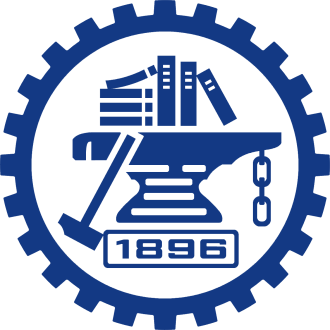
申请上海交通大学硕士学位[论文](#_OFDM技术简介_1)

基于病例数据的智能诊断支持系统

****

**学校：** 上海交通大学

**院系：** 电子信息与电气工程学院

**班级：** B1403392

**学号：** 1140339075

**硕士生：**张棪

**专业：** 计算机科学与技术

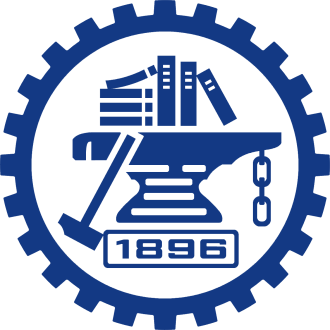
**导师：** 曹健

**上海交通大学电子信息与电气工程学院**

**2017年1月**

**A Dissertation Submitted to Shanghai Jiao Tong University for the Degree of Master**

A Case-Based Intelligent Diagnosis Support System

****

**Author:** Yan Zhang

**Specialty:** Computer Science

**Advisor:** Jian Cao

School of Electronic Information and Electrical Engineering

Shanghai Jiao Tong University

Shanghai, P.R.China

January, 2016

基于病例数据的智能诊断支持系统

摘 要

随着医学的不断发展，各专业分化越来越细，专家们专注于自己所学的专门学科，在深刻钻研专业知识的同时势必造成知识面的缩小。在恶性肿瘤等复杂的疾病治疗过程中，为了使患者能获得更完善更全面的医疗服务，医院等医疗机构以患者为中心，联合与疾病相关的各个学科的医生和专家，针对同一个患者进行讨论、沟通和协商，共同为患者制定完善的治疗方案。近年来，这种多学科综合的治疗方式已经成为了临床治疗领域的一个新的趋势。

这种理念和方式不仅为患者提供了更优质的医疗服务，同时也促进了不同学科的医生之间的交流，对医学的发展有着重要的积极影响。然而与庞大的医疗需求相比，医疗资源十分紧缺，多学科综合治疗的推广需要大量人力物力提供支持。此外，由于缺乏标准的电子化的系统对多学科诊疗的过程进行记录和分析，对一个患者的诊疗过程所包含的知识和信息，并不能为将来的诊疗过程提供帮助，这也是提高诊疗效率的一个很大的瓶颈。

针对这些问题，本文尝试将机器学习技术应用到医疗领域，在前期建立的高质量乳腺癌病例数据库基础上开发了一个能高效准确地推荐治疗方案的智能诊断支持系统，并投入了实际使用。此外，本文还针对现实数据中存在的概念漂移现象，提出了适用于概念漂移的推荐算法，进行了相关实验和分析。

关键词：多学科、决策支持系统、基于案例推理、概念漂移

A CASE-BASED INTELLIGENT DIAGNOSIS SUPPORT SYSTEM

ABSTRACT

With the rapid development of medicine and the growing number of subfields, experts focus on their own specialized subfields for deep study of expertise, which also leads to knowledge narrowness at the same time. During the diagnosis of complicated diseases such as malignant tumors, hospitals and other medical institutions gather experts of various related disciplines for discussion, so as to jointly formulate the perfect treatment plans for patients. This is called multi-disciplinary treatment and has become a new trend in clinical treatment in recent years.

Multi-disciplinary treatment not only provide patients with better quality of medical services, but also promote the expertise communication of doctors among different disciplines, which has a positive impact on the development of medicine. However, multi-disciplinary treatment requires a lot of manpower and material resources, and due to the lack of standard electronic system for recording the inexplicit knowledge of the diagnosis, it is hard to provide help for future diagnosis of similar cases.

To solve these problems, this paper attempts to apply the machine learning technologies to the medical field. Based on a high-quality database of breast cancer cases developed before, we develop an intelligent diagnosis support system to recommend treatment plans accurately and efficiently. In addition, this paper proposes an adaptive recommendation algorithm, which can work in the context of concept drift.

KEY WORDS: Multi-Disciplinary, Decision Support System, Case-based Reasoning, Concept Drift

目 录

[摘 要 I](#_Toc471762271)

[ABSTRACT II](#_Toc471762272)

[第一章 绪论 1](#_Toc471762273)

[1.1 研究背景和意义 1](#_Toc471762274)

[1.2 国内外研究现状 2](#_Toc471762275)

[1.3 本文主要工作 4](#_Toc471762276)

[1.4 本文组织结构 5](#_Toc471762277)

[第二章 相关技术 7](#_Toc471762278)

[2.1 基于规则的推理 7](#_Toc471762279)

[2.1.1 基于规则推理的优点 8](#_Toc471762280)

[2.1.2 基于规则推理的缺点 8](#_Toc471762281)

[2.1.3 决策树算法 9](#_Toc471762282)

[2.2 基于案例的推理 11](#_Toc471762283)

[2.2.1 基于案例推理的优点 12](#_Toc471762284)

[2.2.2 基于案例推理的缺点 13](#_Toc471762285)

[2.2.3 k-NN算法 14](#_Toc471762286)

[2.3 本章小结 16](#_Toc471762287)

[第三章 智能诊断支持系统的架构 17](#_Toc471762288)

[3.1 引言 17](#_Toc471762289)

[3.2 工作流引擎 19](#_Toc471762290)

[3.2.1 基于ECA规则的工作流模型 19](#_Toc471762291)

[3.2.2 工作流引擎结构 20](#_Toc471762292)

[3.2.3 系统角色与工作流之间的交互 21](#_Toc471762293)

[3.3 病例数据库 22](#_Toc471762294)

[3.3.1 病例数据库介绍 22](#_Toc471762295)

[3.3.2 病例数据表结构 22](#_Toc471762296)

[3.3.3 病例数据库系统功能 27](#_Toc471762297)

[3.4 系统核心模块 28](#_Toc471762298)

[3.4.1 投票模块 29](#_Toc471762299)

[3.4.2 推荐模块 30](#_Toc471762300)

[3.5 本章小结 30](#_Toc471762301)

[第四章 智能诊断支持系统的推荐模块 31](#_Toc471762302)

[4.1 基于规则的推荐模块 31](#_Toc471762303)

[4.2 基于病例的推荐模块 31](#_Toc471762304)

[4.2.1 相关记号 31](#_Toc471762305)

[4.2.2 离散属性的相似度度量 32](#_Toc471762306)

[4.2.3 目标类别 34](#_Toc471762307)

[4.3 推荐结果的准确率 35](#_Toc471762308)

[4.3.1 准确率分析 35](#_Toc471762309)

[4.3.2 综合比较 37](#_Toc471762310)

[4.4 推荐结果融合 38](#_Toc471762311)

[4.4.1 推荐结果的粒度 39](#_Toc471762312)

[4.4.2 推荐结果之间的支持度 41](#_Toc471762313)

[4.4.3 推荐结果的综合评估 41](#_Toc471762314)

[4.5 本章小结 43](#_Toc471762315)

[第五章 适用于概念漂移的推荐算法 44](#_Toc471762316)

[5.1 引言 44](#_Toc471762317)

[5.2 概念漂移 44](#_Toc471762318)

[5.3 自适应分类器 45](#_Toc471762319)

[5.3.1 概念敏感的探测器 46](#_Toc471762320)

[5.3.2 冲突记录表 47](#_Toc471762321)

[5.3.3 概念漂移的检测和自适应 49](#_Toc471762322)

[5.4 系统实验 52](#_Toc471762323)

[5.4.1 衡量指标 52](#_Toc471762324)

[5.4.2 比较算法 52](#_Toc471762325)

[5.4.3 数据集 53](#_Toc471762326)

[5.4.4 结果分析 54](#_Toc471762327)

[5.4.5 参数分析 56](#_Toc471762328)

[5.5 本章小结 58](#_Toc471762329)

[第六章 智能诊断支持系统的具体实现 59](#_Toc471762330)

[6.1 相关工具和软件 59](#_Toc471762331)

[6.1.1 操作系统 59](#_Toc471762332)

[6.1.2 数据库 59](#_Toc471762333)

[6.1.3 Web框架 59](#_Toc471762334)

[6.1.4 消息队列中间件 60](#_Toc471762335)

[6.2 相关界面 60](#_Toc471762336)

[6.2.1 病例数据库的功能界面 60](#_Toc471762337)

[6.2.2 智能诊断支持系统的用户界面 62](#_Toc471762338)

[第七章 总结与展望 64](#_Toc471762339)

[7.1 主要工作与创新点 64](#_Toc471762340)

[7.2 后续研究工作 64](#_Toc471762341)

[7.2.1 分歧度预测 64](#_Toc471762342)

[7.2.2 用户行为分析 66](#_Toc471762343)

[参 考 文 献 67](#_Toc471762344)

[致 谢 71](#_Toc471762345)

[攻读硕士学位期间已发表或录用的论文 72](#_Toc471762346)

# 绪论

## 研究背景和意义

传统的医学诊断的决策方法中，医生大都使用医疗仪器对患者进行检查，根据检查的结果，结合患者的各种既往病史，凭借自身的医学知识和临床经验对病情进行分析判断，进而制定相应的治疗方案。大多数疾病的症状和问题都可以用生物学和医学上的术语准确地进行描述，医生利用他们在医学院学习获得的知识和技能来解决相关问题。这种决策方法很大程度上取决于主治医生的业务水平，主观因素影响较大，由于在一个特定的领域中积累了大量的经验，专家们因此成为了临床诊断的主要力量，经常长时间进行工作。与此同时，随着医学的不断发展，各专业分化越来越细，各科室的医生更倾向于专攻自己所研究的学科，在深刻钻研专业知识的同时势必造成了知识面的缩小，即便是某个领域的专家，在诊断的过程中也很难靠一人之力兼顾患者的各方面情况。

多学科方法（Multi-Disciplinary Approach）在基于对当前复杂情况有新的认知和理解的前提下，包括了适当地从多个学科获取额外的知识，发现和定义那些存在于一般情况以外的问题，最终找到相应的解决方案这些过程。在医疗保健领域中，这种方法被广泛使用 [1]。随着跨越不同学科，跨越不同医院的合作越来越频繁，医疗保健正在变得日益复杂。为了解决复杂的临床和护理需求，有条件的医院等医疗机构通常会组建一个多学科的团队（Multi-Disciplinary Team），以应对情况复杂的病例。

采用多学科团队进行医疗诊断的方式可以被称为多学科综合治疗（MDT，Multi-Disciplinary Treatment），是近年来临床治疗领域中一个新的趋势，尤其是在肿瘤治疗这样复杂的领域中，它更是一个不可缺少的理念。它的核心思想是以患者为中心，召集内科、外科、放疗科等多个学科的医生和专家，针对同一个患者进行讨论。各学科的专家发挥各自所长，互相沟通协商，共同为患者制定更完善的治疗方案。

然而就国内而言，医疗资源与庞大的医疗需求相比，显得十分紧缺。一般来说，开展多学科讨论的过程相对较为繁琐，需要各个学科的医生共同参与，对每个患者的情况进行多次讨论，兼顾各方面可能存在的问题之后，最终制定一个综合治疗方案。这一过程为了达到最佳的治疗效果，消耗大量的医疗资源，对诊疗的效率有很大的制约，而许多医院甚至并不具备实施多学科讨论的医疗资源和能力。这些原因都限制了患者接受多学科讨论的机会，可能导致患者的各方面情况无法被兼顾，从而使患者得不到最佳的治疗方案。

上述情况的主要问题在于稀缺的医疗资源和庞大的医疗需求之间的矛盾，由于在多学科讨论的活动中，对所有的接受讨论的病例，专家们都要完整地执行相应的流程，造成了一定的资源浪费。这就需要寻求一个既能保证诊疗效果又能提高诊疗效率的基于病例数据的智能诊断支持系统，对那些和以往讨论过的病例相似的，能快速得出讨论结果的患者，系统可以直接推荐最佳诊疗方案，从而将专家们从繁琐的重复工作中解放出来。

这样的系统具有学习能力，可以根据当前患者的各项指标，从数据库中检索历史病例以作为依据，进行治疗方案的推荐。对有能力进行多学科讨论的大医院而言，一个由系统推荐的治疗方案可以作为参考，为讨论的过程提供帮助。对医疗条件相对较差的医院，在共享病例数据的基础上，智能诊断支持系统可以基于国内其他领先的医院产生的历史病例数据，推荐诊疗方案及相关信息以供主治医生参考，降低疑难病症的误诊率。

## 国内外研究现状

关于计算机化的决策支持系统（DSS，Decision Support System），来自许多学科的学术研究者们已经对其有了几十年的研究 [2]。时至今日，可以将历史上的决策支持系统分为五个广义上的类别：

* **模型驱动的决策支持系统**。模型驱动的决策支持系统强调对模型的访问和操作，简单的量化模型在系统中提供了最基本的功能。模型驱动的决策支持系统根据用户提供的有限的输入数据和相关参数帮助决策者分析情况，但是一般情况下，由于数据规模有限，这类系统并不需要大型的数据库提供支持。
* **通信驱动的决策支持系统**。通信驱动的决策支持系统使用网络和通信技术来促进决策相关的协作和通信。在这些系统中，通信技术是架构中最主要的组件。在20世纪80年代初，学术研究人员开发了一种新类别的软件来支持群体决策，称为“群体决策支持系统”（GDSS，Group Decision Support System） [3]。早期的群体决策支持系统有亚利桑那大学开发的Group Systems、明尼苏达大学研究人员开发的SAMM系统等，其中Group Systems最终演变成了一个成熟的商业产品。通常，群件（Groupware），公告板（Bulletin Boards），音频和视频会议是用于通信驱动决策支持的主要技术。伴随着网络通信技术的迅速发展，通信驱动的决策支持系统在同步方面的研究得到了极大的发展。
* **数据驱动的决策支持系统**。一般来说，数据驱动的决策支持系统强调访问和操纵包括系统内部数据、外部数据和一些实时产生数据在内的时间序列数据。在系统的设计中，可以只包含简单的文件系统，它可以被查询和检索工具访问，这为决策支持系统提供了最基本的功能；也可以使用数据仓库系统，利用预设的自动化脚本提供更为复杂和高级的功能。此外，数据驱动的决策支持系统还可以与在线分析处理模块搭配，通过与大量历史数据集合的分析相关联，提供高水平的决策支持和额外的功能 [4]。
* **文档驱动的决策支持系统**。文档驱动的决策支持系统使用计算机存储和处理技术来提供文档检索和分析。大型文档数据库可以包括扫描文档、超文本文档、图像、声音和视频等。搜索引擎是与文档驱动的决策支持系统相关联的主要决策辅助工具，全球网络技术大大增加了文档的可用性，并促进了文档驱动的决策支持系统的开发。
* **知识驱动的决策支持系统**。知识驱动的决策支持系统可以向用户提出建议或推荐方案，这类系统是具有解决专业问题能力的人机交互系统。这里的“知识”包括了对特定领域背景知识的认知能力、对该领域内的问题的理解能力、以及解决这些问题的能力。例如，[5]和[6]中介绍的专家系统一般被认为是知识驱动的决策支持系统。

