似然估计方法和最大后验概率估计的方法，若存在隐含变量，则同样需要借助 EM 算法实现参数估计。然而贝叶斯学派认为，待确定参数仍然属于随机变量，因此通常将参数看作是待推测的随机变量，

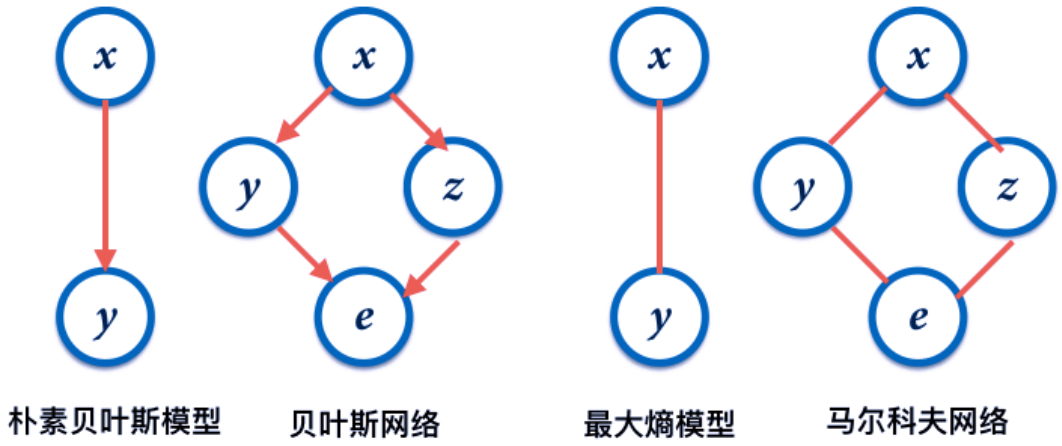

图 3.1 有向图与无向图示意图。贝叶斯网络是有向图，马尔科夫网络是无向图，并且有向图的联合概率能够写作各项条件概率之积，无向图的联合概率能写作最大团的随机变量函数之积

进行“推断”。这一小节将主要专注概率图模型的推断过程。按照推断方法进行分类，概率图模型的推断可以分为精确推断和近似推断。精确推断虽然能够计算出期望变量的条件分布的准确值，但是其计算复杂度随着变量数目的增加而极大增加。因此在实践中，更常用近似推断近似推断大致可以分为两类，采样和变分推断。本节将主要关注变分推断的实现过程。

变分推断的基本方法是用一个已知的相对简单形式的概率分布来近似需要推断的复杂分布。这个近似分布一般具有良好的特性，以简化运算过程。例如用简单的分布 $q(z;\lambda)$ 来拟合一个较复杂的分布 $p(z|x)$ 的过程，该过程的优化目标就是：

$$\lambda^* = \arg\min_{\lambda} \ divergence(p(z|x), q(z;\lambda))$$

上式收敛之后，就可以直接用 $q(z;\lambda)$ 来代替 $p(z|x)$。

接下来，作为将要提出的 DS-CTR 模型的几个基本单元以及以上基本内容的应用，将详细介绍几种基本的概率图模型以及其简单的推导过程。

第二节　概率矩阵分解

矩阵分解（Matrix Factorization，MF）指的是将一个矩阵分解为多个矩阵乘积。常见的矩阵分解有三角分解、QR 分解、约旦分解、SVD 分解等。在一般的

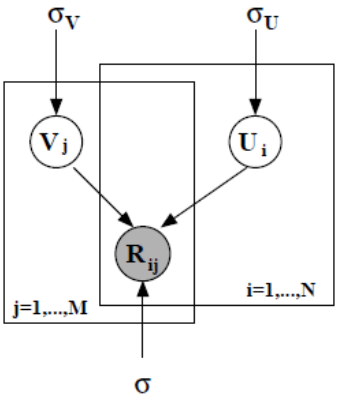

图 3.2　概率矩阵分解示意图。

推荐系统中，矩阵分解的过程就是讲评分矩阵 $R \in \mathbb{R}^{n \times m}$ 分解为用户特征矩阵 $U = U_1, U_2, \cdots, U_n$ 与项目特征矩阵 $V = V_1, V_2, \cdots, V_m$，优化目标就是使得预测值接近真实值，即：

$$min \sum_{i=1}^{n} \sum_{j=1}^{m} (R_{ij} - U_i^T I_j)^2$$

其中 R_{ij} 指的是原始用户 i 为商品 j 的评分，U_i 和 V_j 分别是用户 i 和商品 j 的特征向量，二者的乘积就是根据矩阵分解结果预测得到的评分数值，这个过程采用梯度下降的方式迭代计算 U 和 V，当二者收敛时就可以得到最终分解出来的矩阵结果。

在传统矩阵分解的基础上，进一步引入概率论知识进行优化，概率矩阵分解的基本思想是[34]，假设用户 U 和项目 V 的特征矩阵都服从高斯分布，首先通过原始的评分矩阵得到 U 和 V 的特征矩阵，之后再利用特征矩阵预测评分矩阵中的未知评分（见图3.2）。即用户、项目特征矩阵以及评分矩阵都分别满足以下分布：

$$\begin{aligned}
&U_i \sim \mathcal{N}(0, \sigma_U^2 \mathbf{I}), p(U|\sigma_U^2) = \prod_{i=1}^{N} N(U_i|0, \sigma_U^2 I) \\
&V_j \sim \mathcal{N}(0, \sigma_V^2 \mathbf{I}), p(V|\sigma_V^2) = \prod_{i=1}^{N} N(V_i|0, \sigma_V^2 I) \\
&R_{ij} \sim \mathcal{N}(U_i^T V_j, \sigma^2) \\
&p(R|U, V, \sigma^2) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left[\mathcal{N}(R_{ij}|U_i^T V_j, \sigma^2)\right]^{I_{ij}}
\end{aligned} \tag{3.1}$$

其中 $N(x|u, \sigma^2)$ 表示变量 x 满足均值为 u，方差为 σ^2 的高斯分布；I_{ij} 是指示函

数，表明如果用户 i 对于商品 j 有过评分，则 $I_{ij}=1$，否则为 0。进一步的，可以写出用户、项目的后验概率分布：

$$\begin{aligned}
p(U,V|R,\sigma^2,\sigma_V^2,\sigma_U^2) &= p(R|U,V,\sigma^2,\sigma_V^2,\sigma_U^2)\times p(U,V)/p(R,\sigma^2,\sigma_V^2,\sigma_U^2)\\
&\sim p(R|U,V,\sigma^2,\sigma_V^2,\sigma_U^2)\times p(U,V)\\
&= p(R|U,V,\sigma^2,\sigma_V^2,\sigma_U^2)\times p(U)\times p(V)\\
&= \prod_{i=1}^{N}\prod_{j=1}^{M}\left[\mathcal{N}(R_{ij}|U_i^TV_j,\sigma^2)\right]^{I_{ij}}\times\prod_{i=1}^{M}\mathcal{N}(U_i|0,\sigma_U^2\mathbf{I})\times\prod_{j=1}^{N}\mathcal{N}(V_j|0,\sigma_V^2\mathbf{I})
\end{aligned}\tag{3.2}$$

上面式子左右两端同取自然对数[①]，得到：

$$\begin{aligned}
\ln p(U,V|R,\sigma^2,\sigma_V^2,\sigma_U^2) = &\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}\ln\mathcal{N}(R_{ij}|U_i^TV_j,\sigma^2)+\\
&\sum_{i=1}^{N}\ln\mathcal{N}(U_i|0,\sigma_U^2\mathbf{I})+\sum_{j=1}^{M}\ln\mathcal{N}(V_j|0,\sigma_V^2\mathbf{I})
\end{aligned}\tag{3.3}$$

以其中 $\ln\mathcal{N}(U_i|0,\sigma_U^2\mathbf{I})$ 为例，为了求解这一项，首先给出用户 i 的概率密度函数，即：

$$\mathcal{N}(U_i|0,\sigma_U^2\mathbf{I}) = -\frac{1}{(2\pi)^{D/2}|\sigma_V^2\mathbf{I}|^{1/2}}\exp(-\frac{1}{2}U_i^T(\sigma_u^2\mathbf{I})^{-1}U_i)$$

最终两边同取自然对数，得到：

$$\begin{aligned}
\ln\mathcal{N}(U_i|0,\sigma_U^2\mathbf{I}) &= \ln(-\frac{1}{(2\pi)^{D/2}|\sigma_V^2\mathbf{I}|^{1/2}})-\frac{U_i^TU_i}{2\sigma_u^2}\\
&= -\ln(|\sigma_U^2\mathbf{I}|^{1/2})-\frac{U_i^TU_i}{2\sigma_u^2}+C_U\\
&= -\frac{1}{2}\ln(\sigma_U^{2D})-\frac{U_i^TU_i}{2\sigma_u^2}+C_U\\
&= -\frac{D}{2}\ln(\sigma_U^2)-\frac{U_i^TU_i}{2\sigma_u^2}+C_U
\end{aligned}\tag{3.4}$$

类似的，对其余两项进行计算，最后略去常数项，得到最终的目标函数：

$$E=\frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{M}I_{ij}(R_{ij}-U_i^TV_j)^2+\frac{\lambda_U}{2}\sum_{i=1}^{N}\|U_i\|^2+\frac{\lambda_V}{2}\sum_{i=1}^{M}\|V_i\|^2\tag{3.5}$$

容易观察得到，式（3.5 ）的第一项与传统矩阵分解目标函数一致，后两项是在贝叶斯学习过程中添加的正则项，因此可以有效的避免过拟合等问题。

[①]取对数不会改变原问题的凸凹性，同时不改变最优解得位置。

第三节　话题模型

一、隐狄利克雷模型的数学原理

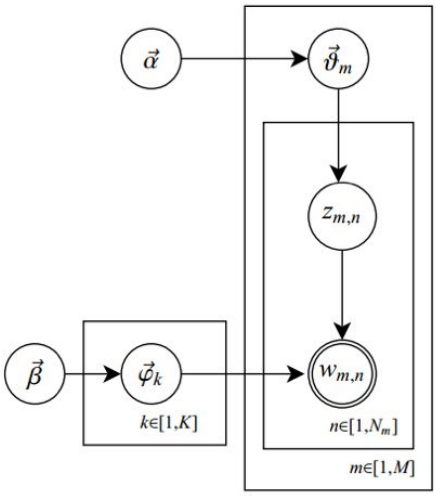

图 3.3　隐狄利克雷主题模型示意图。

话题模型（Topic Model）指的是一类有向图模型，这类模型主要用于处理离散数据，例如文本等。在自然语言处理、机器学习等领域都有非常广泛的应用。本节将重点介绍 Latent Dirichlet Allocation，即隐狄利克雷分配模型（LDA）。LDA 最早是由 Blei 等人于 2003 年正式提出[35]。在这一模型中（见图3.3），包含三个重要元素：词、文档和话题。词表征要处理的对象当中的离散单元，例如文章中的英文单词等；文档表征由一组词不计顺序而构成的集合，是需要处理的数据对象，例如论文等，这种表示方式也称作词袋模型。话题表征的是一系列相关的词，并包含每一个词所出现的概率值。例如，假设待处理的数据集（文档库）中包含 K 个主题和 M 篇文档，所有文档中的词都可以在包含 N 个词的词典。简便起见，将这些文档所构成的数据集表示为 M 个 N 维向量，即 $W = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_M\}$，$K$ 个 N 维向量表征 $\beta_k (k = 1, 2, \cdots, K)$ 表示话题，并且记第 m 篇文档 $w_m \in \mathbb{R}^N$ 当中的第 n 个分量的词频记为 $w_{m,n}$，类似的记第 k 个话题 $\Phi_k \in \mathbb{R}^N$ 的第 n 个词的词频为 $\phi_{k,n}$，最后记生成的文档 m 中词 n 的话题为 $z_{m,n}$。在 LDA 模型假设每篇文档不仅仅只包含一个话题，因此用向量 $\Theta_m \in \mathbb{R}^K$ 表示生成的第 m 篇文档中包含的话题比例，例如 $\theta_{m,k}$ 就表示文档 m 中包含主题 k 的比例。在以上各变量定义的情形下，LDA 实现文本建模的过程如下：