关于使用计算机科学来进行医学诊疗方面，早在1966年，国外就已经有研究者提出了计算机辅助诊疗的概念 [7]。尽管计算机技术在临床医学中的应用方式不尽相同，但是从很早开始人们就已经意识到，计算机的计算能力可以帮助人们从疾病和症状中进行快速的筛选，节省大量的时间，从而对医疗从业者提供巨大的帮助。而国内对智能化医疗理论的研究起步于20世纪70年代，早期的系统原理基本上都是基于规则的决策推理。近年来，电子病历的推广为这方面的研究提供了数据基础，机器学习的发展则提供了算法支持 [8]，这使得国内医学诊疗的智能化水平进一步提高到了一个新的高度。

数据集的质量和大小对机器学习相关研究而言是至关重要的。在医学研究领域，由于不同的医疗机构各自的标准不同，整合不同机构的数据比较困难，因此数据的一致性和样本容量的大小一直都是相关研究的瓶颈。国外的研究起步较早，已经形成了包含大量数据的高质量数据库，例如在癌症研究领域有美国国家癌症研究中心主导的SEER（Surveillance, Epidemiology, and End Results）项目和由美国外科医师学院和美国癌症协会联合赞助的NCDB（National Cancer Database），它们为学术研究者们提供了大量的数据支持。

与此相比，国内目前关于智能诊断的研究中数据集样本容量太小。近期有学术研究者使用了国内某医院的电子病历数据进行决策支持系统的研究 [9]，但是研究中仅使用了该医院的977条住院记录，数据的特征数量也不多。稍早一些的研究 [10]也面临了相同的问题，其使用的整个数据集中只包含了698个病例。过小的样本容量可能会带来很多问题，例如造成过拟合等现象，以至于无法反映数据中蕴含的真实规律等。

除此之外，国内的许多研究 [11] [12]仅针对其所提出理论，在小数据集上做了测试，均没有形成完整实用的系统投入到实际生产环境之中。

## 本文主要工作

本文的研究以病例数据信息为基础，主要分为三个部分。

首先，在研究初期，我们基于医院真实的病例诊疗记录，构建了一个标准化的高质量病例数据库作为后续研究的数据基础。由于国内的各个医院对于一些医学指标的命名和度量方式差异很大，还没有形成一个统一的标准，不同医院所产生的病例数据的格式、所包含的医疗指标的格式和内容等都因不同程度的差异而无法相互兼容，必须通过显示的规则进行转换才可能被统一使用。相比之下，一个标准化的数据库则非常容易扩展。有了统一的标准作为基础，各个医院就可以将各自的数据进行整合与共享，这样一来，可供医学研究的数据集规模将会大大增加。此外，在标准的基础上增加额外的特征用来表示新的医疗指标信息，也不会影响到数据的兼容性。

其次，在研究中我们对机器学习相关算法进行了深入的了解，并且熟练地进行运用，最终顺利开发了一个基于病例数据的智能诊断支持系统。通过研究，我们成功地将智能诊断支持系统投入了实际使用，应用到多学科讨论的过程中，为患者推荐诊疗方案以供参加讨论的医生参考，从而帮助患者制定一个更为完善的综合治疗方案。为了提高该系统的各方面性能，我们采集了系统在试运行阶段产生的数据，根据推荐结果和实际的讨论记录数据进行分析，研究了智能诊断系统中的各个参数对多学科讨论过程的影响，例如推荐方案的准确度、医生决策的分歧度等。

此外，为了进一步提高推荐方案的准确度，保证系统长期运行的可靠性，本文对实际存在的概念漂移现象进行了阐述，提出了一种适用于概念漂移的推荐算法。为了验证该算法的可行性，我们为实验构造了两个数据集，一个是人为构造的具有概念漂移现象的模拟数据集，另一个是使用病例数据库中导出的数据整理的真实数据集。本文同时在两个数据集上做了相应的实验，并对结果进行了分析和比较。

## 本文组织结构

本文共分为七个章节，各个章节的内容安排如下：

第一章绪论，包含了本文研究的背景和意义，介绍了决策支持系统的研究背景，以及当下国内外在医学诊疗智能化这个领域的研究现状，最后总结了本文研究的主要工作。

第二章相关技术，介绍了本文的研究过程中所涉及的相关技术。本章所述技术主要和推荐模块相关，因而对基于规则的推理和基于案例的推理分别进行了详细的介绍和比较。此外，对于这两种不同的推理方式，本章还具体介绍了其中的典型算法：决策树算法和k-NN算法。

第三章介绍了智能诊断支持系统的架构。智能诊断支持系统是本文的最主要成果之一，这一章对系统的结构进行了具体的介绍，包括所使用的工作流引擎、使用的外部病例数据库和核心模块。其中的外部数据库为上海交通大学乳腺癌病例数据库，这是一个高质量的乳腺癌病例数据库，为该系统提供了数据支持。本文研究中的实验使用的真实数据集，就是从这个数据库中整理得到的。因此本章还具体介绍了该数据库的详细情况，包括数据表的结构，数据的详细字段，数据库系统提供的录入、查询等基本功能和数据分析等其他高级功能。

第四章智能诊断支持系统的推荐模块，分别介绍了系统中使用的基于规则和基于病例的推荐算法。由于该系统已经投入实际使用并在试运行阶段产生了详细的数据，本文因此对这些数据进行了详细统计，对推荐模块的性能表现进行了展示和分析。此外，本章还介绍了不同的推荐方法得到的结果的粒度不同的现象，并且提出了综合考虑不同粒度的推荐结果，从而提高预测准确率的方法。

第五章适用于概念漂移的推荐算法，是为了提高智能诊断支持系统的推荐准确率而进行的研究。这一章介绍了概念漂移的概念，详述了现实场景中实际存在的概念漂移现象，并且提出了一种可以处理概念漂移的自适应分类算法，以用于进一步提高推荐模块的性能。此外，为了验证算法的有效性，我们分别在模拟数据集和真实数据集中进行了实验，对实验数据进行了具体分析。

第六章智能诊断支持系统的具体实现，对该系统的具体实现方式进行了详细的介绍，包括所使用的操作系统、开发工具、第三方的运行库等。除此之外，本章还展示了系统的相关界面。

第七章总结与展望，对全文进行了总结，并展望未来，列举了一些值得继续深入研究的方向和问题。

# 相关技术

## 基于规则的推理

在知识表示方法和推理方法中，符号化的规则是其中最流行的方法之一 [13]。这种方法的流行原因主要源于其自然性，符号化的规则有利于理解其所代表的知识。对于一个符号化的规则，最简单的形式如下：

|  |  |  |
| --- | --- | --- |
|  |  | (2‑1) |

其中表示一条规则中包含的条件，而则表示这条规则的所指向的结果。对于一条规则，它的条件相互之间使用逻辑关系连接起来，例如与、或、非等，因此构成了一个逻辑函数。一旦规则之中的条件被满足，规则指向的结果就会被推理得到，这个过程称作规则的触发。一般来说，规则代表的是某个领域中的通用的知识。

专家系统是基于规则推理（Rule-based Reasoning）的系统中最著名的一个例子 [5] [6]。第一个基于规则的专家系统是DENDRAL [14]，它是一个有机化学领域的应用。另一个早期流行的专家系统是MYCIN [15]，它用于医学诊断领域，后来又由此衍生出第一个带命令行的专家系统EMYCIN [5]。专家系统在很早就被投入商业使用，DEC公司（Digital Equipment Corporation）开发了第一个商用的专家系统XCON [16]，用于辅助配置计算机系统。

典型的专家系统主要包括以下几个部分：规则库、推理引擎、运行存储和解释机制。推理引擎使用规则库和所遇到的具体问题之中所蕴含的知识进行工作，这些知识被保存在了运行存储中以供后续使用。推理引擎主要根据推理方法的不同，分为后向链接型和前向链接型。后向链接型的推理引擎通过预期的目标指引进行规则和数据的匹配，而前向链接型的引擎则通过数据和规则的匹配，进而得到指向的结果。解释机制根据推理过程得到的结果提供详细的解释，其过程通常是简单地追踪推理引擎的推理步骤，并对每一个步骤给出可读性强的解释，从而阐明了得到当前结果的详细过程。

符号化的规则通常通过和特定领域的专家进行交流，根据专家的经验翻译之后得到。不过规则也可以通过机器学习的方法 [17]从实际的数据中提取出来，例如使用ID3算法 [18]从通过数据直接构造规则。

符号化的规则作为一种知识表示形式有许多优点，当然也存在一些非常关键的缺点 [6] [19]。

### 基于规则推理的优点

基于规则推理的优点主要包括以下几点：

* **对知识的简洁表示**。对于一个领域中的知识，使用规则可以将其划分成许多互相独立的小块，用一组互不干扰的规则简单地对其进行描述。
* **表示的自然性**。规则是一种非常自然的知识表示方法，它看上去和自然语言的表达方式非常相似，这使得它的可理解性非常高。同时，规则可以自然地模拟专家在各个领域中对问题的思考方式。
* **模块化**。每一条规则都独立地表示了一部分的知识，可以在不考虑任何技术细节，不干扰任何别的规则的情况下，直接被加入知识库，或者从知识库中删除。这个特性为基于规则的推理系统的开发过程赋予了极大的灵活性。
* **提供解释**。在一些特定的领域，例如医疗，提供解释是非常必要的功能，因此能为得到的结论提供很直观的解释是一个非常重要的特性。这个特性其实是基于规则推理的自然性和模块化共同作用下的结果。

### 基于规则推理的缺点

基于规则推理的缺点主要包括以下几点：

* **知识获取方式的瓶颈**。规则提取的普遍做法是通过与专家交流，这是一个非常麻烦和耗时的活动。最大的问题在于专家是否能将自己的知识很清晰地表达出来翻译成规则，以及是否有足够的专家有足够的时间进行这样的活动。因此，从专家那里获取到的知识可能是不完整的，甚至可能由于沟通和理解的过程中产生的偏差，只有一部分是正确的。虽然使用机器学习的方法可以从实际数据中提取规则，从而解决一部分知识获取的问题，但是这需要大量数据的支持，对于一些领域中出现频率较低的边缘条件下的情况，这种方法并不能起到多大的作用 [20]。此外，特定的领域知识可能非常复杂，因而需要大量的规则进行支持，这些都是值得关注的问题。
* **规则的脆弱性**。当输入数据中存在缺失属性的时候，规则引擎会有极大的可能性无法推理得出最终的结论。对于一个特定的规则，一系列的属性必须是已知的，这样才能判断是否符合规则所对应的逻辑函数。此外，如果输入的数据中包含非法的格式或者异常的值，规则引擎也可能无法正确地工作。
* **推理效率问题**。在一些情况下，推理引擎地性能可能无法满足要求，特别是在规则库特别庞大的情况下。虽然有一些快速匹配算法，例如Rete算法 [21]，可以改善前向链接型推理引擎的性能，但是基于规则的推理引擎通常无法很好的解决可扩展性的问题。任何情况下，推理引擎都会对输入的数据在整个规则库中进行匹配，即使相同的输入已经在之前出现过多次，相同的匹配过程也进行了多次，推理引擎还是会从零开始按照既定的规则进行推理，执行相同的操作。这无疑是低效的，尤其是在每次处理都相当费时的情况下，这种机械的行为模式大大增加了运行时间。
* **大型规则库的维护问题**。随着规则库的规模变得越来越大，对规则的维护难度也逐渐提高。在规则库中可能会出现冗余的规则、冲突的规则、条件重复或者缺失的不完全规则等等，它们的存在对推理效率和正确性都有很大的损害。为了处理这类问题，一系列复杂的验证方法必须被加入系统中，并在规则库发生变化的时候执行检查。
* **无法自我更新**。由于缺乏从过去处理的案例中获取信息的能力，一个基于规则的系统是无法自我更新的。也就是说，解决问题的经验无法被利用起来，为后续的案例提供更好的解决方案。事实上，这些经验在推理过程中可以起到决定性的作用，特别是在遇到规则库中没有覆盖到的边缘案例的情况下。而在实时运行的过程中，除非有相关的专家不断地进行人工干预补充规则库中的内容，否则系统是无法获取到最新的经验的。

### 决策树算法

决策树算法是基于规则推理的典型算法之一。决策树是一种提供决策支持的工具，它因其树状的模型以及提供决策支持的功能而得名。在决策树中，内部节点表示在一个属性上的测试，而叶子节点代表最终的类别。决策树模型可以被用来解决许多基本问题，诸如多阶段决策，表查找，最优化等。它的树状模型天然还原了做决策的过程，将复杂的决策过程按照实际的决策的步骤有条理地拆分成了一系列简单的选择，因而能以一种直观的形式从头到尾完整地解释从做出最终决策的整个过程。

最早的时候，决策树是人为构建的。随着机器学习的发展，决策树被广泛地以树状预测模型地身份应用于机器学习地过程中，决策树的构建也从最初的按照经验人为构建转变为根据训练样本集合自动生成。在机器学习和数据挖掘领域，不乏一些需要展示或阐明处理过程的场景，如专家决策系统等。大多数复杂的机器学习算法由于其计算过程晦涩难懂，无法满足这类需求，而决策树模型可以学习训练数据中的规律，并将这些规律以直观的树状图表现出来，正是最合适的解决方案之一。

构造决策树的过程和人做决策的行为模式相似。给定一个数据集S，包含多种属性的值以及所属的分类，首先要做的是使用一些统计方法选择一个属性A作为根节点，根据属性A的值将数据集S划分为多个子集。在这些子集上重复上述过程，直到满足一个特定的终止条件，如子集中的超过一定比例的数据都属于同一个类别，最终叶子节点就代表它对应的集合中大部分数据的类别。