- $\vec{\alpha} \to \vec{\theta}_m \to z_{m,n}$：这一过程表征在生成文章 m 时，首先根据狄利克雷分布

随机采样选出这篇文章的主题分布（每个主题占的比例）θ_m，之后根据 θ_m 再进行话题指派，服从多项分布随机采样得到生成得到的文档中第 n 个词的 topic 编号 $z_{m,n}$

- $\vec{\beta} \to \vec{\phi_k} \to \omega_{m,n} \mid k = z_{m,n}$：这一过程表征在 $k = z_{m,n}$ 的条件下，考虑由第 k 个 topic 生成的词，首先由话题 $\vec{\beta}$ 根据狄利克雷分布随机采样得到在该话题下的词频分布 $\vec{\phi_k}$，之后再由 $\vec{\phi_k}$ 进行词的指派，即服从多项分布的采样得到文档 m 中的第 n 个词频 $\omega_{m,n}$

实际上，以上两个步骤不必要依次进行，例如常常先生成 N 个文档的主题，之后再基于前面得到的每个词的 topic 编号，生成每一个词，即：

$$\begin{aligned} p(\vec{w}, \vec{z} \mid \vec{\alpha}, \vec{\beta}) &= p(\vec{w} \mid \vec{z}, \vec{\beta}) p(\vec{z} \mid \vec{\alpha}) \\ &= \prod_{k=1}^{K} \frac{\Delta(\vec{\phi}_K + \vec{\beta})}{\Delta(\vec{\beta})} \prod_{m=1}^{M} \frac{\Delta(\vec{\theta}_m + \vec{\alpha})}{\vec{\alpha}} \end{aligned} \tag{3.6}$$

LDA 的核心部分包含两个 Dirichlet-Multiomial 共轭结构，根据共轭先验概率的定义（3.3 ），利用共轭结构的形式不变性质，能够在先验分布当中给予参数极其明确的物理含义，同时这个物理含义可以有效的延续到后续分布中并进行解释，从先验变换到后验过程中从数据中得到的补充知识也容易有物理解释。

定义 3.1 （**狄利克雷分布**） 满足狄利克雷分布的概率密度函数为：

$$f(x_1, x_2, ..., x_k; \alpha_1, \alpha_2, ..., \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} {x_i}^{\alpha^i - 1}$$

其中 $B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha^i)}{\Gamma(\sum_{i=1}^{k} \alpha^i)}, \sum_{i=1}^{k} x^i = 1$

定义 3.2 （**多项分布**） 多项分布，是简单的二项分布拓展到多维的情况. 多项分布指的是在一次实验中，随机变量的取值不再是二值的（即 0 或者 1）的，而是有多种离散取值可能 $1, 2, 3..., k$。多项分布的概率密度函数为：

$$P(x_1, x_2, ..., x_k; n, p_1, p_2, ..., p_k) = \frac{n!}{x_1!...x_k!} {p_1}^{x_1} ... {p_k}^{x_k}$$

定义 3.3 （**共轭先验分布**）贝叶斯概率理论当中指出，如果后验概率 $P(\theta|x)$ 和它的先验概率 $p(\theta)$ 具有相同的分布律，则将先验分布和后验分布称为共轭分布，与此同时，先验分布也被叫做似然函数的共轭先验分布。

二、LDA 参数估计

LDA 建模过程需要完成了两个任务，首先是依据给定的文档集合，确定超参数 α 和 β 的值，另一方面能够根据前面确认的超参数来确定新文档中的隐藏变量 θ 和 z 的分布。为了进一步明确各参数含义，生成了表（3.1）。对于给定文

表 3.1　LDA 模型参数

参数	含义
$\mathbf{W}$	所有文档构成的集合，唯一的已观测变量
M	总的文档数目
N	词典中词的总数
K	主题 Topic 的数目，一般人为设定
α	狄利克雷分布的参数
β	$K \times N$ 矩阵，每一行表示主题-词的分布，该分布服从狄利克雷分布
θ	文档中主题的分布，该分布服从狄利克雷分布
z	一篇文档中的某个主题，服从多项分布

档库 $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_M\}$ 进行建模的时候，需要对每一篇文章，根据已观测变量（词），来估计文章的主题分布。因此建模的结果应该使得原有的已观测变量以可能的最大概率出现。即寻找合适的参数 α^* 和 β^* 使得似然函数极大化，:

$$l(\alpha^*, \beta^*) = \sum_{d=1}^{M} logp(\mathbf{w}_m|\alpha, \beta) \tag{3.7}$$

为了解决参数之间的耦合关系以及未知变量，选择 EM 算法和变分推断来进行参数估计。首先需要求解隐含变量 θ 和 z 的期望:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

由于 $p(\mathbf{w}|\alpha, \beta)$ 难以计算，因此利用变分推断的方法，简化 LDA 模型如图 3.4，用分布 q 来近似原始模型，这里假设 θ 和 z 相互独立。则由上图可以得到二者的分布 q 为:

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\phi) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{3.8}$$

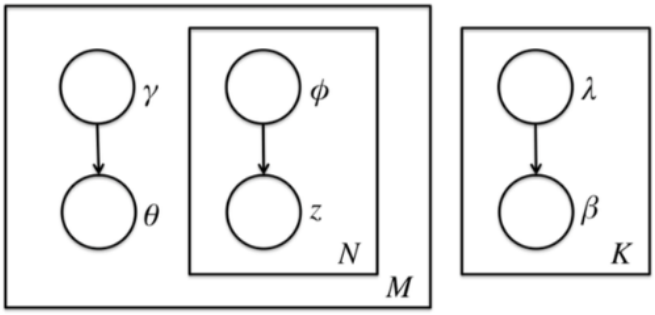

图 3.4　简化的 LDA 参数估计示意图。

其中，γ 是 Dirichlet 参数，多项式分布参数为 $(\phi_1, ..., \phi_N)$，为了达到推断的目的，需要分布 q 尽可能接近 p，因此优化的目标转化为，通过变分推断迭代算法求得合适的参数 γ^* 和 ϕ^* 使得：

$$(\gamma^*, \phi^*) = argmin_{(\gamma,\phi)} D(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)). \tag{3.9}$$

至此完成的 EM 算法中的 E 步算法，在接下来的 M 步，根据上述求得的 γ^* 和 ϕ^*，最大化 $L(\gamma, \phi|\alpha, \beta)$，以求得数 α 和 β。EM 算法结束，即可得到估计的 LDA 文本模型。

第四节　协同主题回归模型

本节将简要介绍 CTR 模型的实现原理（见图3.5）。

协同主题回归模型是 LDA 模型的扩展与应用，如图 3.5 所示，首先假设主题数目为 K，则主题的先验参数记为：$\beta = \beta_{1:K}$，在此基础上，设用户的正则化系数为 λ_u，推荐的项目正则化参数为 λ_v，该模型的生成过程可以描述为：

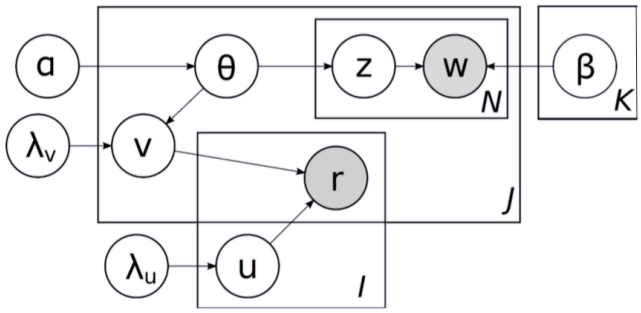

图 3.5　CTR 的图模型。

- 引入用户 i 的特征向量（隐变量）：$u_i \sim N(0, \lambda_u^{-1} I_K)$
- 对于每一个项目 j:

  (a) 主题的后验参数记为 $\theta_j \sim Dirichlet(\alpha)$

  (b) 设偏移量为 $\epsilon \sim N(0, \lambda_v I_K)$, 项目的隐变量 $v_j = \epsilon_j + \theta_j$

  (c) 利用 LDA 模型框架，生成主题 $z_{jn} \sim Mult(\theta_j)$，生成单词 $\omega_{jn} \sim Mult(Dirichlet(\beta_{z_{jn}})$
- 对于用户-项目对 i, j，评分值 $r_{ij} \sim N(u_i^T v_j, c_{ij}^{-1})$

类似于 LDA 算法的参数估计过程，在 CTR 模型中，使用 EM 算法和变分推断，估计参数的近似值。

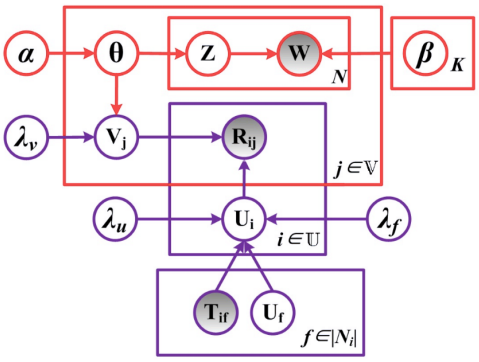

图 3.6　CTR-SMF2 的图模型结构。[36]

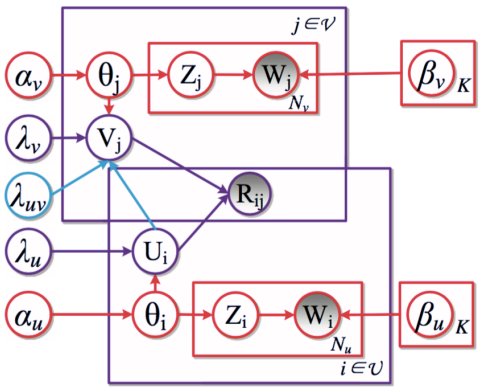

图 3.7　TRCF 的图模型结构。[22]

为了进一步提高 CTR 的性能，Wang 等人提出 CTR-SMF[37][11] 来进一步考虑用户之间的相似性关系，Chen 等人提出传统社交矩阵分解方法在所有对象之间分配一个单一的先验，而忽略了它们之间的不同，进而提出了 CTR-SMF2（见图3.6），它通过分配基于社交网络的每个对象的不同的先验来考虑相似的网络信息[36]，即引入：

$$min_{i=1}^{m} \sum_{j=1}^{m} T_{ij} \left\| U_i - U_j \right\|^2 \tag{3.10}$$

上式中，T_{ij} 指的是基于对用户有影响的上下文（contexts）信息（例如用户对某个所有评分的均值）而构建的用户相似性矩阵。更进一步，Chen 等人提出通过用户和商品的标签的相似性将用户与商品连接起来的方法，并提出 TRCF 模型（见图3.7）[22]。尽管 CTR 模型已经在诸多领域得到广泛应用，但是在药物敏感性推荐方面，仍然没有有效的将此方法进行运用，尤其在医学领域，药物和细胞的

相似性信息则很容易通过分析细胞基因表达特征和药物化学结构分析而得到[38]，下一节，将对 CTR 模型进行进一步的改进，并提出双层加强协同主题回归模型（Dual-layer Strengthened Collaborative Topic Regression, DS-CTR）。

第五节　DS-CTR 模型的建立

在前面，以常见的商品推荐系统为背景，详细的介绍了几种概率图模型，在本节将推荐系统中的用户-商品关系与药物-细胞系敏感性进行对比，将药物类比与商品推荐中的用户端，而将细胞系类比为商品推荐中的商品，在进行 LDA 建模过程中，药物和细胞系的基因特征将被视为“文本库”中的“词”，利用这些观测到的“词”，对细胞系和药物进行建模（如图3.8 所示）。此时，原问题可以描述为：假设目前有 n 种药物，记为 $U = \{u_1, u_2, \cdots, u_n\}$, m 个细胞系，记为 $V = \{v_1, v_2, \cdots, v_m\}$, p 种药物靶向基因 $P = \{p_1, p_2, \cdots, p_n\}$ 以及 q 细胞的癌变基因 $Q = \{q_1, q_2, \cdots, q_n\}$, 其中药物的特征矩阵 $U \in \mathbb{R}^{K \times n}$ 以及细胞系的特征矩阵 $V \in \mathbb{R}^{K \times m}$，它们的列向量 U_i 和 V_j 分别代表着 K-维药物 i 的特征向量和细胞

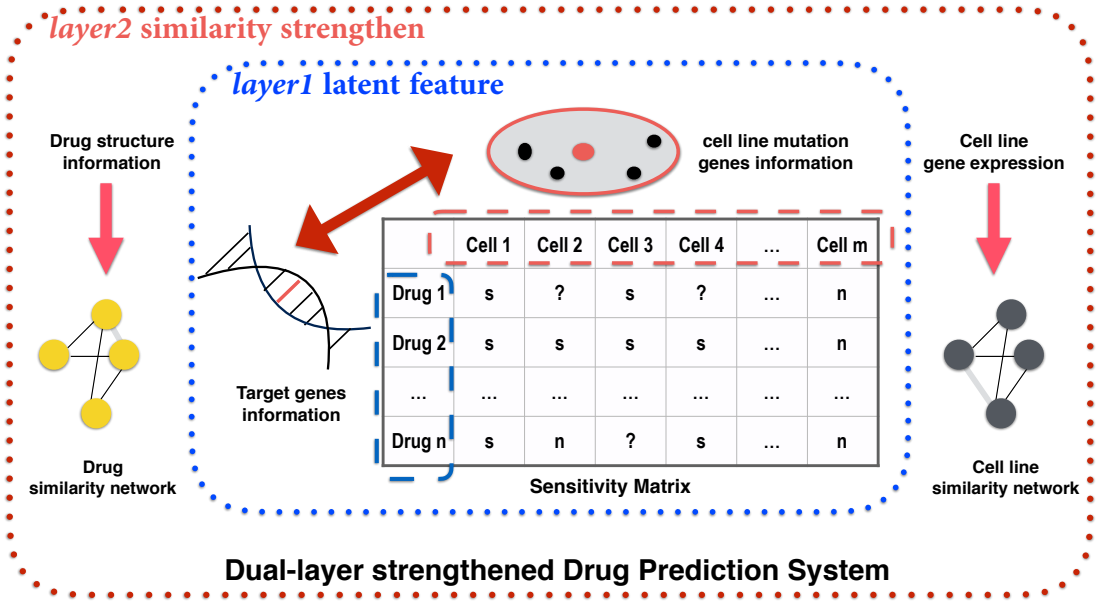

图 3.8　药物敏感性预测模型 DS-CTR 的计算框架示意图。在第一层，根据药物的靶向基因信息与细胞系的变异基因信息整合并捕获二者的关系；在第二层，利用药物结构信息和细胞系变异基因信息分别构建相似性网络，利用相似性网络强化推荐结果的相关性

系 j 的特征向量。令 $A \in \mathbb{R}^{n\times n}$ 和 $B \in \mathbb{R}^{m\times m}$ 分别代表药物和细胞系相似性网络。$R \in \mathbb{R}^{n\times m}$ 矩阵表征了数据集当中药物对细胞的敏感性矩阵，其中 R_{ij} 是以数值表示的药物 i 对于细胞系 j 的敏感性。定义置信参数 c_{ij} 表征 R_{ij} 的精准度，它的定义如下：

$$c_{ij} = \begin{cases} a & r_{ij} = 1 \\ b & r_{ij} = 0 \end{cases} \tag{3.11}$$

注意这里 a 和 b 两个参数满足 $a > b > 0$。如果 c_{ij} 更大，就表示更加信任 R_{ij} 所提供的信息。请注意当 $r_{ij} = 0$，此时可能存在两种情况：1）药物 i 对于细胞系 j 不敏感或者 2）该药物-细胞系对的药物敏感型实验并未进行，数值为空。综上所述，我们重点关注于充分利用药物和细胞系的基因组学数据和相似性信息，预测新的药物-细胞对的敏感性信息 (可见图 3.8)。

第六节　DS-CTR 模型的数学推导

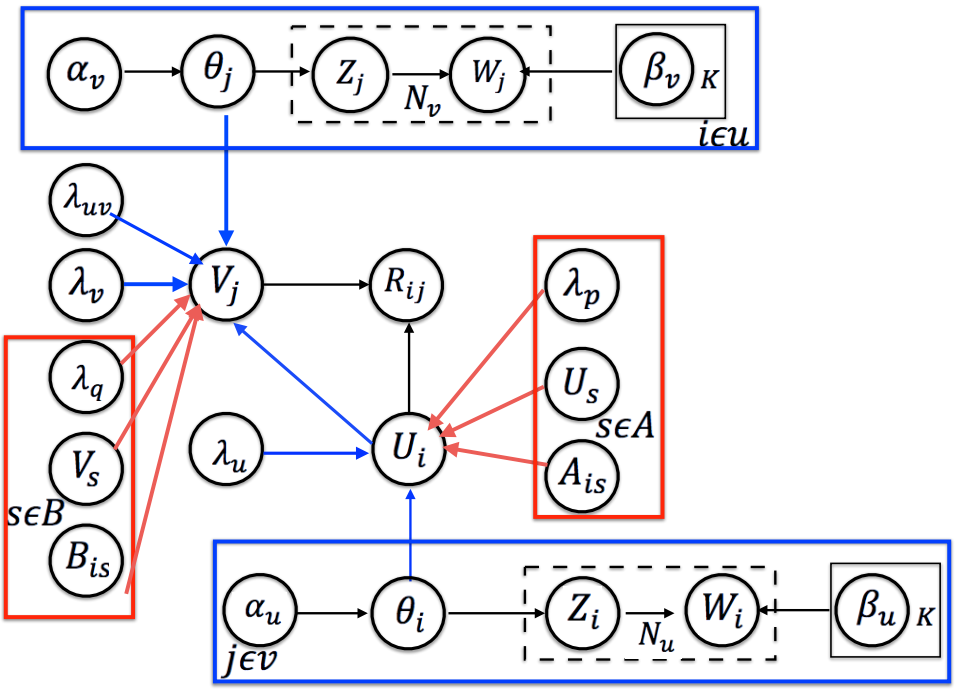

图 3.9　DS-CTR 图模型结构示意图。

基于前面对于图模型的讨论和分析，所提出的模型 DS-CTR 是一个分层的贝叶斯模型，该模型同时学习用户和项目的潜在特征空间。总体上，DS-CTR 模型分为两层：第一层是主题层，第二层是关系层。在第一层中，分别对每种药物和细胞进行基因组信息的分组，利用 LDA 分别对每种药物和细胞变异基因的语义信息进行挖掘（见图 3.9中的红色线）[22]。而在第二层中，使用双边矩阵分解来处理敏感性矩阵和相似性网络信息，以加强预测结果（见图 3.9中的蓝色线）。最后，将两层信息运用到矩阵分解的过程中，对灵敏度信息进行分解。之后通过设计 EM 算法，同时更新、优化损失函数。在这项研究中，使用两种不同的相似性来加强推荐结果。在第一层中，当用户和项目通过基因组信息和敏感性矩阵链接时，它们的潜在特征在某种程度上会显示出相似性[22]。同时基于结论：通过基因组变化显示的分子生物标志物具有来鉴别患者特质是有可能的[22]，使用在细胞系和药物两边同时使用 LDA 方法提取特征向量，并考虑相似药物对于细胞系作用的影响（图中从 U_i 指向 V_j 的箭头）。而在第二层中，强调了当药物与另一种药物具有密切关系（相似性值较高）时，它们的敏感性结果也应该彼此相似的，同样同种药物对于具有相似特征的变异细胞系也应该具有相近的敏感性。因此通过矩阵分解的方法将药物或细胞的相似性纳入到系统中以强化敏感性预测结果。传统的药物敏感型预测模型仅通过基因组信息来桥接药物和细胞，因此提出的 DS-CTR 可以比传统的药物反应预测模型更准确地捕获相关性。DS-CTR 的模型推导如下：假设用户和项目都有 k 个主题，则（LDA 部分的参数定义与前面章节保持一致）：

1. 挖掘药物的“潜”信息（semantic information）。

对于药物 u_i

(a) 得到药物类型所占比例（主题参数）

$$\theta_i \sim Dirichlet(\alpha_u);$$

(b) 得到药物的特征向量，$U \sim p(U)$，其中

第一层.

$$p(U) \propto N(\theta_i, \lambda_u^{-1} I_K)$$

第二层.

$$p(U) \propto \prod_{i=1}^{n} \prod_{s \in A_i} N(U_s, \lambda_p^{-1} A_{is}^{-1})$$

综合第一、二层，得到:

$$p(U) \propto N(\theta_i, \lambda_u^{-1} I_K) \prod_{i=1}^{n} \prod_{s \in A_i} N(U_s, \lambda_p^{-1} A_{is}^{-1}) \tag{3.12}$$

(c) 对于药物 u_i 的靶向基因 ω_{in_u}

i. 描述靶向基因的类型

$$z_{in_u} \sim \ Mult(\theta_i)$$

ii. 最终生成的靶向基因为:

$$\omega_{in_u} \sim \ Mult(\beta_{z_{in_u}});$$

2. 挖掘细胞的“潜”信息，并捕获到药物和细胞之间隐含的关系信息对于细胞系 v_j (a) 得到细胞系类型所占比例（主题参数）

$$\theta_j \ \sim \ Dirichlet(\alpha_v);$$

(b) 得到细胞系特征向量，$V \ \sim \ p(V)$, 即:

第一层.

$$p(V) \propto N(\theta_i, \lambda_v^{-1} I_K)$$

$$p(V) \propto \prod_i I_{ij}^R N(U_i, \lambda_{uv}^{-1} I_K)$$

第二层.

$$p(V) \propto \prod_{i=1}^{n} \prod_{s \in B_i} N(V_s, \lambda_q^{-1} B_{is}^{-1})$$

综合一、二层的结果

$$p(V) \propto N(\theta_i, \lambda_v^{-1} I_K) \prod_i I_{ij}^R N(U_i, \lambda_{uv}^{-1} I_K) \prod_{i=1}^{n} \prod_{s \in B_i} N(V_s, \lambda_q^{-1} B_{is}^{-1}) \tag{3.13}$$

(c) 对于细胞系 v_j 的变异基因 ω_{jn_v}:

i. 描述变异基因的类型

$$z_{jn_v} \sim \ Mult(\theta_j);$$

ii. 最终得到变异基因

$$\omega_{jn_v} \sim\ Mult(\beta_{z_{jn_v}});$$

3. 得到敏感性（数值）.

对于药物-细胞系作用对 $(i,\ j)$,

预测的敏感性（数值）结果:

$$R_{ij}\ \sim\ N(U_i^T V_j,\ c_{ij}^{-1}). \tag{3.14}$$

在上述推导过程中，I_K 是一个 $K \times K$ 的指示矩阵（identity matrix）; I_{ij}^R 指的是一个指示符矩阵，如果 u_i 对于细胞 v_j 有抑制作用，它的值为 1，否则它的值为 0; $N(x|\mu, \sigma^2)$ 表示的是一个均值为 μ 并且方差为 σ^2 的高斯分布。矩阵 *A* 和矩阵 *B* 分别表示药物相似性网络和细胞系相似性网络，例如 B_{ij} 表示细胞系 v_i 和细胞系 v_j.