图2-1决策树的一般结构

Fig.2-1 Sample structure of a decision tree

决策树算法的优点：

* 易于理解和解释。决策树直观地将规则表示出来，与日常生活中决策的过程非常相似，这使得用户可以只通过决策树本身就可以快速地理解其中的规律。
* 适用于多种数据类型。决策树模型可以同时处理数值型的连续数据和类别型的离散数据，而许多复杂的机器学习算法往往要求单一的数据属性。
* 计算开销小。无论是手动创建还是自动生成，决策树模型的构建都可以离线完成，而且数据中的规律已经抽象成简单的规则，大大缩减了存储空间。而给定一个完整的决策树，对样本进行测试的时间复杂度几乎是常数级别的，因而在实际应用中的开销非常小。

决策树算法的缺点：

* 模型不易维护。产生新的样本的时候，一般无法简单地更新模型或者是直接添加新的规则以适应新的规律。由于新的规律可能会对很多节点造成影响，一般不得不重新构建整个决策树。
* 容易过拟合。当样本的特征较多或者决策树的深度较大的时候，抑或是样本中的离群值（outlier）太多的时候，容易产生过拟合的情况。

## 基于案例的推理

对基于案例的推理（Case-based Reasoning）的研究可以追溯到上世纪七十年代末 [22] [23]，相关研究中提出了名为MOPs（Memory Organization Packets）的数据结构用于存储知识和组织案例，并在此基础之上发展出了基于案例推理系统的先驱者IPP [24] [25]和CYRUS [26]。

基于案例的推理方法将大量历史案例和相应的解决方案存储在案例库中，当遇到一个新案例需要处理的时候，就从中寻找相似的记录作为参考。当一个新的输入案例需要处理的时候，基于案例推理的系统将会按照CBR循环 [27]里的四个阶段进行推理：

* 检索（Retrieve）阶段
* 重用（Reuse）阶段
* 修正（Revise）阶段
* 保留（Retain）阶段

在检索阶段，系统在案例库中存储的案例里找出和新案例最相关的历史案例。在这个阶段中，索引技术和相似度度量是重要的部分。建立索引可以高效地检索到相关病例，因而大大缩短执行时间。相似度度量则决定了案例之间相关程度的评估方式，其中一种简单的方式是最近邻匹配，从案例库中找出距离最近的一个或者多个历史案例作为参考。

在重用阶段，将会基于这些最相关的历史案例，并结合当前案例的领域知识等生成一个解决方案，作为新案例解决方案。

在修正阶段，系统验证所提出的解决方案的正确性。其中，系统和外界用户的交互过程就发生在这个阶段。

在保留阶段，系统将会判断当前案例的重要程度，并决定是否要从当前的案例中学习知识。换言之也就是说，是否要把当前的案例加入到案例库中，以供将来参考。

在其他类型的推理系统中，知识库在系统的运行期间更倾向于保持不变。而在基于案例的系统的运行期间引入新知识的过程，增强了其推理能力，这是它的主要优势之一。

### 基于案例推理的优点

基于案例推理的优点主要包括以下几点：

* **表达专业知识的能力**。在基于规则推理中会遇到的，难以将一些专业的知识直接翻译成通用规则的情况下，使用相关的案例来表达这些知识是一个很好的方式。
* **表示的自然性**。使用相关案例是一个很简单的知识表示方式，同时对用户而言，它也非常易于理解。
* **模块化**。每一个案例都是一个离散的、独立的知识单元，它们可以被插入知识库，或者从知识库中删除，而不对其他案例造成直接问题。
* **知识易获取**。在基于案例的知识表示方式中，知识的获取通常并不是一个问题，在绝大多数的应用领域，案例的获取非常方便。然而事实上，的确存在一些不容易获取到案例的领域。
* **自我更新能力**。在碰到新的案例，并且需要学习其中蕴含的新知识的时候，基于案例推理的系统可以在运行时就将当前的案例插入到案例库中，使得系统的有效性得到了维护。自我更新的过程同时也更利于案例库的维护。
* **健壮性**。一个基于案例推理的系统可以处理异常或是缺失的数据输入。对于缺失信息来说，系统可以通过评估相似度从相关的案例中获取信息。对于异常的输入来说，系统自我更新的机制极大地提高了处理能力，以利于处理预期之外的案例。
* **推理效率**。在基于案例的推理系统中，去调整一个已经存在的案例以适应新的问题，通常要比从头开始构建一套规则去解决一个问题要更有效率 [28]。不过值得注意的是，这并不在所有情况下都适用。

有研究 [29]报告称，对于大量服从简单规律的数据而言，使用基于规则的推理可以使得规则库规模极小，它比使用一个具有大规模案例库的基于案例的推理系统要高效得多。

### 基于案例推理的缺点

基于案例推理的缺点主要包括以下几点：

* **缺乏表达通用知识的能力**。具体的案例本来就是用来表示专门的知识的，因而和基于规则的系统相比，无法表示通用知识是它的一个重要的缺点。
* **知识获取问题**。诚然在大多数情况下，如果领域中有充足的案例可以被收集记录的时候，知识获取根本不是一个问题。但是不得不考虑到具体案例无法采集、或者说案例的数量不够的情况，这将会带来很多问题 [30] [31]。缺少足够数量的案例，会阻碍一个基于规则推理的系统的构建以及推理过程。不仅如此，基于案例的系统通常需要相当大的案例背景知识 [28] [32]，缺乏案例背景知识其实是对整个系统的有效性有所损害的。在Mitra和Basak的研究 [32]中表明，为了使系统代表最新的知识，需要针对领域而言特定的背景知识，从而使得系统能更精确地进行知识的获取。而由于这些特定的适应性知识一般是基于规则的，这将会从基于规则的推理系统中引入知识获取困难的问题，因而影响到基于案例推理的系统的开发和操作。正是出于这个原因，一些研究者将这一过程排除在了CBR循环之外，而另一些研究者则提出了自动获取这方面知识的机制 [33]。
* **可解释性问题**。虽然可以提供一些具体的案例，来表示推理系统得出相应结论的原因，但是和基于规则的系统相比，这种解释显得并不那么直观，而且它很难对整个推理的过程给出详细的解释。

### *k*-NN算法

-NN算法（k-Nearest Neighbors） [34]是基于案例推理的典型算法之一。在模式识别中，-NN算法可以用于分类和回归，是非参数统计 [35]的一种算法，换言之，它适用于样本分布情况未知的情景。无论是应用在分类还是回归中，-NN算法以特征空间中最近的个训练样本作为输入，并以它们作为依据产生相应的输出结果。

对于-NN分类器而言，输出结果是类别的归属关系。一般来说，一个对象的类别会以它的近邻的类别作为依据，按照少数服从多数的投票方式决定。在最近的个近邻中，出现次数最多的类别将被认为是这个对象的类别。考虑在如的极端情况下，一个对象的类别可以简单地由在特征空间中距离它最近的对象所决定。

上述方法可以使用公式2‑2来进行表示，其中表示训练集合中的第个样本，表示样本的特征所构成的向量，表示这个样本所属的类别；是一个集合，它表示目标类别的值域；集合中包含了在特征空间中距离最近的个样本，也就是样本的近邻集合；是指示函数算子，当其中的条件满足时其值为1，否则其值为0。

|  |  |  |
| --- | --- | --- |
|  |  | (2‑2) |

对于-NN回归器而言，输出结果是对象的属性值。一般来说，会在特征空间中选取个距离当前对象最近的对象，将它们的目标属性平均值作为这个对象的目标属性值。上述方法如公式2‑3所示，其中表示训练集合中的第个样本，表示在特征空间中距离最近的个样本的集合。

|  |  |  |
| --- | --- | --- |
|  |  | (2‑3) |

-NN是一种基于案例的学习算法，属于惰性学习，是机器学习算法中最简单的一种。在这种算法中，无论是分类和回归，都可以给近邻样本的贡献赋予不同的权重，使得越近的样本产生的贡献越大。例如借助线性插值的思想，我们可以简单地根据近邻的距离，给每个近邻的贡献赋以的权重，这样一来即使选取的过大，距离相对较远的样本也就不会对结果产生很大的影响了。

-NN算法的优点如下：

* 直观且有效。通过相似样本来进行分类和回归的思想，符合人们的常识，而且其结果非常直观。
* 训练成本低。-NN算法直接在样本集合中计算距离，选取需要的样本，模型简单几乎不需要学习。
* 不受样本分布的影响。由于它是非参数统计的算法，没有复杂的模型，使用的时候可以完全忽略概率等和分布相关的因素

-NN算法的缺点如下：

* 可解释性不强。-NN算法和决策树算法相比，仅根据一个相似样本的集合，产生的结果无法被很好地解释。
* 计算开销较大。由于缺乏模型和相应的参数，仅根据训练样本集合直接进行计算，当样本容量大量增加的时候，存储空间和计算时间都将大幅度增加。
* *k*值影响大。在-NN算法中必须指定*k*值，不论是使用静态的值，还是变化的动态值，它对准确率有着直接的影响，因而值的选择是一个很重要的问题。

## 本章小结

本章介绍了本文的研究过程中所涉及的相关技术。本章所述技术主要和推荐模块相关，因而对基于规则的推理和基于案例的推理分别进行了详细的介绍和比较。此外，对于这两种不同的推理方式，本章还具体介绍了其中相应的典型算法：决策树算法和k-NN算法。

# 智能诊断支持系统的架构

## 引言

在这一小节，我们将介绍智能诊断支持系统的系统结构。为了便于理解整个系统的功能，我们还将结合该系统介绍多学科讨论的一般过程。



图3-1智能诊断支持系统结构图

Fig.3-1 Structure of the MDT decision support system

图3-1展示了智能诊断支持系统的结构概览。为了使得系统能够跨平台使用，我们用RESTful Web服务 [36]的形式实现了该系统，其顶层是一个基于Web的友好的用户界面。在用户界面的背后，隐藏着系统的两大核心模块，分别提供了控制决策过程和推荐治疗方案的功能。两个模块都需要详细的病例信息，因而我们在病例数据库系统中设计了开放的API，使得这两个模块可以直接从外部数据库中读取到患者的个人资料和医疗指标等详细信息。治疗方案推荐模块从外部数据库中直接读取原始数据，将所需要的属性依照一定的格式转换为向量形式，并保存在内部数据库中以供后续使用。决策过程控制模块负责控制和记录整个多学科讨论的决策产生过程，在系统底层与一个基于ECA规则的工作流系统 [37] [38]交互。在工作流系统中，每一个被讨论的病例都会初始化一个工作流的实例，所以在系统内部的数据库中还维护了每个病例讨论的进展信息，用于监控和分析多学科讨论过程的进展和规律。



图3-2多学科讨论的决策产生过程流程图

Fig.3-2 Flow chart of the MDT diagnosis assisted by the system

如图3-2所示，在智能决策支持系统的辅助下，规范化的决策过程变得非常的高效。对于一个亟待讨论的病例，主持者（在系统中比一般用户拥有更高的权限的管理员）首先创建一个讨论过程的实例。一个讨论过程的实例包含了很多信息，例如病例的详细医疗指标信息、参与多学科讨论的医生列表、讨论的类型以及讨论的预定时间等。在主持者激活了这个讨论实例之后，相关医生可以在系统中自己的工作区界面内看到病例的信息，从而分别对患者的治疗方案进行第一次投票。待所有的医生都完成了第一次投票之后，系统会给出所有人的投票结果统计，以及推荐模块根据病例信息提供的治疗方案作为参考。流程上，如果投票结果和推荐结果一致，则该方案就直接作为讨论的结果，由系统生成讨论报告表。如果医生们的意见与推荐结果相左，或是医生们的投票结果之间产生了分歧，则对此患者进行详细的现场讨论。讨论中，医生将详细阐述各自赞同的治疗方案以及支持的理由，讨论结束后在系统中进行第二次决议投票。根据两次投票的结果统计、讨论中各学科医生的见解以及系统的推荐方案，由主治医生确定最后所采用的治疗方案。

## 工作流引擎

### 基于ECA规则的工作流模型

工作流模型主要由两部分组成，分别是控制流和数据流。

控制流部分定义了在工作流的过程中，需要被执行的各个活动的具体内容和类型，以及活动之间的逻辑关系。活动与活动之间的逻辑关系包括相互依赖、执行顺序等，约束了一个活动的执行条件和执行时间。

ECA（Event Condition Action）规则是用于事件驱动 [39]的系统中的活动规则结构，其中规则主要由Event，Condition，Action三部分组成：

* Event决定了触发这条规则的事件，例如活动结束，活动错误，活动中断等。
* Condition是决定了是否要执行后面的动作的一个逻辑测试。
* Action则表示这条规则中的后续动作，它定义了接下来具体要执行的操作。

对于一条ECA规则，当它被Event事件触发之后，会通过Condition逻辑测试的结果，来判断时候要继续执行后续的Action动作。通过ECA事件驱动，以及表示“与”和“或”的逻辑关系节点，控制流可以对事务模型进行建模。

数据流部分定义了在工作流过程中，涉及的相关数据的传输过程。这些数据不仅包括起始的参数和输入，还包括工作流执行过程中生成的中间数据。数据流一般沿着控制流的方向，控制着这些数据的传输交换，将数据从之前的活动传输给下一个需要该数据的活动。在特殊情况下，数据流也可以不依照控制流的方向，将一些数据直接传入后续的活动中。

### 工作流引擎结构

工作流引擎是工作流系统的核心，负责所有工作流实例的执行、管理与运行时操作，是系统中负载最大的部分。



图3-3工作流引擎中工作流的生命周期

Fig.3-3 Lifecycle of workflow processes in workflow engine

在工作流引擎中，每个工作流的一般执行过程可以被分为三个部分，分别是实例化、事件驱动和应用调用。它们执行的任务相互独立，设计上可以独立划分为不同的模块，各自向模块外部提供API，模块之间通过互相调用API进行数据交换。

如图3-3所示，当获得一个实例请求时，工作流引擎首先根据工作流模型，从数据库中读取相应的数据，并将其实例化成一个具体的工作流实例。基于ECA规则，工作流从起始状态开始以事件驱动的方式路由与执行，并根据工作流中的具体动作，调用通过相应的API调用各类服务。这些服务执行任务并将结果返回给应用调用模块，进而传递回事件驱动模块。通过实例化模块、事件驱动模块和应用调用模块，对于一个工作流请求，工作流引擎实例化了相应的工作流，并在工作流的生命周期中完整地执行了相关的任务。