在 DS-CTR 模型中，利用参数 λ_u, λ_v 来平衡通过基因组信息和敏感性等信息提取出的药物和细胞系“潜”信息对于预测药物敏感性的贡献多少[16]。同样的，利用参数 λ_p, λ_q 来平衡药物相似性网络和细胞相似性网络对于模型性能的影响。参数 λ_{uv} 则平衡了药物和细胞系内在关系对模型性能的影响大小。

为了使用贝叶斯推理方法，首先根据获得的药物和细胞的概率分布:

$$p(U|\lambda_u,\ \lambda_p, U_s, A) \propto p(U|\lambda_u) \times \prod_{i=1}^{n} \prod_{s \in A_i} p(U|U_s,\ A_{is}^{-1}\lambda_p\ )$$

$$\begin{aligned} &p(V|\lambda_v,\ \lambda_{uv}, \lambda_q, U, V_s, B\ ) \\ &\propto p(V|\lambda_v) \times \prod_i I_{\mathrm{ij}}^R p(U|U_i, \lambda_{uv}^{-1}) \times \prod_{i=1}^{n} \prod_{s \in B_i} p(V|V_s, \lambda_q^{-1} B_{is}^{-1}) \end{aligned}$$

与此同时，观测到的敏感性信息的条件概率可以描述为:

$$p\left(R|U,\ V, C\right) = \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\mathrm{N}\ \left(R_{ij}\middle|U_i^T V_j, \sigma_R^2\right)\right]^{I_{ij}^R}$$

最后，使用贝叶斯推理方法，给定灵敏度矩阵，双侧相似性网络和双侧基因组信息的特征向量，可以得到下面的后验概率的表达式:

$$\begin{aligned} &p\left(U, V|R,\ A,\ B, C,\ \lambda_u,\ \lambda_v,\ \lambda_p,\ \lambda_q, \lambda_{\mathrm{uv}}\right) \\ &\propto p(R|U, V, C) \times p(U|\lambda_u,\ \lambda_p, U_s, A) \times\ p(V|\lambda_v,\ \lambda_{\mathrm{uv}}, \lambda_q, U, V_s, B\ ) \end{aligned} \tag{3.15}$$

第七节 DS-CTR 模型参数学习

基于前面几个章节的内容，对于给定的参数和信息矩阵来直接计算药物和细胞特征向量 U_i 和 V_j 的完全后验是困难的。在这种情况下，可以利用坐标上升算法来学习最大后验概率（MAP）[21][22][36]。首先固定超参数的值，然后最大化两特征向量的后验概率，这相当于在给定 λ_u，λ_v，λ_p，λ_q 和 λ_{uv} 的情形下，使 U，V，A，B，$\theta_{1:n}$，$\theta_{1:m}$ 和 R 相关的似然函数最大化，即:

$$\begin{aligned} L = & -\frac{\lambda_u}{2}\sum_i (U_i-\theta_i)^T(U_i-\theta_i) - \frac{\lambda_v}{2}\sum_j (V_j-\theta_j)^T(V_j-\theta_j) \\ & -\frac{\lambda_{uv}}{2}\sum_{ij}(U_i-V_j)^T(U_i-V_j) - \frac{\lambda_p}{2}\sum_{s\in A} A_{is}\|U_i-U_s\|_F^2 \\ & -\frac{\lambda_q}{2}\sum_{r\in B} B_{ir}\|V_i-V_r\|_F^2 - \sum_{ij}\frac{c_{ij}}{2}\left(R_{ij}-U_i^TV_j\right)^2 \\ & +\sum_i\sum_{n_u}\log\left(\sum_k \theta_{ik}\beta_{k,\,\omega_{in_u}}\right) + \sum_j\sum_{n_v}\log\left(\sum_k \theta_{jk}\beta_{k,\,\omega_{jn_v}}\right) \end{aligned} \tag{3.16}$$

上式的前三项表征 DS-CTR 模型的第一层，接下来的两项表征第二层。在这里将狄利克雷先验设置为 1。固定超参数值，使用梯度下降的方法来迭代优化矩阵分解变量 U_i 和 V_j，以及主题后验参数 θ_i 和 θ_j。为了完成这样的工作，首先需要在给定的 θ_i 和 θ_j 估计值的情形下更新 U_i 和 V_j。对3.16中关于 U_i 和 V_j 的项分别求导并让等式右边等于 0，最终计算得到:

$$\begin{aligned} U_i \leftarrow & \left(\lambda_u I_K + VC_iV^T + \lambda_p\sum_j A_{ij}I_K + \lambda_{uv}\sum_j I_{ij}^R I_K\right)^T \\ & (\lambda_u\theta_i + VC_iR_i + \lambda_p\sum_j A_{ij}U_j + \lambda_{uv}\sum_j I_{ij}^R V_j) \end{aligned} \tag{3.17}$$

$$\begin{aligned} V_j \leftarrow & \left(\lambda_v I_K + UC_iU^T + \lambda_q\sum_i B_{ji}I_K + \lambda_{uv}\sum_j I_{ij}^R I_K\right)^T \\ & (\lambda_v\theta_j + UC_jR_j + \lambda_q\sum_i B_{ji}V_i + \lambda_{uv}\sum_i I_{ij}^R U_i) \end{aligned} \tag{3.18}$$

其中 C_i 是一个对角矩阵，并且它对角线上的元素为 $c_{ij}(j=1,\cdots,m)$，并且上式中的 $R_i=(r_{ij})_{j=1}^m$ 是药物 *i* 的评分向量。对于细胞系 *j*, C_j, R_j 有类似的定义。上式表明了参数 λ_u, λ_v, λ_p, λ_q 和 λ_{uv} 对于药物和细胞系特征向量都具有一定的

Input: 药物敏感性矩阵 **R**, 药物靶向基因集合 T_1, 细胞系变异基因集合 T_2, 药物相似性网络 A, 细胞系相似性网络 B, 正则化参数 λ_u, λ_v, λ_p, $\lambda_q$and λ_{uv}

Output: 药物和细胞系特征矩阵 U 和 V

药物和细胞的靶向或变异基因的"主题"分布 θ_u 进而 θ_v, 药物和细胞的靶向或变异基因分布 β_u 和 β_v;

repeat

  初始化 U, V

  更新药物特征矩阵 U

  更新细胞系特征矩阵 V

  更新超参数 θ_u 和 θ_v

  更新超参数 β_u 和 β_v

  计算似然函数

until 收敛或者超过迭代次数超过设定值;

算法 3.1: DS-CTR 模型的学习算法

影响。得益于来自细胞系和药物两边的信息，可以得到关于 U_i 和 V_j 对称的更新等式（3.17 和 3.18）。在的模型中, 使用五个参数来平衡不同信息源对于模型整体预测效果的影响。例如，当参数 λ_p 值更大时，药物相似性网络在整个模型中的影响更大。特别的，当 $\lambda_p = \lambda_q = 0$, 的 DS-CTR 模型将退化为 TRCF 模型[16], 更近一步，如果 $\lambda_{uv} = 0$，的模型将退化为 CTR-SMF2 模型 [40]。当所有参数都被设置为 0，则 DS-CTR 模型将退化为矩阵分解。

在同时的到药物和细胞系的特征矩阵 U 和 V 之后，需要进一步学习得到主题后验参数 θ_i 和 θ_j。在这里选择和 TRCF 模型一样的方法[16]。对于 θ_i，首先定义 $q\left(z_{in_u} = k\right) = \Phi_{in_u k}$，接着将等式3.16 中关于药物的项分离出来，并应用琴生不等式[21]，得到:

$$\begin{aligned} L\left(\theta_i\right) &\geqslant -\frac{\lambda_u}{2}\left(U_i - \theta_i\right)^T\left(U_i - \theta_i\right) + \sum_{n_u}\sum_{k}\Phi_{in_u k}\left(\log\theta_{ik}\beta_{k,\omega_{in_u}} - \log\Phi_{in_u k}\right) \\ &= L(\theta_i, \phi_i) \end{aligned} \tag{3.19}$$

注意其中 $\Phi_i = (\Phi_{in_uk})_{n_u=1,k=1}^{N_u \times K}$, N_u 是药物 i 所包含的靶向基因的数目, 同时 Φ_{in_uk} 满足 $\Phi_{in_uk} \propto \theta_{ik}\beta_{k,\omega_{in_u}}$。接下来使用梯度投影（projection gradient[16] ）的方法来优化 θ_i。对于 θ_j，也可以用同样的方法。

另一方面，利用在文献[16] 中的方法来更新 β:

$$\beta_{kw_i} \propto \sum_i \sum_{n_u} \Phi_{inu} kI[\omega_{in_u} = \omega]$$

在优化得到所有的项，包括 $U^*, V^*, \theta_{1:n}{}^*, \theta^*_{1:m}, \beta^*_u$ 和 β^*_v，最终可以得到预测的敏感性（数值）:

$$R^*_{ij} \approx (U^*_i)^T V^*_j.$$

为了更直观的表现上述过程，DS-CTR 参数的学习过程在算法3.1 中进行了总结。

第四章　DS-CTR 模型的验证

上一章系统的介绍为了解决传统药物敏感性预测方法的不足而新提出的双层协同主题回归模型（dual-layer strengthened collaborative topic regression，DS-CTR）。为了适应药物敏感性预测问题中多视角信息，DS-CTR 模型对原有 CTR 模型进行了改进，并与传统的主题模型中的概念进行对照；接着详细推导了 DS-CTR 模型的数学原理；本章将详细介绍了测试用数据集以及数据的预处理工作；第二、三节展示了在数据集上的测试结果和论文验证工作。

第一节　算法测试

在本节，详细介绍了算法的测试过程，主要包括：测试用数据集的介绍和预处理；CTR 类模型的对比；DS-CTR 与传统药物推荐系统方法的对比；还包括引用相关论文验证模型的实际预测效果最终得到实验结论。这一节，所有的处理结果都利用 MATLAB（MathWorks Inc., Natick, MA）以及配备有 Linux 操作系统的计算机上完成（Intel(R)Core(R) CPU 1.4 GHz）。

一、数据集介绍

在的实验中，选择癌症基因与药物敏感性数据集（GDSC，2016 年 7 月版本）。为了在癌症细胞中发现治疗性生物标志物和并开展进一步的临床前验证，GDSC 数据集提供了大量的药物应答信息和相关的基因组数据集以识别假定的治疗生物标志物[16]。1001 个癌细胞系和 265 个抗癌药物被包括在这个版本中，更详细的数据集介绍可在[32] 中获得。药物反应结果记录为 $\log IC_{50}$ 值的形式[14]。组织类型（Tissur type）也作为细胞系的注释信息在数据集中提供。药物则将其主要的治疗靶点在数据集中予以注释[13]。此外，细胞系的基因表达数据也在 Iorio 等人的工作中提供。在 GDSC 数据集中，研究者基于以下原则生成敏感性矩阵[39][32]：对于药物 $\boldsymbol{i}$ 和细胞系 $\boldsymbol{j}$，定义敏感性（数值）r_{ij}

$$r_{ij} = \begin{cases} 1 & if\ drug_i\ is\ sensitive\ to\ cell_j \\ 0 & else \end{cases} \tag{4.1}$$

除了从 GDSC 数据集获得药物-细胞系敏感性信息以及基因特征（靶向基因

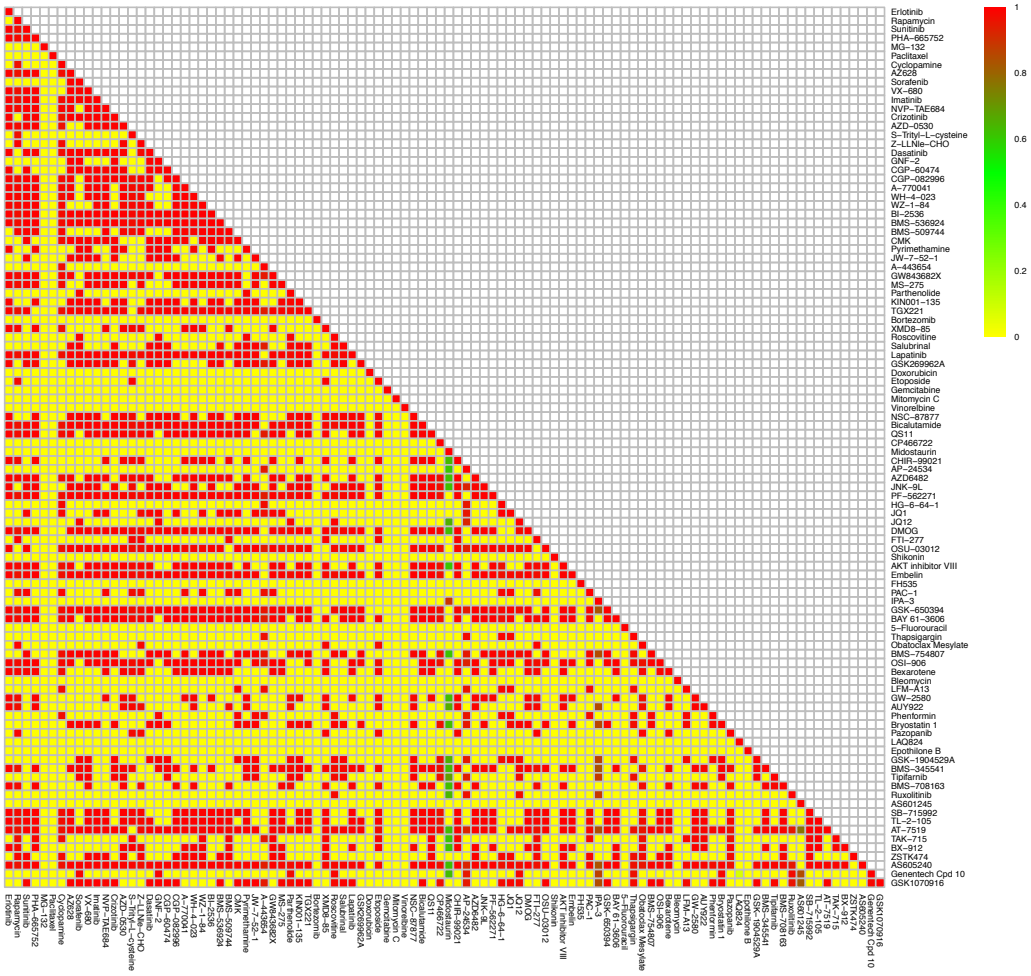

图 4.1　药物相似性网络（这里只展示 100 个药物相似性）。

或变异基因）信息，接着利用 PhbChem[32] 系统下载关于药物化学结构的信息。至此，可以利用 Zhang 等人在 [25] 中的方法，利用药物机构信息生成药物相似性网络以及利用细胞系的基因表达信息构建细胞系的相似性网络。为了使相似性网络以及敏感性信息可视化，在图4.2和图4.1分别绘制出细胞相似性和药物相似性网络的热图（heatmap），以及在图4.3中展示了 GDSC 中提供的药物-细胞系敏感性数据的可视化热图。汇总以上信息，可以得到本实验所用到的数据概览（见表 4.1）。

在本实验中，为了评估所提出模型的表现，选择使用 5 折交叉验证的方法（5-fold cross-validation）。进一步的，将每一个数据集分成三部分，其中的 80% 用作训练集，10% 用作测试集，剩余的 10% 用作验证数据集。将利用训练集训练

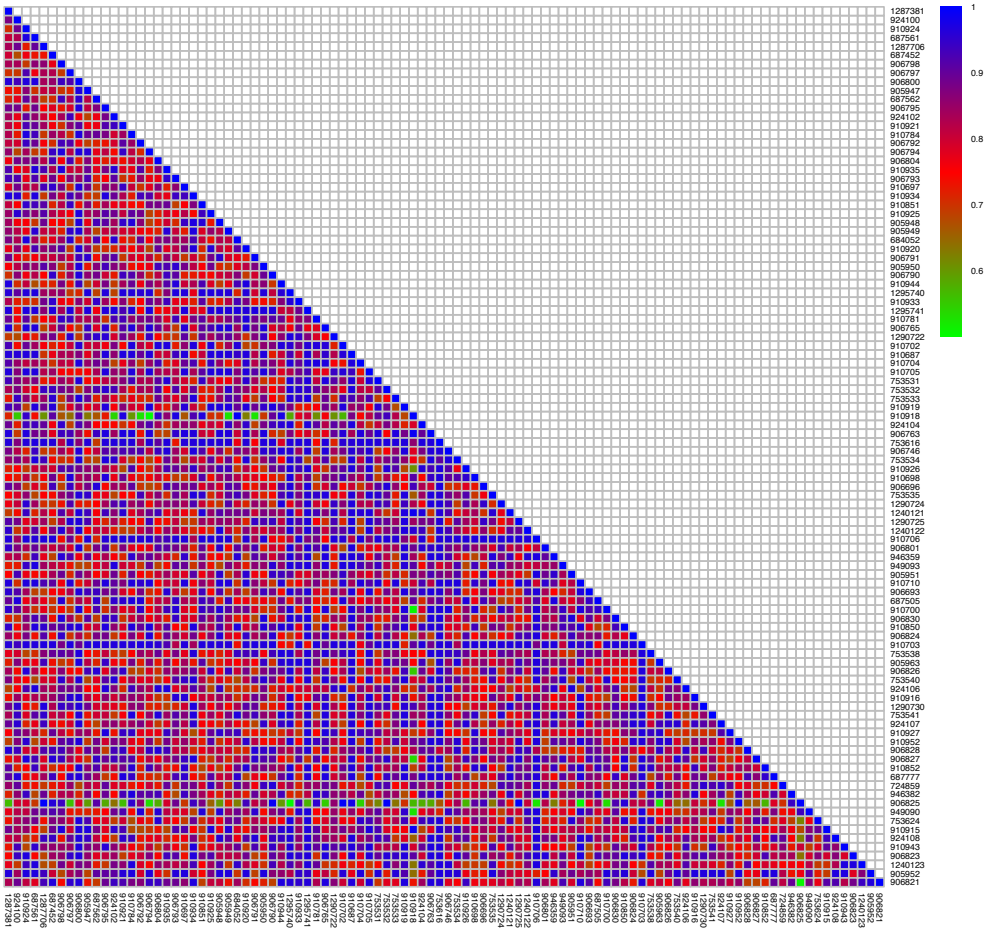

图 4.2　细胞系相似性网络（这里只展示 100 个细胞系相似性）。

出模型，之后利用验证机优化模型参数，最后利用测试集测试模型性能。

表 4.1　本实验用数据概览（单位：个）

数据源	药物	靶向基因	药物相似性	数据源	细胞系	变异基因	细胞相似性
药物	265	125571	65536	细胞系	1001	127035	1002001

二、评估指标（Evaluation Metric）

DS-CTR 模型用于预测未知的药物-细胞系敏感性关系，因此按照类似任务中常用的评估指标[11]，选择使用 ROC 和 AUC 来评估模型性能。ROC 图常用在

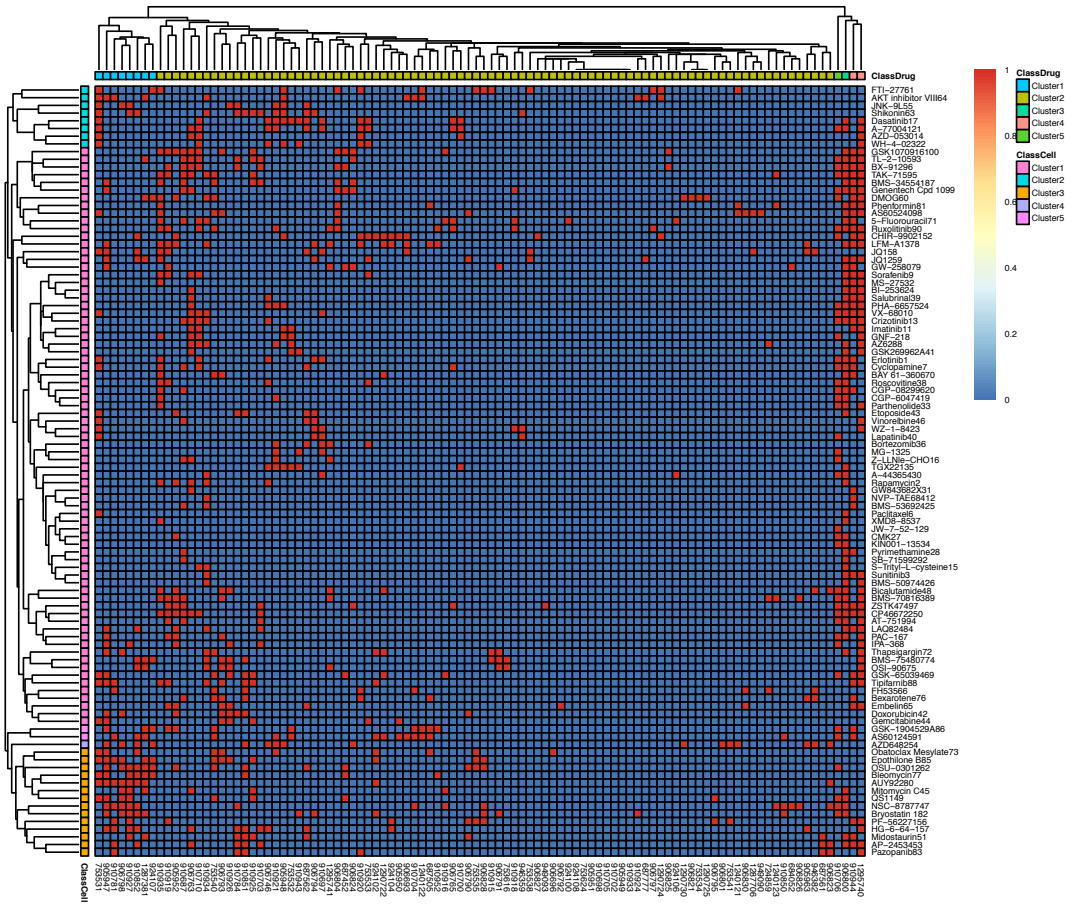

图 4.3　GDSC 数据集以 IC50 值形式提供的药物-细胞系敏感性的可视化热图（部分）。

医疗决策（medical decision）领域，并且近年来也在机器学习和数据挖掘领域有了广泛的应用[40]。受试者工作特征曲线（ROC）常常用于可视化分类器的性能，与 Precision-Recall 曲线（PRC）相比，在类别分布中，ROC 对变化不敏感，例如，即使测试集中的正面到负面事件的比例发生变化，ROC 也不会改变，相反，对于 PRC 而言，则会发生很大的变化[16]。在这种情况下，设置两个度量来评估模型性能。ROC 的 x 轴和 y 轴分别是特异性（specificity）和灵敏度（sensitivity），分别用 FN（False Negative）表示预测为负样本而实际上是正样本的情形；FP（False Positive）表示预测为正样本然而实际上是负样本的情形；TN（True Negative）表示预测为负样本并且实际上也是负样本的情形以及 TP（True Positive）表示预测值和实际值均为正样本的情形，在此基础上，ROC 的 x 轴和 y 轴分别可以定义

如下：

$$\begin{aligned} \text{specificity} &= \frac{\text{FP}}{FP+TN} \\ \text{sensitivity} &= \frac{\text{TP}}{FN+TP} \end{aligned} \tag{4.2}$$