### 系统角色与工作流之间的交互

智能诊断决策系统中根据不同用户的职能区分了不同的角色，例如输入员、管理员、超级管理员等。系统中不同角色与底层工作流系统的交互不全相同，如图3-4所示。



图3-4智能诊断决策系统角色与工作流系统的交互示意图

Fig.3-4 Interaction between workflow system and MDT system roles

对于一个病例的讨论过程，系统管理员可以控制它在不同阶段的开始和结束，对应到工作流中即为，管理员的动作控制了这些阶段中触发ECA规则的Event事件。由图3-4可知，当不止一个讨论的实例被管理员激活的时候，这些讨论是并行进行的，也是就说系统的普通用户（也就是参与讨论的各个学科的医生）可以分别参与不同讨论的投票。由于第二次投票之前需要各学科的专家各抒己见进行全方位的讨论，第二次投票的开始流程是由管理员触发的，这使得第二次投票的过程是串行进行的，同一时间只有一个讨论实例进行到第二次投票的阶段。对于管理员而言，开始第二次投票的动作对于当前讨论而言最后一次和工作流系统交互，但是对于普通用户而言，投票的过程是他们和工作流的最后交互。当工作流流程结束之后，针对当前病例的讨论结果也就产生了。

## 病例数据库

### 病例数据库介绍

算法的研究和智能诊断支持系统的开发都需要一个高质量的病例数据库，因此我们选择上海交通大学乳腺癌数据库（SJTU-BCDB）作为数据源。该数据库以上海交通大学医学院附属瑞金医院为核心，联合国内多家医院共同参与，旨在为国内医院的乳腺癌研究提供高质量的数据和服务。数据库中记录了在合作医院中就诊的乳腺癌患者的基本信息、医疗指标、治疗方案和治疗效果等超过100个属性的详细信息，目前数据库中病例总数超过7000例。

### 病例数据表结构

病例数据库中，每一条病例包含了丰富的基本信息和医疗指标等，因此数据根据不同的类别，在数据库底层使用了不同的数据表来表示。这些类别包括系统回顾和既往史、患者临床特征、影像学检查、活检病理学检查、新辅助治疗、手术及病理、手术病灶、术后辅助治疗和随访记录，它们各自的属性详情如表3-1所示，各表之间的关系如图3-5所示。



图3-5 病例数据库的物理数据模型图

Fig.3-5 Physical data model of the database schema



图3-6 患者临床特征数据结构图

Fig.3-6 Schema of table `clinicals`



图3-7 活检病理学检查数据结构图

Fig.3-7 Schema of table `pathologies`



图3-8 影像学检查数据结构图

Fig.3-8 Schema of table `imagings`



图3-9 系统回顾和既往史、术后辅助治疗数据结构图

Fig.3-9 Schema of table `histories` and `adjuvants`



图3-10 手术及病理、手术病灶、随访记录数据结构图

Fig.3-10 Schema of table `operations`, `operation\_lesions` and `followups`



图3-11 新辅助治疗数据结构图

Fig.3-11 Schema of table `neoadjuvants`

表3-1病例相关数据表概览

Table 3-1 Overview of patient related tables

|  |  |  |  |
| --- | --- | --- | --- |
| **病例相关信息** | **实体表名** | **与病例对应关系** | **数据结构** |
| 系统回顾和既往史 | Histories | 1:1 | 图3-9 |
| 患者临床特征 | Clinicals | 1:1 | 图3-6 |
| 影像学检查 | Imagings | 1:1 | 图3-8 |
| 活检病理学检查 | Pathologies | 1:1 | 图3-7 |
| 新辅助治疗 | Neoadjuvants | 1:1 | 图3-11 |
| 手术及病理 | Operations | 1:1 | 图3-10 |
| 手术病灶 | Operation\_lesions | 1:\* | 图3-10 |
| 术后辅助治疗 | Adjuvants | 1:1 | 图3-9 |
| 随访记录 | Followups | 1:\* | 图3-10 |

### 病例数据库系统功能

#### 病例维护

一般乳腺癌患者的治疗流程按顺序包括以下几个阶段：入院、体检、新辅助治疗、手术、术后辅助治疗。有条件的医院还会定期主动对出院的患者通过电话、问卷等不同形式进行随访，了解患者的后续病情发展（是否痊愈或者复发等），更好地应对病情二次复发的同时，也收集了诊疗的实际效果，为医学研究分析提供了数据基础。

在病例维护模块中，病例数据库系统按照患者正常的就诊流程，抽象出了病例的完整生命周期，并在这个生命周期的每一个阶段都提供了相应的接口，使得系统的输入员和管理员可以对每一个病例的信息进行维护和更新。这保证了数据库系统中的病例始终是最新最准确的，大大提高了数据的质量。

#### 病例查询

在病例查询模块中，系统提供了基本查询、高级查询和自定义查询三种功能。

* 基本查询：根据姓名、身份证号、住院号、年龄、手术时间等基本信息查找病例。
* 高级查询：根据常见的医疗指标如ER、PR、cT分期、cN分期等，查找某一类特定的病例。
* 自定义查询：以“条件组”为单位，自定义查询条件，将不同的限定条件按照AND、OR的关系进行组合、嵌套处理，生成相应的SQL语句提供复杂的查询。

#### 数据分析

除了病例维护和病例搜索、导出等常见功能之外，病例数据库系统还提供了数据分析功能。数据分析模块分成三个部分如下：

* 基本统计
* 专业统计
* 数据库使用情况统计

其中基本统计包括了数据库中病例的总体情况，包括按录入时间、病例BMI、年龄等为依据的分布统计，并以柱状图和饼状图直观地展示其中的规律。

为了使医学研究者能更便捷地观察并分析数据库中的病例，除了基本统计之外，系统还增加了专业统计功能。在专业统计范围内，我们增加了一些核心的医疗指标作为依据，如手术方式、术后病理、ER、PR、CerbB-2、术后辅助治疗方式等，提供了病例的分布统计功能。

此外，根据乳腺癌领域的专业知识，系统中还对病例的信息进行自动分析，根据相应的公式和规律统计出特定类型的患者并直观地展示出其分布情况，系统可以直接根据病例的一系列医疗指标，根据既定的规则计算出不同分子分型的病例数量以及分布。

系统中不仅包含了患者就诊前后的各项信息，患者出院以后的随访记录也在系统中进行维护，这使得诸如OS和DFS一类的生存统计分析可以直接由系统中的记录直接计算产生，避免了繁琐的人工检索和计算。在医学研究的过程中，研究者可以直接在系统中筛选相应的病例进行统计，直接产生直观的统计图，避免了探索规律的过程中对数据进行繁琐的处理和计算。

## 系统核心模块

在整个智能诊断支持系统的结构中，最为核心的部分是投票模块和推荐模块。投票模块控制着整个多学科讨论的流程，推荐模块通过推荐治疗方案潜移默化地影响着决策的结果。本节对投票模块、推荐模块以及它们在系统中互相协作共同工作的方式进行介绍。

图3-12是投票模块和推荐模块共同参与的多学科讨论过程的顺序图，从图中可以清晰地看到系统管理员负责触发投票的开始和结束，普通用户负责对病例的治疗方案进行投票，以及查看相应的推荐方案。



图3-12投票和推荐过程顺序图

Fig.3-12 Sequence diagram of voting and recommendation modules

### 投票模块

投票模块控制着整个多学科讨论的流程。在系统中，投票模块与底层工作流交互，通过工作流系统维护每一个病例的多学科讨论实例的状态，进而管理着每一个讨论的生命周期。

在投票模块中记录了每一个讨论实例的所有具体信息，包括讨论的创建时间、参与讨论的用户信息、用户每次投票的时间和内容等。利用这些数据，投票模块还提供了投票结果的分析和统计功能，包括投票结果的分布，两次投票的结果变化情况等。这些功能不仅提供了对投票过程的直观展示，还为将来更深入的分析和研究提供了数据基础。

### 推荐模块

系统中使用的推荐模块按照推荐方法不同，又分为基于规则的推荐模块和基于病例的推荐模块。基于规则的推荐模块依据既定的规则，一步一步得出最终的结果；基于病例的推荐模块查找相似的病例，根据历史病例采用的治疗方案进行推断。它们分别独立地对病例进行治疗方案推荐，推荐的结果均会在系统的用户界面中展示，以供参考。

由于推荐模块需要详细的病例信息作为支持，两种类型的推荐模块均依赖于病例数据库。对于一个病例，推荐模块根据病例唯一标识的住院号从病例数据库中查询详细数据，并根据推荐方法所需要的信息进行格式化，最终计算出推荐的治疗方案。

## 本章小结

本章对智能诊断系统的结构进行了具体的介绍，包括所使用的工作流引擎、基于规则的推荐模块和基于病例的推荐模块。此外，在推荐模块这一部分，还介绍了不同的推荐方法得到的结果的粒度不同的现象，并且提出了综合不同粒度的推荐结果，从而提高预测准确率的方法。除此之外，由于该系统已经投入实际使用，在试运行阶段产生了详细的数据，本章最后进行了详细的统计和分析。

# 智能诊断支持系统的推荐模块

## 基于规则的推荐模块

NCCN（National Comprehensive Cancer Network）是致力于患者护理，研究和教育的领先的癌症中心联盟。为了提高癌症护理的质量，疗效和效率，使患者能够过上更好的生活，NCCN致力于成为定义和推进高品质、高价值癌症护理的世界领先者。NCCN提供的核心资源是NCCN肿瘤学临床实践指南（NCCN Guidelines） [40]，这些指南是肿瘤学临床对策的公认标准，是任何医学领域中最全面，最经常更新的临床实践指南。

上海交通大学医学院附属瑞金医院乳腺疾病诊治中心，整合乳腺外科、肿瘤放化疗科、病理科、放射诊断科、超声诊断科、核医学科和整形外科、临床营养科等多学科群的优势，是致力于乳腺疾病预防、诊断与治疗的医疗、教学和科研机构。作为国内乳腺疾病诊疗的权威机构之一，瑞金医院乳腺疾病诊治中心制定的了比NCCN实践指南更为细致的乳腺癌诊疗指南，具有很重要参考价值。

在系统中，我们参考了NCCN实践指南中的乳腺癌部分，和上海交通大学医学院附属瑞金医院乳腺疾病诊治中心制定的乳腺癌诊疗指南（RJ Guidelines），基于两套规则分别生成了相应的决策树，实现了诊疗方案的推荐功能，投入到了实际使用中。

## 基于病例的推荐模块

在这一节，我们介绍了基于-NN算法的诊疗方案智能推荐模块，内容包括对诊疗方案推荐问题进行了抽象，提出了一种针对离散属性改进的相似度度量方法，并详细阐述了推荐模块是如何工作产生推荐方案的。

### 相关记号

集合是一个包含个样本的训练数据集，其中每个样本有个不同的属性。我们用表示这个属性的集合，用表示第个属性，其中。对一个属性而言，样本可以取值的范围用集合表示，样本目标类别取值范围用集合表示。例如，样本集合中的第个样本可以如下表示：

|  |  |  |
| --- | --- | --- |
|  |  | (4‑1) |

其中以向量的形式表示了样本在各个属性上的取值。

按照上述定义，我们有以下约束条件：

|  |  |  |
| --- | --- | --- |
|  |  | (4‑2) |
|  |  | (4‑3) |

相似度度量方式是影响-NN分类器性能的最重要的因素之一。对于两个样本和，我们用来表示它们在特征空间上的距离。从这个距离度量出发，我们可以使用公式4‑4进行转换，从而得到一个值域为的相似度的度量 [41]。

|  |  |  |
| --- | --- | --- |
|  |  | (4‑4) |

对于数值型的属性，我们可以使用马氏距离 [42]作为其相似度的度量方式，其中 是协方差矩阵。

|  |  |  |
| --- | --- | --- |
|  |  | (4‑5) |

### 离散属性的相似度度量

对于类别型的离散属性，其值之间并没有特定的大小关系，所以不能像数值型的属性一样简单地使用马氏距离来直接度量样本之间的相似度。通常，为了计算样本和之间的相似度，需要先分别定义每个属性上的距离，最后各个属性加权综合计算得出最终的结果。

|  |  |  |
| --- | --- | --- |
|  |  | (4‑6) |

其中是样本和在属性上的相似度，因子表示属性所占的权重。

在我们的推荐方法中，我们基于*overlap*方法 [43]，在具有离散属性的样本上定义了一种新的相似度度量方法。在*overlap*方法中，为了计算两个样本的相似度，仅仅需要简单地逐一比较样本在各个属性上的值是否一致，最后根据一致程度来计算总体的相似度。如公式4‑7所示，如果两个样本和在属性上的值相同，则样本和在属性上的相似度为1，否则为0。

|  |  |  |
| --- | --- | --- |
|  |  | (4‑7) |
|  |  | (4‑8) |

然而在实际的情况下，数据集的质量以及属性覆盖程度很难达到完美的理想情况，样本在某个属性上的值的缺失非常常见，使用*overlap*方法度量相似度过于粗糙。对于三个样本、和，考虑以下条件下它们之间的相似度：

|  |  |  |
| --- | --- | --- |
|  |  | (4‑9) |
|  |  | (4‑10) |

注意到样本和是完全不一样的样本，它们在所有属性上的值均不同。而对于样本，我们其实并不知道它和样本在各个属性上的值是否不同，但是能肯定的是，最坏情况下，它也不过和类似，在所有属性上都与不同而已。基于这层考虑，的值应该是的值的下界，我们应该给赋予一个更为折中的值。

因此，我们在其基础上做了一些改进，产生了一种新的相似度度量方法。我们称之为*modified overlap*，它的定义如下：

|  |  |  |
| --- | --- | --- |
|  |  | (4‑11) |
|  |  | (4‑12) |

在这种度量方法之中，如果两个样本在对应的属性上具有相同（不同）的值，它们将会有一个较高（低）的相似度，当两个样本的属性有值缺失时，它们在该属性上就会有一个介于两者之间的中等的相似度。这样的改动看似很微小，但是将属性缺失的情况和属性值不同的情况加以区分，使得相似度的度量方法较*overlap*方法更为精细一些。

### 目标类别

通常，在-NN分类器中一般使用少数服从多数的投票方法，在个最近邻中找到一个出现次数最多的类别作为最终推荐结果。实际上，距离待推荐的样本距离越近的邻样本，对最终的结果应该具有更高的贡献。有了相似度这个度量指标，我们可以把最终目标类别的确定过程也做得更加精细。