第二节　数据集测试结果

一、CTR 模型对比

由于提出的模型的核心结构是基于协作过滤的回归，因此在本小节中，按照模型所使用的来自药物或细胞的不同类型的信息，将所提出的 DS-CTR 模型与以下三种退化推荐（Collapsed Model）模型进行比较。同样利用 GDSC 数据集，当只考虑细胞系基因组信息时，应用最原始的 CTR 模型（在这里标注为 R+G）[21]。与此同时，通过利用 CTR 的改进版本 [46]，添加了一方的相似性信息（药物相似性），并且在每种药物之前分配了不同的先验信息，CTR 模型与社会矩阵分解构成 CTR-SMF2 模型，在这里标记为（R + G + D）以表征此模型结合了敏感性矩阵、基因组信息和药物相似性信息[36]。此外，可以通过基于标签和评分的协同过滤模型[22] 来获得双侧（细胞系和药物）基因组信息和它们之间潜语义的关系（详见第三章第 3 节的介绍），标记为（R + (GD)）以显示该模型考虑药物-细胞系之间潜在关联的模型特征。最后提出的模型 DS-CTR 模型被标记为（R + (GD) + C）以强调添加了额外的细胞系相似性网络。

另一方面，在比较过程中，使用网格搜索的方法为所有四个模型的超参数找到最佳起始值。对于基于主题的模型（CTR，CTR-SMF2，TRCF），将主题数 K 武断的设置为 50，以使用 LDA 生成主题分布。对于原始 CTR 模型，将置信参数设置为 a=1 and b=0.01，同时设 $\lambda_u = 0.01$，$\lambda_v = 10$。对于 CTR-SMF2 模型，将置信系数设置为 a=1 和 b=0.1，另外设置 $\lambda_u = 1$，$\lambda_v = 1$，$\lambda_f = 1$。对于 TRCF 模型，将置信系数设置为 a=1 和 b=0.1，与此同时设置 $\lambda_u = 10$，$\lambda_v = 1$，$\lambda_{uv} = 0.01$。对于所提出的 DS-CTR 模型，将置信系数设置为 a=1 和 b=0.1，同时设 $\lambda_u = 10$，$\lambda_v = 1, \lambda_{uv} = 0.0001$，$\lambda_p = \lambda_q = 0.001$。

图4.4给出了通过 5 折交叉验证 ROC 和 AUC 比较的结果，通过这些结果能够得到以下结论：1）考虑额外信息（相似性，基因组）的推荐模型，从 AUC 和 ROC 评测指标来看，显著提高了预测性能。特别是基于相似性网络的协同滤波

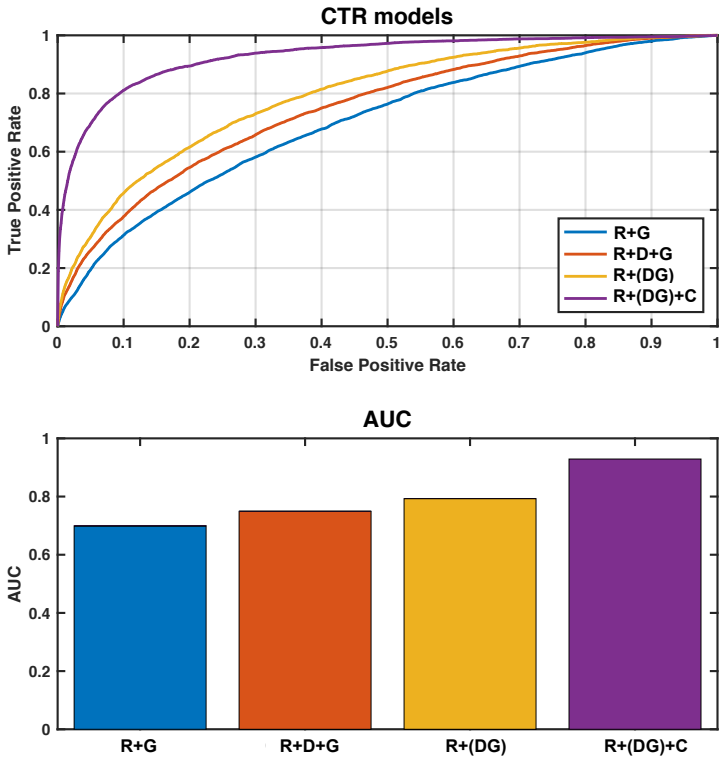

图 4.4　协同过滤模型的 AUC 与 ROC 指标的对比。其中的缩写：R+G，结合药物敏感性信息和基因组信息的模型，即 CTR 模型；R+D+G，在以上模型的基础上结合药物的结构相似性信息（即 CTR-SMF2 模型）；R+(DG)，考虑药物靶向基因和细胞系变异基因之间的“潜”关联，即 TRCF 模型；R+(DG)+C，本文提出的 DS-CTR 模型，在上述模型的基础上，考虑细胞系的相似性以强化预测结果的相似性。该图中上方的图是 ROC 的比较，下方的图是 AUC 值的比较。

算法（例如 R + D + G）和基于双侧基因组信息的协同滤波算法（例如 R +（DG））具有比传统协同滤波主题回归模型（CTR）更好的性能，并且它们的 AUC 值分别比 CTR 高 7% 和 15%。2）同时考虑利用来自细胞和药物的双侧相似性网络和基因组信息的，所提出的方法 DS-CTR 具有最佳性能。与 R +（DG）和 R + G 相比，提出的方法在 GDSC 数据集上可以实现 AUC 值 24% 和 33% 的提升。3）根据 ROC 的研究结果，在相同的真阳性率下, 提出的模型与其他三种方法相比，具有最低的假阳性率，这表明的模型能够提供最准确的预测结果。从图 4 可以更清楚地看出，固定特异性值为 0.1，然后比较相应的灵敏度值。结果表明，提出的方法提供了更高效的实现预测结果中的正面例子排名高于负面例子[40]。

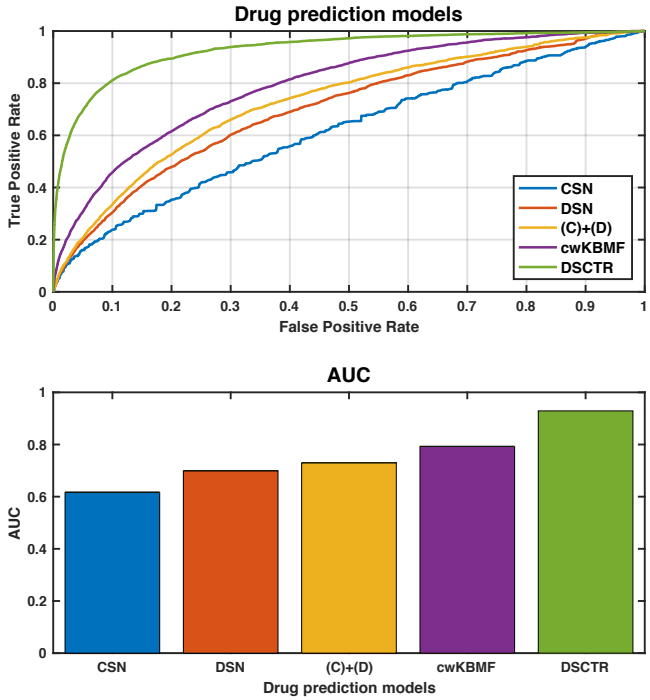

图 4.5　本文提出的模型 DS-CTR 与其余几种常用的药物敏感性预测间的 AUC 以及 ROC 指标的对比。其中的缩写：CSN，细胞相似性网络；DSN，药物相似性网络；C+D，细胞相似性网络与药物相似性网络的加权组合模型；cwKBMF，核化贝叶斯矩阵分解的方法；DS-CTR：本文提出的双层强化协同主题回归模型模型。该图中上方的图是 ROC 的比较，下方的图是 AUC 值的比较。

二、经典药物推荐模型对比

正如在第二章的介绍中提到的，研究人员已经提出了多种药物敏感性预测方法[9,23–27][16]。在本小节中，提出的模型将和第二章中涉及到的四种先进方法进行比较，所有四种最先进的分析方法都只考虑基因组信息。在 Muhammad Ammad-ud-din 等人的工作中[14]，为了完成预测任务，研究人员提出了核化的贝叶斯矩阵分解方法（cwKBMF）。该方法整合了来自药物和细胞的基因组信息，用于预测药物反应，并在侧面数据视图中编码样本之间的相似性网络。Zhang 等人的工作中强调了包括药物相似网络（DSN）和细胞系相似网络（CSN）在内的相似性网络在药物敏感性预测方面的重要性[10]。正如在第二章的最后一节中提到的那样，该模型需要分别构建 CSN 和 DSN 网络，最终通过加权的方式（C+D）整合为一个预测模型进行最终的预测任务。在 CSN，DSN，C + D 和 cwKBMF 模型

中，用 90% 的数据用于训练，10％用于测试，而对于 DS-CTR，按照前面提到的那样分割数据集。DS-CTR 模型与以上四种模型（CSN，DSN，C＋D，cwKBMF）的对比结果可以在图五中看到。所有上述方法都符合默认配置（满足原论文的配置要求）。

首先根据 ROC 曲线来分析五折交叉验证的结果。显然，当曲线靠近左边界和上边界时，表明预测系统会出现更高的灵敏度和特异性。特别是从图4.5中，当将特异性固定为 0.1 时，提出的方法在所有方法中具有最高的灵敏度。DS-CTR 方法明显优于其他方法（AUC = 0.9289）。具体而言，当比较 CSN，DSN 和 C + D 方法时，基于 DSN 的预测比单独 CSN 更好（17%）。此外，二者的加权组合即更多的信息整合相较于 CSN（20％）或 DSN（5％）都更加优越，这再次表明多视角信息对预测的巨大帮助[41]。同时，考虑药物和细胞基因组相似性的 cwKBMF 表现优于 Zhang 等人的方法（5%）。就 AUC 指标而言，提出的方法分别比上述四种方法分别高 49%，31% 和 26%，这表示当组合多视角信息时所提出的基于 CF 的模型的优越性。

表 4.2　预测的药物-细胞作用结果的前 20 位

排序	药物	细胞系编号	敏感性	排序	药物	细胞系编号	敏感性
1	CX-5461	908158	1	11	KIN001-260	909702	1
2	QL-XI-92	1330960	1	12	Nutlin-3a	908156	1
3	Tubastatin A	909715	1	13	Tubastatin A	1330982	1
4	Tubastatin A	906856	1	14	KIN001-260	908158	1
5	QL-XI-92	908156	1	15	Tubastatin A	906800	1
6	CX-5461	909715	1	16	Tubastatin A	1331034	1
7	Y-39983	949155	1	17	Y-39983	909715	1
8	PIK-93	909715	1	18	CX-5461	1331033	1
9	QL-XI-92	1331033	1	19	QL-XI-92	1331037	1
10	KIN001-260	909715	1	20	CX-5461	909702	1

表 4.3　前 10 位新预测得到的药物敏感性结果

序号	药物	细胞系	序号	药物	细胞系
1	CX-5461	1327774	6	AC220	906856
2	CH5424802	908156	7	CX-5461	688026
3	Tubastatin A	907275	8	CX-5461	908457
4	TG101348	1331037	9	KIN001-260	907275
5	GSK1070916	910706	10	CX-5461	906862

第三节　论文验证

通过利用 DS-CTR 模型，首先生成预测结果并将其与由 GDSC 数据集提供的已知敏感性关系进行比较。表4.2 给出了前 20 位的预测结果。本表中的灵敏度列显示了数据集提供的实验结果。第三列的索引值表示由 GDSC 项目定义的特定细胞系。有关索引的更多信息，请参阅[39] 和[32]。可以清楚地看到，所有前 20 位预测结果都与原始数据集一致。同时，附录中也提供了前 200 名的预测结果（表A.1）。它们都表明了模型的优越性能。

本文所提出的 DS-CTR 模型也能够有效的发现新型药物敏感性关系。为了评估的模型预测结果，还生成了未包含在 GDSC 数据集中的前 10 个新型预测药物-细胞系作用对，然后检索相关的研究证据或数据库。表4.3 总结了的预测结果，在此表中，如果 GDSC 数据集中没有给出这一对的敏感性，但已经通过最新的科学研究观察到，将用粗体标记。例如，在的实验中，药物 AC220 和细胞系 EoL-1-细胞之间的敏感性并未在 GDSC 数据集中给出，而根据的推荐结果，它们可能会具有一定的敏感性关系。Reiter 等人在 2017 年 9 月在 Nature 上发表的研究表明，研究人员注意到，FLT3-D835Y 或 FLT3-ITD 蛋白保留在核周 ER 中，在加入 AC220 后，研究者能够观察到与 FLT3-WT（EOL-1 包含在其中）或 FLT3-N67 6K 相似的细胞膜定位。对于这部分实验的详细介绍，请参考文献[42]。这一证据部分支持了 AC220 和 EOL-1 细胞之间的相互作用。在补充材料（表A.2）中包含了所有前五十个新的预测对，这些预测可能会对实际的药物-细胞系敏感性实验起

到帮助。综上所述，所提出的 DS-CTR 模型系统的预测出细胞-药物敏感性作用对，并能够为实际医学试验提供必要的生物学预测。所提出的方法 DS-CTR 结合两层贝叶斯网络以增强结果的准确性，与临床前试验相比，利用所提出 DS-CTR 结合两层以加强结果的准确性能够使新型药物 - 细胞系作用对的开发更快也更便宜，这也正是精准医疗所追求的个性化医疗的一部分。

第五章　结论

精准医疗或者个性化医疗作为新型医疗模式和医学概念，相比于传统的治疗方案，具有明显的优势，也因此得到了越来越多研究者的关注。本文专注于解决个性化医疗当中的药物敏感性预测问题，即通过分析药物和癌变细胞系的相关信息，对于未知的药物-细胞系敏感性关系进行预测，以期为临床试验提供更多的依据并降低新药的研发成本。本文首先介绍了问题的背景知识，之后详细介绍了传统的药物敏感性预测方法以及商品推荐系统中常用的图模型的背景知识，最后利用这些背景知识，引出所构建的 DS-CTR 药物敏感性预测模型，并利用 GDSC 数据集对该模型进行全面的测试。本章将总结本文的研究贡献，之后将关注与本模型中可能存在的问题，进而为未来的研究方向提供参考。

第一节　全文总结

开发或者发现具有特定药理学性质的新分子特性是一种昂贵并且且损耗率高的过程。例如，考虑到实体的毒性或者药物的有效性，会出现许多 II 期或 III 期临床试验失败的临床案例，这也导致新药的研发成本大大增加[41]。基于即使单一药物也可以作用于多个靶点以及多靶点多途径作用与细胞等的事实，研究者对药物重新定位产生了极其浓厚的兴趣，同时这也将大大降低开发新药的成本[6][19]。此外，精确医学的目标是通过参照癌症患者的特定分子谱来选择合适的治疗方法进而实现有效的个体化治疗[43–45]。个性化治疗最终导致对系统性预测技术的研究，该预测系统能够通过分析现有药物和细胞系信息来预测未知的敏感性作用对并以此来促进药物的开发过程，从而为进一步的临床实验提供一定的依据。然而，癌症的发生和发展涉及到多种分子机制，而且来自同一患者的癌细胞，甚至来自不同患者的肿瘤的多样性都使得该问题非常复杂[46]。近年来，肿瘤的分子特性受到研究者的广泛关注，研究者在大规模高通量筛选工作中取得了显著的进展。这些研究揭示了人类癌症细胞系和药物的敏感性关系[6–8]，为研究者的进一步研究提供了依据。此外通过分析某些相关的人体细胞系和相应药物的药理谱，研究者系统地解决了预测生物标志物识别的问题，如癌症细胞系百科全书（CCLE）[8] 和癌症基因组计划（CGP）[9]。因此，基于这些研究，建立能

够精确地将癌症基因组信息转化为肿瘤生物学和治疗知识的癌症细胞应答的高精度预测系统[11] 将变得更容易和更可靠。

基于现有药物和细胞系基因组信息，研究者已经做了大量的工作来预测药物敏感性。尤其是在筛选抗癌药物和肿瘤细胞系方面取得了非凡的进展，为建立具有异质性来源数据的综合预测和整体方法提供了良好的契机，同时也促进了对药物重定位的研究[10] 和个性化医疗的研究。然而，在预测模型中的两个关键问题极大地挑战构建系统的难度。从建模的角度来看，药物预测模型有必要将多视角信息结合在同一个模型中[19]。同时，不同于普通的商品推荐系统，药物敏感性预测系统需要更高的有效性和准确性[47]。本文提出了一种新的基于协同滤波的推荐框架 DS-CTR 来解决这一问题，并弥补了以往在主题回归模型当中所存在的不足。结合我们所参考的文献资料，本文是首先提出利用双边（药物和细胞系）基因组信息和相似性网络来预测药物敏感性网络的。

第二节　未来研究展望

尽管我们在本实验中取得了比较理想的预测结果，但也应当注意所提出的方法的局限性，未来的研究展望主要集中在以下几个方面：

1）计算复杂度问题。DS-CTR 模型的一个显著特征是其学习过程。本模型所利用的梯度下降算法会随着数据量的扩大极大增加计算负荷[47] 并使得运算过程耗时较长，这直接导致了本模型在应用于大规模数据集时可能会遇到一定的困难。DS-CTR 模型是以计算复杂度为代价而提供了更高精度的预测结果。随着生物医学领域中信息量的不断积累和多视角信息的日益增多，快速算法是实际应用中迫切需要解决的问题。

2）模型使用的数据问题。在我们的模型中，本文选择药物的结构信息和细胞系的基因表达来衡量其相似性，并作为额外的强化信息加入模型中。然而实际的临床试验也为研究者提供了类似于 DNA 拷贝数和致癌基因突变、ATC 编码等多样化信息。尽管在本实验中，利用 DS-CTR 已经取得了高精度的预测结果，但利用其余类型的信息进行预测并达到或许更满意的预测效果，仍然值得研究者进一步研究。

3）模型参数问题。在 DS-CTR 中，结合 LDA 的方法来挖掘药物与药物、细

胞系与细胞系之间的关联。首先需要假定主题的数目，在本文，认为假定它的值为 50，并在此假设下取得良好的预测效果，本文没有讨论主题数目对模型性能的影响；与此同时，考虑到数据集包含信息的完整性，本文只选用了 GDSC 数据集进行训练与测试，未来应当结合更新的数据集进行测试，已验证模型的完备属性。

参 考 文 献

[1] 张会鲜, 何琪杨. 基于精准医学的抗肿瘤靶向药物敏感性预测及其研发应用. Chinese Journal of New Drugs, 2015, 24(16).

[2] Cohen R L, Settleman J. From cancer genomics to precision oncology—tissue's still an issue. *Cell*, 2014, 157(7): 1509–1514.

[3] 项亮. 推荐系统实践. 北京: 人民邮电出版社, 2012: 39–44.

[4] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. *Computer*, 2009, 42(8).

[5] Rendle S, Freudenthaler C, Gantner Z, et al. Bpr: Bayesian personalized ranking from implicit feedback//Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. AUAI Press, 2009: 452–461.

[6] Barretina J, Caponigro G, Stransky N, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 2012, 483(7391): 603.

[7] Basu A, Bodycombe N E, Cheah J H, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, 2013, 154(5): 1151–1161.

[8] Garnett M J, Edelman E J, Heidorn S J, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 2012, 483(7391): 570.

[9] Zhang N, Wang H, Fang Y, et al. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS computational biology*, 2015, 11(9): e1004498.

[10] Geeleher P, Cox N J, Huang R S. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome biology*, 2014, 15(3): R47.

[11] Wang Y, Fang J, Chen S. Inferences of drug responses in cancer cells from cancer genomic features and compound chemical and therapeutic properties. *Scientific reports*, 2016, 6: 32679.

[12] Menden M P, Iorio F, Garnett M, et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one*, 2013, 8(4): e61318.

[13] Ammad-ud din M, Khan S A, Malani D, et al. Drug response prediction by inferring pathway-response associations with kernelized bayesian matrix factorization. *Bioinformatics*, 2016, 32 (17): i455–i463.

[14] Ammad-Ud-Din M, Georgii E, Gonen M, et al. Integrative and personalized qsar analysis in cancer by kernelized bayesian matrix factorization. *Journal of chemical information and modeling*, 2014, 54(8): 2347–2359.

[15] Cichonska A, Rousu J, Aittokallio T. Identification of drug candidates and repurposing opportunities through compound–target interaction networks. *Expert opinion on drug discovery*,

2015, 10(12): 1333–1345.

[16] Ammad-ud din M, Khan S A, Wennerberg K, et al. Systematic identification of feature combinations for predicting drug response with bayesian multi-view multi-task linear regression. *Bioinformatics*, 2017, 33(14): i359–i368.

[17] Berlow N, Haider S, Wan Q, et al. An integrated approach to anti-cancer drug sensitivity prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2014, 11(6): 995–1008.

[18] Sos M L, Michel K, Zander T, et al. Predicting drug susceptibility of non–small cell lung cancers based on genetic lesions. *The Journal of clinical investigation*, 2009, 119(6): 1727–1740.

[19] Cobanoglu M C, Liu C, Hu F, et al. Predicting drug–target interactions using probabilistic matrix factorization. *Journal of chemical information and modeling*, 2013, 53(12): 3399–3409.

[20] Iorio F, Knijnenburg T A, Vis D J, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 2016, 166(3): 740–754.

[21] Wang C, Blei D M. Collaborative topic modeling for recommending scientific articles// Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 448–456.

[22] Chen C, Zheng X, Wang Y, et al. Capturing semantic correlation for item recommendation in tagging systems.//AAAI. 2016: 108–114.

[23] De Niz C, Rahman R, Zhao X, et al. Algorithms for drug sensitivity prediction. *Algorithms*, 2016, 9(4): 77.

[24] Zou H, Zhang H H. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 2009, 37(4): 1733.

[25] Sokolov A, Carlin D E, Paull E O, et al. Pathway-based genomics prediction using generalized elastic net. *PLoS computational biology*, 2016, 12(3): e1004790.

[26] Breiman L. Random forests. *Machine learning*, 2001, 45(1): 5–32.

[27] Riddick G, Song H, Ahn S, et al. Predicting in vitro drug sensitivity using random forests. *Bioinformatics*, 2010, 27(2): 220–224.

[28] 周志华. 机器学习. Qing hua da xue chu ban she, 2016.

[29] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1977: 1–38.

[30] Birbil Ş İ, Fang S C, Sheu R L. On the convergence of a population-based global optimization algorithm. *Journal of global optimization*, 2004, 30(2-3): 301–318.

[31] Wu C J. On the convergence properties of the em algorithm. *The Annals of statistics*, 1983:

95–103.

[32] Wang Y, Xiao J, Suzek T O, et al. Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, 2009, 37(suppl_2): W623–W633.

[33] Jordan M I, Bishop C. An introduction to graphical models. *unpublished book*, 2001.

[34] Mnih A, Salakhutdinov R R. Probabilistic matrix factorization//Advances in neural information processing systems. 2008: 1257–1264.

[35] Blei D M. Probabilistic topic models. *Communications of the ACM*, 2012, 55(4): 77–84.

[36] Chen C, Zheng X, Wang Y, et al. Context-aware collaborative topic regression with social matrix factorization for recommender systems.//AAAI. 2014: 9–15.

[37] Purushotham S, Liu Y, Kuo C C J. Collaborative topic regression with social matrix factorization for recommendation systems. *arXiv preprint arXiv:1206.4684*, 2012.

[38] Ma H, Yang H, Lyu M R, et al. Sorec: social recommendation using probabilistic matrix factorization//Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 931–940.