给定一个待推荐的样本，我们并不直接使用少数服从多数的投票来简单地确定样本的目标类别，而是使用相似度作为训练集合中每个相邻样本的权重。对每一个候选的类别，将支持该类别的样本的权重相加作为评分，具有最大权重和的那一个类别就被认为是最终的推荐结果。

|  |  |  |
| --- | --- | --- |
|  |  | (4‑13) |

其中表示训练集合中的第个样本，表示所有类别的集合，表示在特征空间中距离最近的个样本的集合。是指示函数算子，当其中的条件满足时其值为1，否则其值为0。

为了衡量推荐结果的可信程度，我们使用来度量，定义如下：

|  |  |  |
| --- | --- | --- |
|  |  | (4‑14) |

当表示在特征空间中距离最近的个样本均和当前样本重合时，距离最小相似度最大，达到最大值1。

## 推荐结果的准确率

智能诊断支持系统发行版已于2016年9月上线，截至2016年11月试运行两个月时间已有336例乳腺癌患者的多学科讨论记录。虽然由于时间原因，数据量较小，我们仍然可以基于这些数据统计出一些有用的信息，以利于系统功能的分析和改进。

医生投票和推荐模块的准确率，将根据不同的辅助治疗类型分别统计。辅助治疗类型分为四类，其缩写和含义表示如下：

* CT：化学治疗方案（Chemotherapy）
* RT：放射治疗方案（Radiotherapy）
* ET：内分泌治疗方案（Endocrine Therapy）
* TT：靶向治疗方案（Targeted Therapy）

### 准确率分析

在各种类型的治疗方案的候选项中，有一项名为“其他”，不属于候选列表中其他选项的罕见方案均会被归为“其他”。对于k-NN这类基于案例的推荐算法而言，方案为“其他”的案例中混杂了各种类型的样本，其实它们在属性值上表现的特征并不一致，从而很难正确地推荐出“其他”这种治疗方案。

因此，在评估准确率的时候，我们对所有多学科讨论（All Cases）和去除了最终方案为“其他”的多学科讨论（Explicit Cases）这两个集合分别进行了统计。

使用-NN算法进行治疗方案推荐的准确率如图4-1所示，其中kNN(n)表示-NN算法的结果中评分最高的个方案中包含了最终的实际方案的情况。从结果中可以看到，由于去除了实际方案为“其他”的案例，对于不同类型的治疗方案，在Explicit Cases上的推荐准确率和All Cases相比均有了明显的提高。

使用规则进行治疗方案推荐的准确率如图4-2所示，可以看到在四种类型的治疗方案上，NCCN规则的准确率都比RJ略高一些。实际上，NCCN规则的粒度较粗，在治疗方案的诊断上，只区分“需要”和“不需要”，而并不精确到具体某种特定的治疗方案，因而在准确率计算的过程中表现更好，这是很正常的现象。而RJ规则的粒度较细，对于治疗方案细分到了使用的药物，其推荐结果相比NCCN规则而言则更具体。由于两种基于规则的推荐方法都明确包含了得出“其他”方案的规则，在All Cases和Explicit Cases上的准确率并无太明显的差别，符合预期的结果。

图4-1 使用-NN算法进行治疗方案推荐的准确率

Fig.4-1 Accuracy of case-based recommendation using -NN algorithm

图4-2 使用规则进行治疗方案推荐的准确率

Fig.4-2 Accuracy of rule-based recommendation using decision tree

### 综合比较

图4-3 用户投票和推荐模块在All Cases中准确率比较

Fig.4-3 Accuracy comparison of voting and recommendation over all cases

图4-4 用户投票和推荐模块在Explicit Cases中准确率比较

Fig.4-4 Accuracy comparison of voting and recommendation over explicit cases

用户投票和推荐模块的准确率比较如图4-3和4-4所示。其中1st vote表示初次投票的结果，用以代表多学科讨论中医生的平均决策水平；1st vote max表示初次投票中准确率最高的用户，用于代表经验最丰富的医生的决策水平。

结果和预期相符，经验丰富的医生由于大量的知识和经验的积淀，对治疗方案的制定有很高的准确度。可以看到，kNN(1)的总体准确率几乎超过了参与统计的医生平均水平，而经验最丰富的医生的决策水平是我们希望推荐模块能达到的目标状态，相比之下还需要进一步提高算法的性能。

## 推荐结果融合

在推荐模块中，系统根据不同的推荐方式，分别提供了三种推荐结果：

* 基于NCCN规则的推荐
* 基于RJ规则的推荐
* 基于相似病例的*k*-NN推荐

NCCN规则和RJ规则都是根据大量临床诊疗病例的经验，经过相关的学术研究分析、通过行业内的专家审核而进行定期更新和维护的，是对一般情况的诊断方式进行指导的规则。它表示的是根据当下得到共识的医学知识所得出的结果，被医疗机构大规模使用，并经过了长时间的验证，被当成常识或者准则，因而正确性较高。而*k*-NN推荐算法根据病例的特征寻找相似病例作为参考，和固有不变的规则相比，可以从病例中学习到新的规律。然而，虽然它合理有效且具有很大的参考价值，但从根本上而言，却只是一种基于相似病例应该使用相似诊疗方案的假设，只是一个没有经过验证的猜测。

对于这种多方法混合推荐的情况，必须面对两个最基本的问题：

* 如果结果相互支持，应该如何进行融合
* 如果结果相互冲突，应该如何进行抉择

基于对这三种推荐方式各自优点和缺点的认知，如果能找到一种方法解决上述问题，对它们的结果进行融合或是抉择，这将会为临床诊疗方案的制定提供一个更具有参考意义的综合结果。

### 推荐结果的粒度

在计算机科学领域，粒度（Granularity）的粗细描述的是数据的相对大小或粗糙程度。在这里，我们用粒度的概念来表示推荐结果所包含的信息量，通俗地说，也就是的它所描述的内容的详细程度。粒度越细，提供的信息越详细，所具有的指导意义越高；粒度越粗，提供的信息越模糊，所具有的指导意义越低。

表4-1列出了*k*-NN、RJ规则和NCCN规则对化疗方式的推荐结果，以及它们和具体所用的化疗药物的映射关系。从中可以看到，在这三种推荐方式的结果中，*k*-NN的推荐结果粒度最细，它的推荐结果几乎可以和具体的化疗药物一一对应；RJ规则的结果次之，它将A-P-C这种方式也归到“其他”这一大类中去了；而NCCN规则的结果粒度最粗，其结果只包含了“是否”需要化疗的信息。

对于使用一种推荐方法得到一定的方案选项的推荐结果，我们记作R=(m, o)。其中，R表示推荐结果（result），m表示推荐方法（method），o表示方案选项（option）。粒度是一个定性而非定量的衡量指标，虽然不同的粒度之间存在大小关系可以进行比较，但是没有严格规定过一个明确的数值定义。对于一种类型的治疗方案而言（例如化疗），用表示这种治疗所有可选的具体方案的集合，我们可以给粒度一个明确的定义。

参照表4-1，给定一个不同推荐方法下的方案映射表T，就可以定义函数表示使用推荐方法得到的结果所对应具体方案的集合。

|  |  |  |
| --- | --- | --- |
|  |  | (4‑15) |

其中表示映射表T中对应推荐方法m的列中，对应选项o所代表的行。例如，。

所包含的元素越少，代表推荐结果排除掉的方案越多，同时也说明了这个推荐结果更具有参考价值。例如，而，显然前者的粒度更细，更具有指导意义。

在这个启发下，可以定义推荐结果的粒度如下：

|  |  |  |
| --- | --- | --- |
|  |  | (4‑16) |

表4-1不同推荐方式中化疗方案映射表

Table 4-1 Mapping of CT schemes among different recommendation methods

|  |  |  |  |
| --- | --- | --- | --- |
| **CT Schemes** | **k-NN** | **RJ Rule** | **NCCN Rule** |
| TEC | TEC | TEC | Need |
| EC-T | EC-T | EC-T |
| TC | TC | TC |
| TCb | TCb | TCb |
| CMF | AC/EC/CMF | AC/EC/CMF |
| EC(AC) |
| ddAC-ddP | ddAC-ddp/ddAC-ddwp | ddac |
| ddAC-wp |
| EC-wP | EC-wP | EC-wP |
| EC-wPCb |
| A-P-C | A-P-C | Others |
| CEF-T | CEF-T/CEF-P |
| CAPE | Others |
| P |
| PCb |
| No Need | No Need | No Need | No Need |

其中，粒度的值满足，的值越大，表示粒度越细（fine），结果越详细，包含的信息越丰富；反之，的值越小，表示粒度越粗（coarse），结果越粗糙，包含的信息量越少。

### 推荐结果之间的支持度

同样基于这个辅助函数，对于两种不同的推荐方法的推荐结果和，我们可以定义它们之间的相互支持程度，不妨称为“支持度”：

|  |  |  |
| --- | --- | --- |
|  |  | (4‑17) |

其中，而表示集合的交集（intersection）运算，表示集合的基数（cardinality）运算。观察公式可以发现，其实支持度正好表示在概率均等的情况下，两个集合中对应的具体方案相同的概率。

### 推荐结果的综合评估

为了得到一个更具有参考意义的综合结果，我们的需要考虑的因素同时包括了可信度、粒度和其他推荐结果的支持度。其中的原因在于：

* 可信度高：对当前推荐结果的准确性更有把握。
* 粒度高：推荐结果的内容更具有参考价值。
* 支持度高：推荐结果和其他方法的结果一致性高，其正确性得到了更多的佐证。

因此，对于推荐结果的定义综合评分如下：

|  |  |  |
| --- | --- | --- |
|  |  | (4‑18) |

其中关于，如果*m*为*k*-NN，，如公式4‑14所示；否则，根据规则的常识性，的值恒为1。观察公式4‑18，对于推荐结果来说，综合评分不仅仅包括了这个推荐结果本身的可信度和粒度，还同时兼顾了其他推荐结果的作用。

举个两个实际数据中的例子，例1的最终化疗方案是EC-T，例2的最终化疗方案为TC。这两个例子中，不同推荐方案产生的结果，以及前文提到的相关评价指标的值如表4-2所示。

对于例1对应的病例，k-NN推荐模块给出了一个可信度高达90%的推荐结果，认为化疗方案应该选用EC-T。同时，RJ规则和NCCN规则给出了粒度更粗的推荐结果，且相应的中均包含了EC-T，给k-NN推荐模块的结果提供了正面的支持。因而，在最后归一化的综合评分中，k-NN的结果凭借其细粒度、高可信度和较高的一致性，打败了其他的推荐结果，准确地得到了EC-T这个正确答案。

而对于例2所示的病例而言，k-NN推荐模块的首选结果是EC-T，但是可信度仅为49%。同时，RJ规则给出了不同的推荐结果，相应的中并未包含EC-T。与例1中的情况比较，RJ规则在例2中的推荐结果包含的选项更少，所以相比之下具有更细的粒度。而NCCN规则依旧给出了一个粒度很粗的推荐结果。按照相应的计算公式，最后归一化的综合评分中，RJ规则以微弱的优势超过了k-NN，成为了综合推荐结果，而它的结果中所包含的TC方案，正是应该选择的最佳化疗方案。

对比两个例子可以发现在本文提出的方法中，当k-NN的推荐结果附带较高可信度的时候，它的绝对优势很难被压倒，通常即为综合推荐结果；而当k-NN的推荐结果可信度较低时，基于规则的推荐结果可以对其产生一定影响，最终虽然牺牲了粒度，却得到了一个更为全面的综合结果，减少了错误推荐的发生。

表4-2综合推荐结果的两个例子

Table 4-2 Two samples of synthetic recommendation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Sample 1 (**EC-T**) | | | Sample 2 (**TC**) | | |
|  | k-NN | RJ Rule | NCCN Rule | k-NN | RJ Rule | NCCN Rule |
|  | **EC-T** | EC-T EC-wP TCb Others | Need | EC-T | **TC** AC/EC/CMF Others | Need |
|  | 1 | 10 | 15 | 1 | 8 | 15 |
|  | 0.902 | 1.000 | 1.000 | 0.491 | 1.000 | 1.000 |
|  | 0.938 | 0.375 | 0.063 | 0.938 | 0.500 | 0.063 |
|  | 0.973 | 0.660 | 0.144 | 0.630 | 0.657 | 0.126 |
|  | **54.7%** | 37.2% | 8.1% | 35.5% | **37.0%** | 7.1% |

为了进一步验证以粒度作为参考依据对推荐结果进行综合的有效性，我们在所有讨论数据上进行了实验，结果如图4-5所示。其中Merge表示对各推荐方法的结果进行综合考虑后融合的结果，可以看到在准确率方面的表现和预期吻合，比k-NN和RJ规则均略有提升。粒度代表了推荐结果包含的信息量，表示以其作为参考所具有的价值，实验中分别根据各方案的推荐结果取平均值求得。可以看到虽然NCCN的准确率依然最高，但是由于其粒度显著地低于其他方法，其参考价值不高；而融合后的结果在获得了较高准确率地同时，仍然保持了较高的粒度，相对NCCN的结果更有指导意义。

图4-5 各推荐方法和融合结果的准确率和粒度比较

Fig.4-5 Comparison of accuracy and granularity among different recommendation

## 本章小结

本章对智能诊断系统的推荐模块进行了具体的介绍，包括涉及到的基于规则的推荐模块和基于病例的推荐模块。此外，本章还介绍了不同的推荐方法得到的结果的粒度不同的现象，并且提出了综合不同粒度的推荐结果，从而提高预测准确率的方法。

# 适用于概念漂移的推荐算法

## 引言

在本章，我们基于k-NN提出了一种自适应的分类器，用于处理真实场景中普遍存在的概念漂移的现象。特别地，本章最后还对分类器预测的准确度进行了可视化的展示，并对比分析了不同参数配置下的性能和其原因。

由于自适应的分类器是基于k-NN分类器的基础实现的，它完整地继承了k-NN模型的非参数统计特性，因而不用考虑输入数据的特征分布等统计特性，可以通过直接简单地调整k值等参数，使得分类器达到较好的准确率。在概念漂移的情景中，数据集较小的情况下，它可以通过检测概念漂移的发生，学习其中的规律进而更新旧样本，以实现在不丢弃旧样本的情况下能够保持较高的准确率。得益于其易用性和不错的准确率，这种自适应的k-NN分类器可以被应用在各种场景之中。

## 概念漂移

在机器学习领域中，概念漂移 [44]一般表示目标变量的统计学属性，例如其服从的分布，或是分布的参数等，潜移默化地发生不可预见的变化的过程。这个现象会导致人们通过模型进行预测的时候，准确度会随着时间的流逝而逐渐下降。事实上，概念漂移有时候也会指目标变量以外的变化现象，例如模型输入的训练数据的分布随着时间的无规则变化 [45]。在本文中，概念漂移这个术语仅限于表示目标变量相关的变化过程。

对于几乎所有概念漂移的检测方法，都有一个不可避免的问题，那就是如何权衡“假警报”和“慢响应”这两个方面的弊端 [46]。如果检测方法过于灵敏，数据中无处不在的噪声就可能引起概念漂移的误报；相反，如果检测方法过于保守，则会导致概念漂移现象发生之后，过了很长一段时间才被发现。不管是概念漂移的误报还是迟报，都会导致当前的分类器工作在一个错误的概念之下，从而导致预测行为的准确度下降。

许多领域的学者都在这个课题上都有一些研究，特别是流式数据挖掘相关的研究者们，他们进行了大量的研究，并且对于一些重点问题，例如如何检测概念漂移的发生，如何对模型或者数据进行调整以适应最新的概念等，均提出了各种各样的解决方案。其中有的研究者对不同种类的概念漂移进行了归纳，并对现有的算法整理并给出了详尽的评述 [44]。还有研究 [47]详细介绍了如何更新模型以适应新的概念，并展示了这个方向上的一系列最先进的算法。

FLORA系列算法 [48]是最早的针对变化数据的监督式增量学习系统之一，它利用了滑动窗口技术，采用了一个FIFO的数据结构来存储最新的样本。为了使模型适应最新的概念，FLORA算法不断地用最新的样本来更新模型参数，并且不断地丢弃旧的样本。对于这一系列的算法，最主要的挑战是如何去选取一个合适大小的滑动窗口。使用一个长度较小的窗口，可以使模型更准确地反映出当前的概念，但是在概念没有发生变化的稳定阶段，它牺牲了算法的准确性等性能。相反，使用一个长度较大的窗口，模型对概念漂移的灵敏度会下降，但是可以提高算法在稳定阶段的性能。

样本加权是应对概念漂移的另一类方法。直观上，随着时间的流逝，一个样本从新到旧，它反映当前的概念的能力应该是逐步下降的，因此对模型的重要性也呈递减变化。实际上，使用窗口的方法也可以被看作样本加权方法的一种特例，这种情况下，在窗口之中的样本权重为1，窗口之外的样本权重为0。权重随着时间衰减的函数是这类方法的一个重要的部分，线性程度下降 [49]和指数程度下降 [50]的函数都有人在使用。

对于检测概念漂移的算法而言，如何权衡“假警报”和“慢响应”这两个方面是一个不得不面对的问题。如果算法过于敏感，错误的警报可能会在没有发生概念漂移的时候产生；反之如果算法过于保守，概念漂移的现象可能在发生之后很长时间才被发现。“假警报”会导致模型错误地去适应一个根本不存在的新概念，而“慢响应”则导致模型在过时的概念下进行分类，两者都会使分类器的准确度下降。

## 自适应分类器

有研究者为了处理概念漂移的现象，在充分考虑到了灵敏程度的影响下，提出了将两个灵敏度不同的分类器组合起来，作为一个综合的分类器来使用的方法 [51]。这个方法使用灵敏的分类器作为概念漂移的指示器，同时使用另一个较为稳定的分类器用于预测。

### 概念敏感的探测器

基于这种思想，为了使我们的-NN分类器可以对新的概念进行自适应，我们同样可以给-NN分类器上附加一个对概念敏感的探测器，例如使用一个具有固定大小的滑动窗口的同类-NN分类器。我们使用来表示这个自适应的分类器（Adaptive classifier），用*CSD*sw来表示一个窗口大小为的对概念敏感的探测器（Concept-sensitive detector）。以所有的历史样本为参考，对新样本进行分类，而*CSD*sw仅仅使用最近的个样本作为参考。

如算法5-1所示，分类器的输入包括一个包含个样本的数据集，初始训练样本数，探测器的滑动窗口大小，用于计算的最近邻居个数。为了初始化和*CSD*sw，算法首先使用作为训练样本对两个分类器分别进行训练（行3~6）。虽然两者使用同样的样本集合进行训练，但是*CSD*sw和不同，当训练样本的数量超过之后，它会在训练新样本的时候自动忘记最早的样本。经过初始化之后，对于一个新来的样本，算法会使用分类器进行预测（行8~9），同时，两个分类器都会对这个样本进行学习（行14~15）。在预测的过程中，如果算法检测到了概念漂移的发生，那么用于预测的分类器将会被更新，从而使得它在新的概念下依然能有效地进行预测（行11~13）。

自适应分类器利用了整个历史数据的知识作为判断依据，而概念敏感的探测器代表的则是最新的数据的知识。这一算法的有效性，或者说设计该算法的初衷，是基于以下两点假设的：

* 基于更多的知识，分类器的预测结果将会更具有可信度。
* 基于更新的知识，分类器的预测结果将会更具有适应性。

平时，由于同时包含了丰富的历史数据和最新的数据作为参考，分类器的准确率要高于的表现。在概念漂移刚开始发生的阶段，新的样本被基于旧的概念进行分类，在和上都可以观察到明显的性能下降。但是可以很快地将过时的样本用新概念下的样本替换掉，从而使得它的准确率能快速地回升到正常的水平。基于这一点，可以通过监视和这两个分类器的准确率，以观察到概念漂移的发生。简单来说，如果的准确率在一段时间内持续凌驾于之上，我们就可以断言有一定程度的概念漂移现象已经发生了，因此需要对分类器学习到的历史样本进行维护，以适应最新的概念，从而能提供更准确的预测功能。

|  |  |
| --- | --- |
| **Algorithm** **5-1** Adaptive k-NN Classifier | |
| **Input:** | , , , |
|  | : training data |
|  | : initial training sample number |
|  | : sliding window size of concept-sensitive detector |
| **Output:** | Predicted class for each object |
|  |  |
| 1: | an adaptive -NN classifier |
| 2: | a concept-sensitive detector of window size |
| 3: | **for**  **to** **do** |
| 4: |  |
| 5: |  |
| 6: | **end for** |
| 7: | **for** **to** **do** |
| 8: |  |
| 9: | Output |
| 10: |  |
| 11: | **if** concept drift detected **then** |
| 12: |  |
| 13: | **end if** |
| 14: |  |
| 15: |  |
| 16: | **end for** |

### 冲突记录表

为了检测概念漂移的发生，一个朴素的方法是直接在一系列的样本中比较和的准确率，如算法5-2所示。这个算法维护了一个大小为的循环列表，以记录在最近个样本中，分类器出错但是没有出错的频繁程度。我们使用一个阈值作为警示线，一旦这类样本超过了一定的比例，就认为它们被一个新的概念所影响，发生了概念偏移的现象。

|  |  |
| --- | --- |
| **Algorithm** **5-2** Accumulative Misclassification Detector | |
| **Input:** | , , , , , |
|  | : a circular list of bits representing misclassification |
|  | : size of list |
|  | : threshold for judging whether concept drift occurs |
|  | : class predicted by |
|  | : class predicted by |
|  | : true class of the current entry |
| **Output:** | Whether concept drift is detected. |
|  |  |
| 1: | **if** **and** **then** |
| 2: |  |
| 3: | **else** |
| 4: |  |
| 5: | **end if** |
| 6: | **if** **then** |
| 7: | **return** |
| 8: | **else** |
| 9: | **return** |
| 10: | **end if** |

这个方法简单易懂，但是有一个致命的缺点。如果训练样本中，属于类别的样本比例小于阈值，那么在类别上发生的概念漂移就永远不可能被单独发现。为了方便理解，可以看下面这个例子。考虑训练集合，对每个样本而言只有一个属性，满足公式5‑1所示条件。

|  |  |  |
| --- | --- | --- |
|  |  | (5‑1) |

假设样本服从均匀分布，过了一段时间之后，概念发生漂移，之后的样本满足条件。显然在新的概念下，特征空间中一些的样本的类别发生了变化，从类别1变成了类别2，所以很有可能将的样本错误地分类。但是由于理论上这些样本在数据集中的比例期望只有，一个值为的，就可以很轻易地使这样的概念漂移变得难以察觉，尽管这是一个很明显的概念漂移。

### 概念漂移的检测和自适应

在算法5-2中，由于对所有的样本使用了同一个共享的循环列表作为冲突记录表，使得它无法检测到局部小范围特征空间中发生的概念漂移。针对这个问题，直观上我们可以把整个特征空间划分成一个个小的区域，在每个区域上维护一个单独的循环列表作为预测错误记录表。更极端地延伸这个想法，为了使每个区域尽可能小，需要划分出尽可能多地区域，我们甚至可以对每一个历史样本都维护一个单独的循环列表。在分类器预测的过程中，我们对每一个历史样本进行追踪，观察是否和当前预测的样本的结果有冲突，进而决定是否因为发生了概念漂移而需要将训练集中的旧样本的类别进行更新，为后续的样本提供更准确的预测。另一方面，如算法5-3所示，它在检测概念漂移的同时也对模型中的历史样本进行了更新和修正，同时实现了概念漂移的检测和自适应。

对于样本而言，如果分类器AC错误地将其分类，而*CSD*sw却做出了正确的预测（行4），那么AC分类器做出错误的预测所依据每一个最近邻，将会被标记与当前样本冲突（行5）。对于AC分类器中的每一个历史样本，如果它的冲突记录表中的条目满足了一定的条件（行6），例如被观察到频繁地与最近地样本冲突，那么它的属性特征所在的区域将会被认为发生了概念漂移。在这种情况下，AC根据冲突记录表，将这个历史样本的类别更新为符合最新概念的类别（行8）。

冲突记录表的本质使一个字典，它用样本产生的时间作为字典的键。对一个历史样本而言，它在冲突记录表中对应的记录包含了在未来的预测过程中使用该样本作为最近邻但是却得到了错误结果的样本和它们所产生的时间。冲突记录表中的任一项定义如下：

|  |  |  |
| --- | --- | --- |
|  |  | (5‑2) |

注意到在公式中有这样的限制，一旦对一次正确的预测做出了贡献，为了最大程度地避免“假警报”现象，在冲突记录表中对应的条目将会被清空（行11）。

|  |  |
| --- | --- |
| **Algorithm** **5-3** Sample-based detection and adaptation method | |
| **Input:** | , , , , , , |
|  | : training dataset of |
|  | : a dictionary of confliction records for training data |
|  | : class predicted by |
|  | : class predicted by |
|  | : current entry |
|  | : returns a set of ’s nearest neighbors |
| **Output:** | Whether concept drift is detected. |
|  |  |
| 1: |  |
| 2: |  |
| 3: | **for** **do** |
| 4: | **if** **and** **then** |
| 5: |  |
| 6: | **if** satisfies concept-drift indicator then |
| 7: |  |
| 8: | apply adaptation for |
| 9: | **end if** |
| 10: | **else** |
| 11: |  |
| 12: | **end if** |
| 13: | **end for** |
| 14: | **return** |

在算法中出现的概念漂移指示器（concept-drift indicator）是这个算法的关键，它直接决定了断言概念漂移现象发生的条件，从而也决定了是否要更新历史样本（行6）。

本文中，对该自适应分类器分别使用了3种不同的指示器：

**基于规则。**直观上，当中反复出现一个类别且仅出现这个类别的时候，我们认为这是一个概念漂移的讯号。基于规则的指示器从这些启发式的规律中抽取规则，作为判断的依据。例如我们可以计算在一个时间窗口内，每一个类别在中出现的频率，通过一个阈值来做出判断。

**基于熵。**如果用来表示中包含的类别列表（list of classes），我们就可以用列表的大小以及列表中元素的熵，作为一个衡量依据。给定一个关于大小的阈值和一个关于熵的阈值，概念漂移可以通过下列条件来判断。

|  |  |  |
| --- | --- | --- |
|  |  | (5‑3) |

**基于评分。**在整个算法中我们最关心的是当前和将来的样本的类别的预测，因此对于冲突记录表中C[t]所包含的内容，时间越近的条目相对而言具有越高的价值。此外，由于过于迅速的响应是“假警报”的主要原因之一，对于一个历史样本，在它产生后很短一段时间之内的冲突不应该过大地影响到这个历史样本本身。因此我们可以根据样本产生的时间去设置权重，从而得到一个和产生时间正相关的评分标准，作为指示器判断的条件，例如：

|  |  |  |
| --- | --- | --- |
|  |  | (5‑4) |

如果中某一个类别的评分显著高于其他类的评分，我们断言受到了概念漂移的影响。

## 系统实验

### 衡量指标

为了比较不同分类算法的性能，我们使用两种衡量指标。

**累积准确率**（Accumulative accuracy）用于表示从分类器初始化开始到当前时刻，每个样本都基于之前的历史样本进行分类预测，这种情况下所有样本的预测准确率，定义如公式5‑5所示，其中N\_init表示训练样本集合的初始大小。

|  |  |  |
| --- | --- | --- |
|  |  | (5‑5) |

**后续准确率**（Next-n accuracy）用于表示在当前时刻，分类器基于此前所有历史样本进行学习，对接下来n个样本的预测准确率平均值，定义如公式5‑6所示。

|  |  |  |
| --- | --- | --- |
|  |  | (5‑6) |

累积准确率的曲线揭示了分类器在长期学习过程中的性能变化趋势，而后续准确率则表现了短期内的准确率变化情况。

### 比较算法

**-NN**：这是一个最基本的近邻分类器，当作基准用于性能比较。在试验中，它将使用我们之前提到的*modified overlap*作为相似度度量方式。

**-NNSW**：这是一个基于-NN扩展的分类器，作为处理概念漂移的分类器的基准。它使用一个固定大小的窗口，在选择近邻作为参考依据的时候，仅考虑最近一段时间内的样本，而不是所有的历史样本。

**A-NN**：这是我们基于算法5-1提出的自适应-NN分类器。对于算法5-3中提到的细节，我们实现了三种不同的概念漂移指示器，并在后续实验中进行了比较。

### 数据集

为了测试和评估本文提出的算法的性能，我们构造了一个模拟数据集，在数据中人为制造了概念漂移现象。此外，我们还准备了一个真实的医疗数据集，并且在该数据集上使用分类器对乳腺癌患者进行化学辅助治疗方案的推荐。