[39] Yang W, Soares J, Greninger P, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 2012, 41(D1): D955–D961.

[40] Fawcett T. An introduction to roc analysis. *Pattern recognition letters*, 2006, 27(8): 861–874.

[41] van der Graaf P H, Benson N. Systems pharmacology: bridging systems biology and pharmacokinetics-pharmacodynamics (pkpd) in drug discovery and development. *Pharmaceutical research*, 2011, 28(7): 1460–1464.

[42] Reiter K, Polzer H, Krupka C, et al. Tyrosine kinase inhibition increases the cell surface localization of flt3-itd and enhances flt3-directed immunotherapy of acute myeloid leukemia. *Leukemia*, 2018, 32(2): 313.

[43] Garraway L A. Genomics-driven oncology: framework for an emerging paradigm. *Journal of Clinical Oncology*, 2013, 31(15): 1806–1814.

[44] Yuan H, Paskov I, Paskov H, et al. Multitask learning improves prediction of cancer drug sensitivity. *Scientific reports*, 2016, 6: 31619.

[45] Ezzat A, Zhao P, Wu M, et al. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2017, 14(3): 646–656.

[46] Allison K H, Sledge G W. Heterogeneity and cancer. *Oncology (Williston Park)*, 2014, 28(9): 772–778.

[47] Zhang L, Liu H, Huang Y, et al. Cancer progression prediction using gene interaction regularized elastic net. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*

(TCBB), 2017, 14(1): 145–154.

附录 A　补充表格

表 A.1　DS-CTR 模型利用 GDSC 数据集 (July, 2016 version)[39] [32] 预测得到的前 200 位细胞系-药物关系对。下表中的顺序指的是实际模型预测得到的结果顺序，排序靠前的关系对被认为敏感性关系更为明确。请注意这里的细胞系一列的数字编号代表着特定的细胞系。详细的信息请参考 GDSC 数据集的官方网站。在敏感性一列中的数字，如果为 1，则表明在 GDSC 数据集中，该细胞对该药物具有敏感性，如果值为 0，则表明关系对的敏感性未知或者在 GDSC 数据集中未涉及。

排序	药物	细胞系	敏感性	排序	药物	细胞系	敏感性
1	CX-5461	908158	1	101	QL-XI-92	1331034	1
2	QL-XI-92	1330960	1	102	CX-5461	713899	1
3	Tubastatin A	909715	1	103	Tubastatin A	909252	1
4	Tubastatin A	906856	1	104	PIK-93	1323913	1
5	QL-XI-92	908156	1	105	Y-39983	1330947	1
6	CX-5461	909715	1	106	XMD15-27	908156	1
7	Y-39983	949155	1	107	QL-XI-92	753548	1
8	PIK-93	909715	1	108	CX-5461	910944	1
9	QL-XI-92	1331033	1	109	QL-XI-92	910947	1
10	KIN001-260	909715	1	110	Tubastatin A	684059	1
11	KIN001-260	909702	1	111	TPCA-1	909252	1
12	Nutlin-3a	908156	1	112	Trametinib	1240149	1
13	Tubastatin A	1330982	1	113	Tubastatin A	1330947	1
14	KIN001-260	908158	1	114	STF-62247	908156	1
15	Tubastatin A	906800	1	115	JW-7-24-1	1327774	1
16	Tubastatin A	1331034	1	116	Y-39983	908156	1
17	Y-39983	909715	1	117	GSK690693	949163	1
18	CX-5461	1331033	1	118	QL-XI-92	1331035	1
19	QL-XI-92	1331037	1	119	CX-5461	908156	1
20	CX-5461	909702	1	120	ABT-263	906856	1
21	CX-5461	1331036	1	121	QL-XI-92	684059	1
22	KIN001-260	1330984	1	122	CX-5461	949158	1
23	KIN001-260	1295740	1	123	XMD14-99	910706	1
24	GSK690693	1331036	1	124	CX-5461	1331037	1
25	Y-39983	905965	1	125	GSK690693	909701	1
26	CX-5461	1240145	1	126	Tubastatin A	908156	1
27	Y-39983	905952	1	127	KIN001-260	910944	1
28	KIN001-260	910906	1	128	Y-39983	905958	1
29	CX-5461	949155	1	129	GSK690693	909703	1
30	Tubastatin A	1323913	1	130	BX-912	910944	1

31	Y-39983	1295740	1	131	Y-39983	1327769	1
32	CX-5461	905958	1	132	GSK690693	683665	1
33	GSK690693	1331033	1	133	GSK690693	908448	1
34	Tubastatin A	910947	1	134	QL-XI-92	910706	1
35	GSK690693	684059	1	135	GSK690693	906870	1
36	CX-5461	1330984	1	136	XMD14-99	909715	1
37	PIK-93	908134	1	137	Y-39983	684059	1
38	CX-5461	684059	1	138	Tubastatin A	1330984	1
39	CX-5461	1331034	1	139	GSK1070916	908156	1
40	ABT-263	908156	1	140	CX-5461	1247871	1
41	KIN001-260	1330947	1	141	GSK690693	1331034	1
42	Tubastatin A	910944	1	142	KIN001-260	910947	1
43	CX-5461	753610	1	143	KIN001-260	908156	1
44	GSK690693	910944	1	144	Tubastatin A	907272	1
45	Nutlin-3a	1247871	1	145	KIN001-260	1297446	1
46	Y-39983	1331036	1	146	KIN001-260	906856	1
47	GSK1070916	1295740	1	147	GSK690693	906846	1
48	CX-5461	908146	1	148	TPCA-1	906800	1
49	QL-XI-92	1330947	1	149	Y-39983	909252	1
50	QL-XI-92	908146	1	150	Trametinib	905974	1
51	STF-62247	1331037	1	151	QL-XI-92	906846	1
52	GSK690693	684057	1	152	CX-5461	1330960	1
53	TG101348	908156	1	153	PIK-93	909252	1
54	Tubastatin A	908158	1	154	A-770041	906800	1
55	KIN001-260	907272	1	155	KIN001-260	907277	1
56	CX-5461	907073	1	156	Ruxolitinib	910706	1
57	Trametinib	909713	1	157	CX-5461	909251	1
58	A-770041	1331037	1	158	GSK1070916	1247871	1
59	Trametinib	924238	1	159	PIK-93	908158	1
60	TG101348	1330947	1	160	Tubastatin A	1297446	1
61	QL-XI-92	908158	1	161	Tubastatin A	1295740	1
62	KIN001-260	1323913	1	162	Tubastatin A	1331037	1
63	CX-5461	1295740	1	163	CX-5461	910546	1
64	CX-5461	1299080	1	164	GSK1070916	909715	1
65	Y-39983	906800	1	165	Tubastatin A	909702	1
66	Trametinib	906855	1	166	AP-24534	1330960	1
67	QL-XI-92	1331036	1	167	Tubastatin A	683665	1
68	GSK1070916	684059	1	168	BIX02189	906870	1
69	A-770041	1330960	1	169	KIN001-260	908134	1
70	PIK-93	1330984	1	170	BIX02189	909715	1
71	QL-XI-92	1295740	1	171	QL-XI-92	909251	1
72	JW-7-24-1	908158	1	172	QL-XI-92	909260	1
73	Y-39983	1330984	1	173	GSK690693	909715	1
74	GSK690693	907275	1	174	PIK-93	909703	1
75	QL-XI-92	1330933	1	175	XMD14-99	1323913	1

76	KIN001-236	909715	1	176	QL-XI-92	906836	1
77	Y-39983	906856	1	177	QL-XI-92	1330985	1
78	Y-39983	908158	1	178	Trametinib	753545	1
79	TPCA-1	909715	1	179	GSK690693	907789	1
80	QL-XI-92	907799	1	180	KIN001-260	909252	1
81	Trametinib	909756	1	181	CX-5461	907320	1
82	GSK690693	1327774	1	182	VNLG/124	909715	1
83	KIN001-260	906800	1	183	Ruxolitinib	1331037	1
84	QL-XI-92	1330950	1	184	BX-912	908156	1
85	CX-5461	909703	1	185	GSK690693	1331040	1
86	T0901317	908156	1	186	QL-XI-92	908448	1
87	GSK690693	906824	1	187	QL-XI-92	907275	1
88	CX-5461	909701	1	188	QL-XI-92	1323913	1
89	KIN001-260	906870	1	189	Tubastatin A	905958	1
90	QL-XI-92	909715	1	190	QL-XI-92	909255	1
91	GSK1070916	909252	1	191	TPCA-1	1330984	1
92	Trametinib	907061	1	192	GSK690693	1330947	1
93	GSK690693	908158	1	193	GSK690693	910688	1
94	CX-5461	906800	1	194	XMD14-99	1331037	1
95	CX-5461	753548	1	195	Nutlin-3a	905952	1
96	TPCA-1	1330947	1	196	QL-XI-92	906800	1
97	QL-XI-92	906870	1	197	CX-5461	949165	1
98	CX-5461	924247	1	198	Y-39983	1331037	1
99	QL-XI-92	910906	1	199	CX-5461	909256	1
100	XMD14-99	908158	1	200	Ruxolitinib	908156	1

表 A.2　DS-CTR 模型利用 GDSC 数据集 (July, 2016 version)[39] [32] 预测得到未知的前 50 位细胞系-药物关系对（所列出的关系对在 GDSC 数据集中未给出明确关系或者细胞对药物不敏感，因此在敏感性一栏中，所有的值都为 0）。下表中的顺序指的是实际模型预测得到的结果顺序，排序靠前的关系对被认为敏感性关系更为明确。

排序	药物	细胞系	敏感性	排序	药物	细胞系	敏感性
1	CX-5461	1327774	0	26	GSK1070916	1327774	0
2	CH5424802	908156	0	27	Tubastatin A	907783	0
3	Tubastatin A	907275	0	28	Trametinib	908440	0
4	TG101348	1331037	0	29	OSI-930	906856	0
5	GSK1070916	910706	0	30	BX-912	949155	0
6	AC220	906856	0	31	VNLG/124	909703	0
7	CX-5461	688026	0	32	Bleomycin (50 uM)	1298215	0
8	CX-5461	908457	0	33	FMK	1330983	0
9	KIN001-260	907275	0	34	KIN001-102	910906	0

10	CX-5461	906862	0	35	T0901317	909252	0
11	CX-5461	1524419	0	36	VNLG/124	1295740	0
12	GSK1070916	1330960	0	37	Trametinib	905956	0
13	STF-62247	907275	0	38	CP466722	909715	0
14	Belinostat	908156	0	39	AV-951	906856	0
15	CAL-101	1327774	0	40	Zibotentan	910944	0
16	NPK76-II-72-1	910906	0	41	Y-39983	1327774	0
17	Zibotentan	1327774	0	42	GDC0941	907275	0
18	CX-5461	949178	0	43	KIN001-102	1330984	0
19	Tubastatin A	1327774	0	44	XMD13-2	907789	0
20	Y-39983	907275	0	45	AV-951	908158	0
21	AZD6244	906820	0	46	Trametinib	753614	0
22	GSK1070916	949158	0	47	Ruxolitinib	910906	0
23	CX-5461	1240130	0	48	Belinostat	1323913	0
24	KIN001-260	1330960	0	49	VNLG/124	909255	0
25	CX-5461	908133	0	50	AV-951	907272	0

在读期间发表的学术论文与取得的研究成果

已投稿论文

1. **Hang Wang**, Jianing Xi, Minghui Wang, Ao Li, Dual-layer Strengthened Collaborative Topic Regression Modeling for Predicting Drug Sensitivity, Under Review, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*(TCBB).

2. **Hang Wang**, Mohamed Irfan Mohamed Refai, B. J. F. van Beijnnum, One Inertial Sensor Based Upper Extremity Usage Measurement and Standard, Submitted, *IEEE Transaction on Neural Systems and Rehabilitation Engineering*(TNSRE).