**SD**（Simulated Dataset）是一个随机生成的模拟数据集，其中的数据共有4个类别，每一个样本有6个离散类型的属性，值域为。我们随机生成了3000个样本，并且将它们按时间跨度划分成了两个部分，分别服从于两个不同的概念。前1000个样本的类别根据样本的属性通过线性函数计算得出，后2000个样本则改为使用另一个不同的线性函数计算。

**BCCD**（Breast Cancer Chemotherapy Dataset）是一个从SJTU Breast Cancer Database中的2770个真实病例中生成的数据集。数据集中的每一个样本对应了数据库中的一个真实的乳腺癌病例，每一个样本有12个离散类型的属性用于表示原始的医疗指标，整个数据集中的样本分别属于8个不同的类别，对应8种不同的化学辅助治疗方案。

这两个数据集的概况如表5-1所示，为了更清晰地表现出概念漂移发生的时间点，我们将两个数据集中的类别分布情况在图5-1和图5-2中可视化地展现出来了。

C:\Users\sechs\AppData\Local\Microsoft\Windows\INetCacheContent.Word\fake_dist_all_min.tifC:\Users\sechs\AppData\Local\Microsoft\Windows\INetCacheContent.Word\fake_dist_200_min.tif

(a) class proportion till current example (b) class proportion in next 200 examples

图5-1 SD数据集上不同类别样本的比例

Fig.5-1 Class proportion of samples in SD

C:\Users\sechs\AppData\Local\Microsoft\Windows\INetCacheContent.Word\ct_dist_all_min.tifC:\Users\sechs\AppData\Local\Microsoft\Windows\INetCacheContent.Word\ct_dist_200_min.tif

(a) class proportion till current example (b) class proportion in next 200 examples

图5-2 BCCD数据集上不同类别样本的比例

Fig.5-2 Class proportion of samples in BCCD

表5-1数据集概览

Table 5-1 Overview of data sets

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **样本个数** | **属性个数** | **类别个数** | **概念漂移时间点** |
| SD | 3000 | 6 | 4 | 1000 |
| BCCD | 2770 | 12 | 8 | Multiple |

### 结果分析

在模拟数据集SD上，我们使用了之前提到的这些分类器进行实验，分别计算了累积准确率和后续100个样本的后续准确率，实验结果如图5-3所示。

可以看到，-NN和A-NN在预测的时候都使用到了所有的历史数据，因此它们在准确率的平均水平上毫无疑问高于-NNSW。在模拟数据集中，我们提前已经知道在的地方发生了概念漂移，这一点在图中也可以被明显地观察到。如图5-3a所示，由于受到概念漂移的影响，所有的分类器的准确率在的位置开始有所下降。基本的-NN分类器通过学习了将近2000个新概念下的样本，花了很长一段时间才将准确率恢复到了原来的水平，与此相对，-NNSW和A-NN对概念漂移的反应更为明显，它们花了较短的时间就使得准确率大幅度地回升。这个效果在图5-3b表现地更为明显，可以看到-NNSW和A-NN的next-100准确率在概念漂移的时间点之后都有一个快速反弹的趋势，而-NN的上升趋势则较为平缓。

C:\Users\sechs\AppData\Local\Microsoft\Windows\INetCacheContent.Word\fake_w200_score_accum.tifC:\Users\sechs\AppData\Local\Microsoft\Windows\INetCacheContent.Word\fake_w200_score_n100.tif

(a) accumulative accuracy (b) next-100 accuracy

图5-3 SD数据集上kNN，kNNSW，AkNN算法的准确率

Fig.5-3 Accuracy of kNN, kNNSW, AkNN over SD

C:\Users\sechs\AppData\Local\Microsoft\Windows\INetCacheContent.Word\real_w50_rule_accum.tifC:\Users\sechs\AppData\Local\Microsoft\Windows\INetCacheContent.Word\real_w50_rule_n100.tif

(a) accumulative accuracy (b) next-100 accuracy

图5-4 BCCD数据集上kNN，kNNSW，AkNN算法的准确率

Fig.5-4 Accuracy of kNN, kNNSW, AkNN over BCCD

在真实数据集BCCD上的结果看上去略有不同。不同于图5-1，从图5-2中并不能观察到明显的概念漂移现象，真实的情景中概念漂移的过程一般往往是缓慢进行而不是突然发生的，即使是如政策法令限制使用某些药物等之类突发事件，也会有一段缓冲时间，以便各个部门采取相关的行动。

仔细观察图5-2b，可以发现在和附近有相对其他时间更为明显的变化，与之对应的，在图5-4b中确实能观察到相应位置上-NN和A-NN的准确率有一个明显的下降，而-NNSW由于其算法特有的局部性，对这种程度较小的变化没有明显的反应。

得益于内部的概念漂移检测机制，A-NN在准确率的回升速度上略高于-NN。而-NNSW的准确率在从一开始就一直高于其他两个分类器，但是从图5-4a中它的准确率曲线的下降趋势中可以看出它随着样本的增加终将被其他两个分类器打败，也间接暴露出了这个分类器只关注最近的一部分历史样本的这一缺点。

### 参数分析

为了研究CSD的窗口大小和概念漂移指示器的实现方式对A-NN分类器的性能影响，我们在SD数据集上使用这两种参数的不同组合进行了一系列的实验。

对于A-NN和-NNSW，滑动窗口的大小取值范围，概念漂移指示器的实现方式则分别基于规则、熵、评分三种方式来实现。

**基于规则的指示器：**用表示类别在种出现的频率，给定阈值，是否断言概念漂移的发生将会依据公式5‑7所示条件。

|  |  |  |
| --- | --- | --- |
|  |  | (5‑7) |

**基于熵的指示器：**实验中设定了阈值以及。

**基于评分的指示器：**在实验中，每一个类别y的评分计算方式如公式5‑8所示。

|  |  |  |
| --- | --- | --- |
|  |  | (5‑8) |

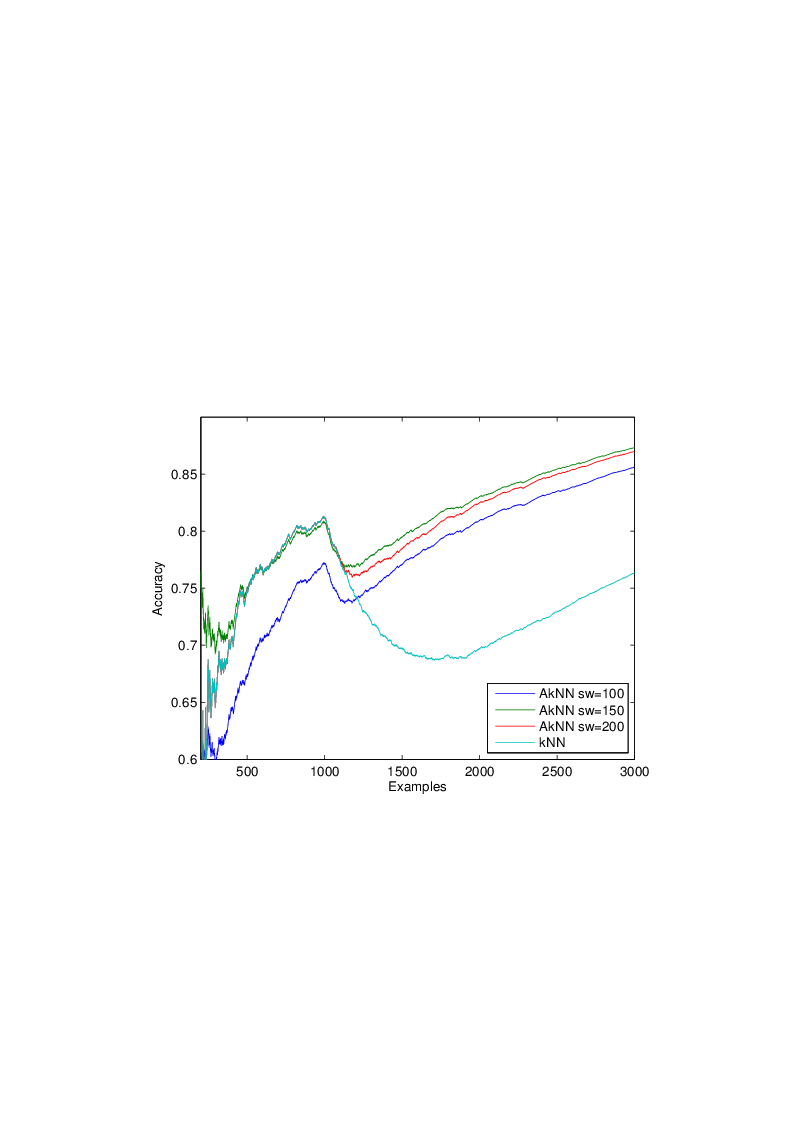
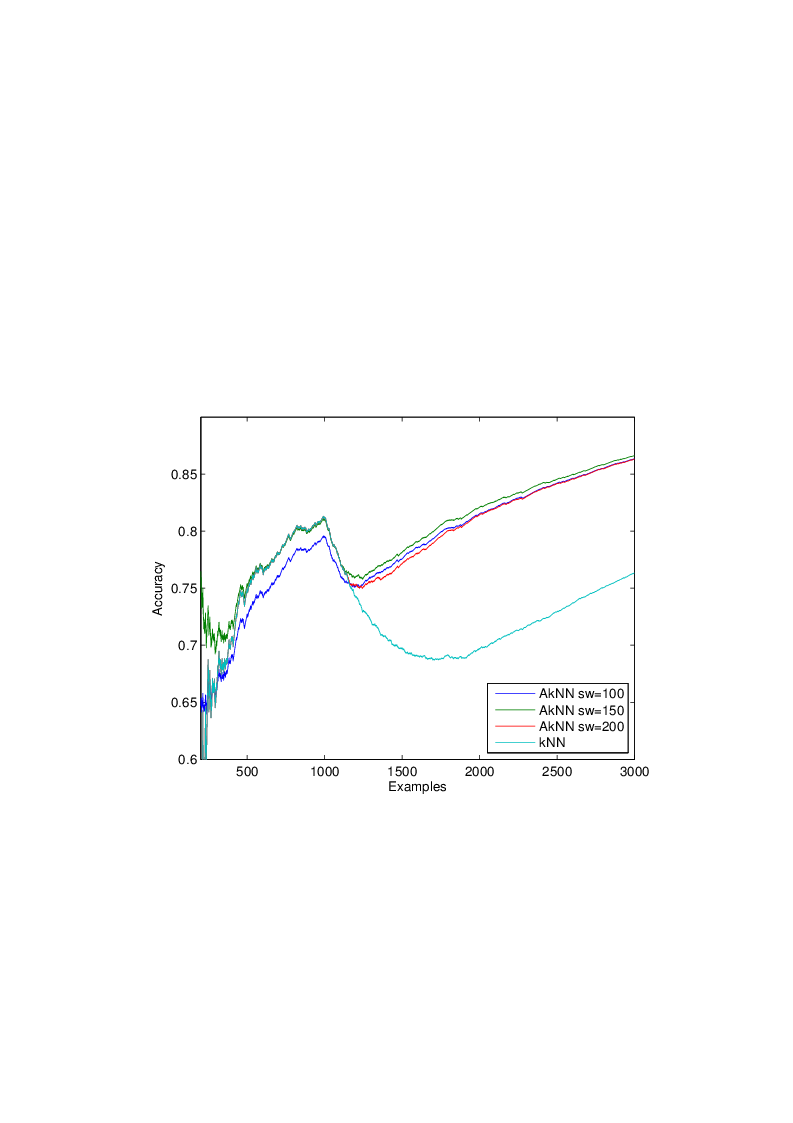
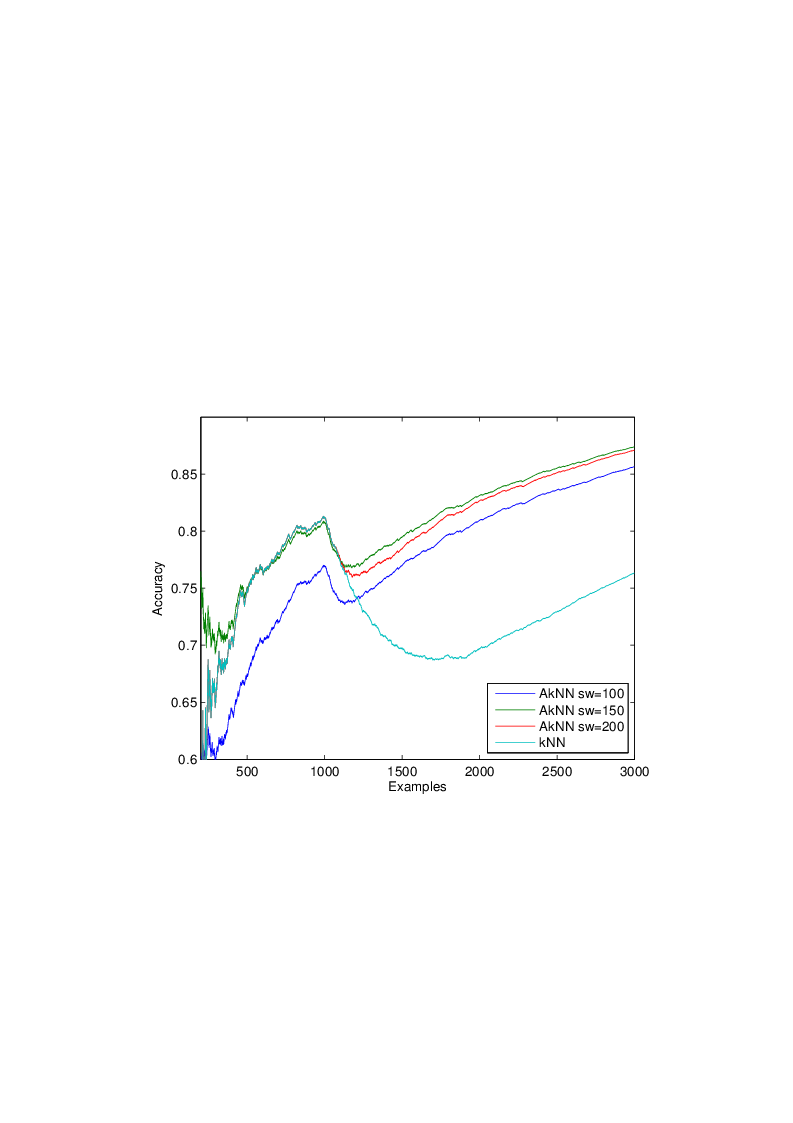
表5-2中展示了不同参数设置下的分类器准确率对比情况，可以看到对于不同的，A-NN总是能达到最高的准确率。此外，为了更清楚地展示不同的参数的影响，我们分组比较了A-NN的累积准确率并以可视化的形式展现在了图5-5和图5-6中。

表5-2 SD数据集上各分类器准确率概览

Table 5-2 Overall Accuracy over SD data set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | -NN | -NNSW | A-NN | | |
| Rule | Entropy | Score |
| sw=100 | 0.7598 | 0.7212 | 0.8565 | 0.8561 | 0.8634 |
| sw=150 | 0.7637 | 0.7451 | 0.8739 | 0.8732 | 0.8662 |
| sw=200 | 0.7634 | 0.7573 | 0.8710 | 0.8699 | 0.8631 |

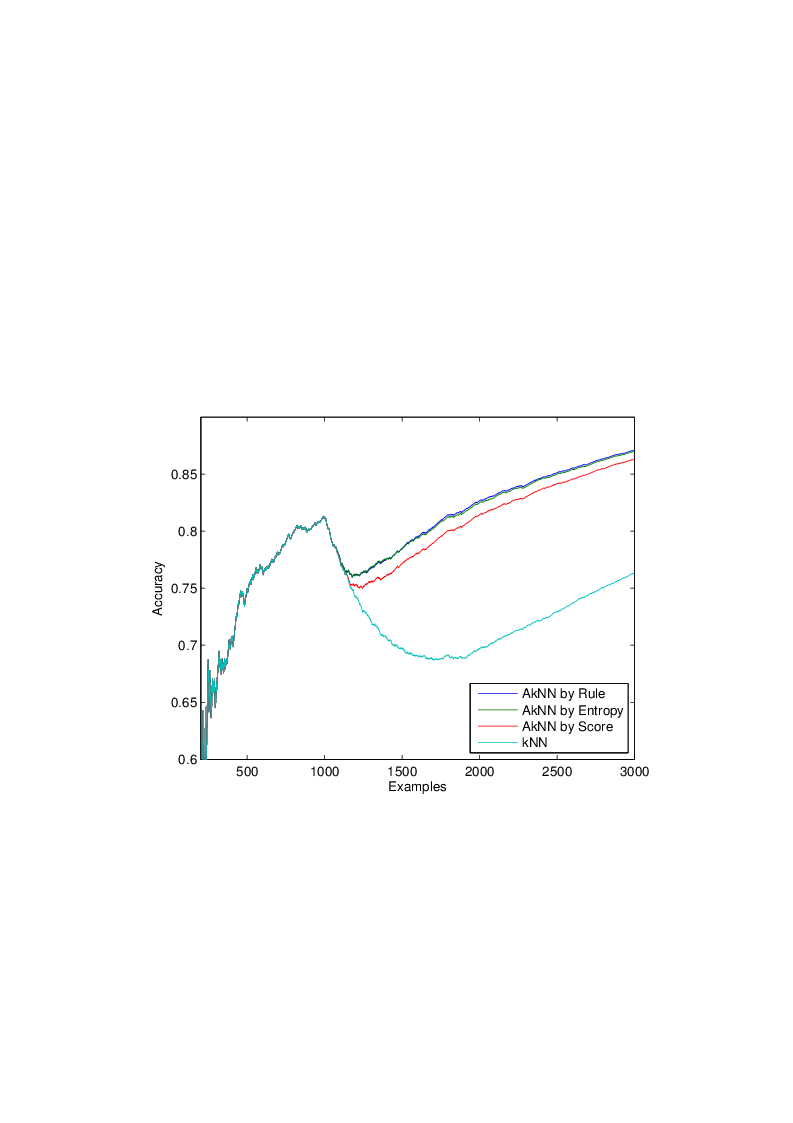
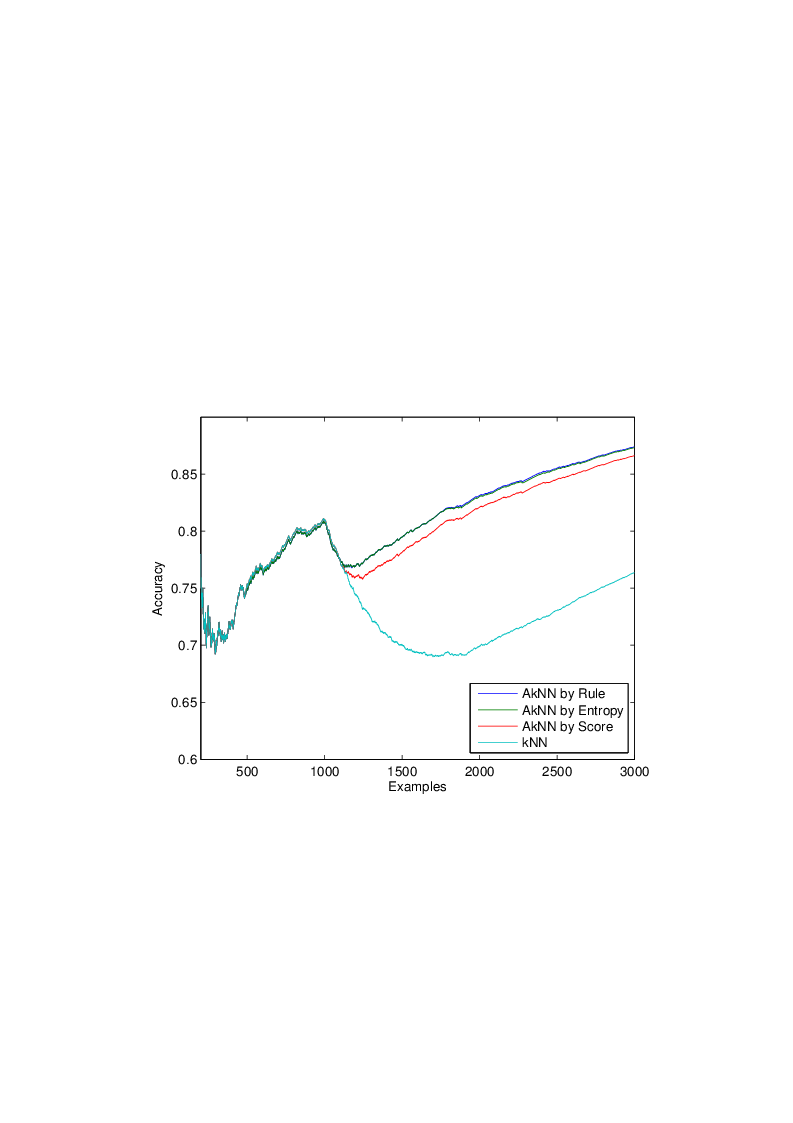
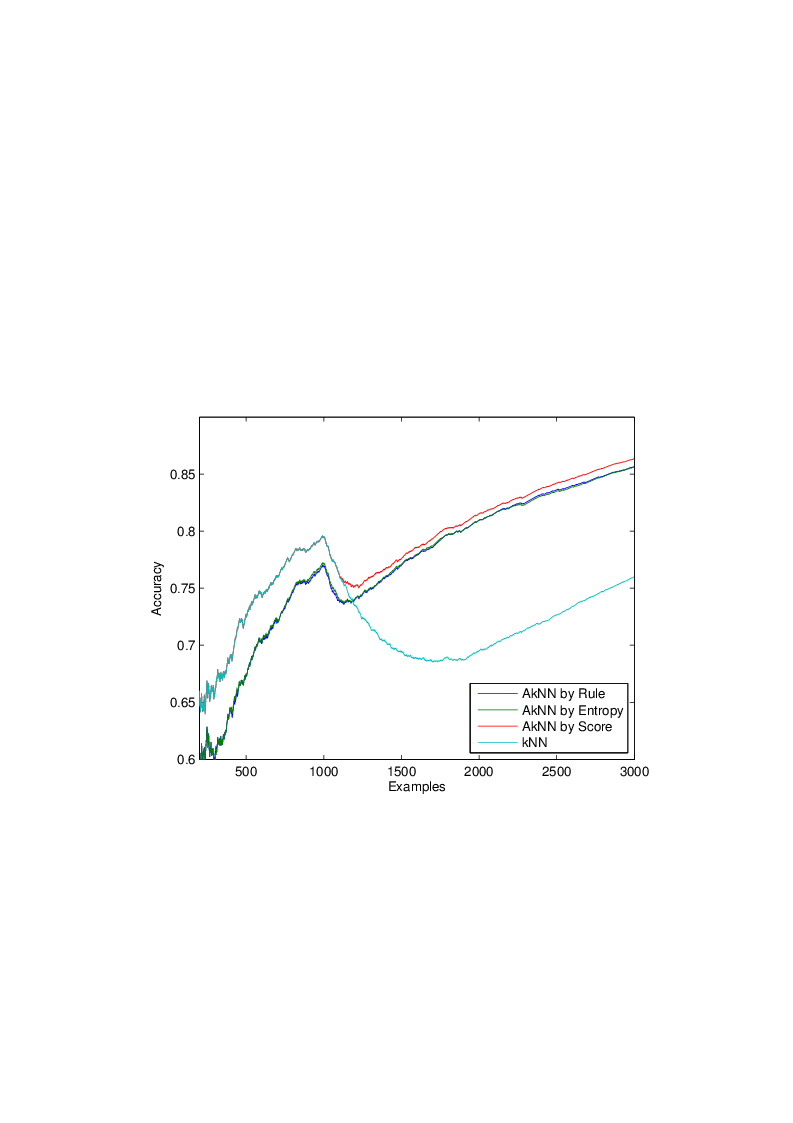
关于窗口大小的影响，一个较小的值可以使ANN分类器对概念漂移的反应更迅速，但是同时也更容易导致“假警报”进而降低准确率；一个较大的sw值使得ANN分类器更稳定，但是同时也降低了它对概念漂移的反应速度。如图5-5所示，在a、b、c三个子图中，ANN（）在一开始的阶段都具有明显偏低的准确率，而ANN（）对概念漂移的反应速度稍慢于ANN（）。



(a) by rule (b) by entropy (c) by score

图5-5 不同CSD窗口大小对累积准确率的影响

Fig.5-5 Accumulative accuracy with different CSD window size



(a) sw=100 (b) sw=150 (c) sw=200

图5-6 不同类型的概念漂移指示器对累积准确率的影响

Fig.5-6 Accumulative accuracy with different concept-drift indicator

不同的指示器实现方式对ANN的影响如图5-6所示。在预期中，基于评分的指示器将会得益于它和时间相关的评分计算方式，从而表现出更好的性能。但是实际上，综合考虑三个子图中的曲线可以发现基于规则和熵的指示器总体上具有更好的性能。不过从图5-6a中可以看到，在这种窗口偏小的配置下，“假警报”相对更容易发生，在这种情况下，基于评分的指示器得益于在开始阶段避免了错误判断而在整体上达到了更高的准确率，这一点是符合最初预期的。

## 本章小结

本章主要介绍为了提高智能诊断支持系统中推荐模块的准确率而进行的研究。本章介绍了概念漂移的概念，详述了现实场景中实际存在的概念漂移现象，并且提出了一种可以处理概念漂移的自适应分类器，用于进一步提高推荐模块的性能。此外，在本章所述的研究中，我们还分别在模拟数据集和真实数据集中进行了相关的实验，并对实验数据进行了比较，对算法的相关参数进行了具体分析。

# 智能诊断支持系统的具体实现

## 相关工具和软件

### 操作系统

Linux是一种自由和开源的Unix-like操作系统，Linux的内核由Linux内核组织（Linux Kernel Organization）负责维护。Linux内核组织是2002年成立的加利福尼亚公共利益公司，向公众免费分发Linux内核和其他开源软件。通常，各社区和厂商将Linux内核打包成供个人计算机和服务器使用的Linux发行版，主流Linux发行版包括Debian、Ubuntu、Fedora、CentOS、Archlinux和openSUSE等。

智能诊断支持系统在Linux操作系统的基础上进行开发，运行在内核版本3.8.0-32以上的Linux发行版之下。

### 数据库

MySQL是一个开源的关系数据库管理系统，由于其免费且具有较高的性能和可靠性，已经成为最流行的开源关系数据库，被广泛地应用在许多中小型网站中。虽然近年来NoSQL飞速发展，特别在大数据相关领域大有取代传统的关系数据之势，但是MySQL由于自身的不断成熟，以及低成本的优势，还是逐渐被用于Google等许多大型的国外网站之中。

智能诊断支持系统中涉及的数据存储，包括工作流的存储、讨论过程的存储和外部的病例数据库，均使用了MySQL作为数据库软件，能在MySQL 5.5及以上版本的环境下正常工作。

### Web框架

Ruby on Rails又被称作为Rails，是一个使用Ruby语言写的开源Web应用框架。Rails基于“不要重复自己”和“约定优于配置”的设计原则，严格按照MVC（Model-View-Control）的结构进行开发，使开发者在使用该框架进行实际的应用开发时，能尽量少写多余的代码。

因此，在开发智能诊断支持系统之时，我们选择了使用Rails作为Web框架。该系统当前稳定运行在Ruby 2.2.1和Rails 4.2.0的环境之下。

### 消息队列中间件

Apache ActiveMQ是Apache软件基金会所研发的开放源代码消息中间件。它使用Java编写，仅依赖于Java虚拟机。

在智能诊断支持系统中，系统在底层与一个工作流系统进行交互，其通信过程需要使用一个消息队列中间件。由于Java虚拟机的跨平台特性，考虑系统到将来移植到其他平台的可能性，我们选择了使用Apache ActiveMQ。

## 相关界面

该智能诊断支持系统已经投入实际使用，病例数据库和智能诊断支持系统的相关界面如下所示。

### 病例数据库的功能界面

病例数据库的基本统计包括了数据库中病例的总体情况，包括按录入时间、病例BMI、年龄等为依据的分布统计，并以柱状图和饼状图直观地展示其中的规律。



图6-12 病例数据库年龄分布统计图示例

Fig.6-12 Distribution of age for patients in database

此外，根据乳腺癌领域的专业知识，系统中还对病例的信息进行自动分析，根据相应的公式和规律统计出特定类型的患者并直观地展示出其分布情况。



图6-13 病例数据库分子分型分布统计图示例

Fig.6-13 Distribution of molecular subtype in database

生存统计分析的结果可以在系统中直接被表示出来，直观地揭示了相关患者的术后生存规律，为医疗研究提供了极大的便利。

图6-14 病例数据库的生存分析图示例

Fig.6-14 Overall/Disease-free survival of patients in database

### 智能诊断支持系统的用户界面



图6-1 系统主页示例

Fig.6-1 Screenshot of homepage the system

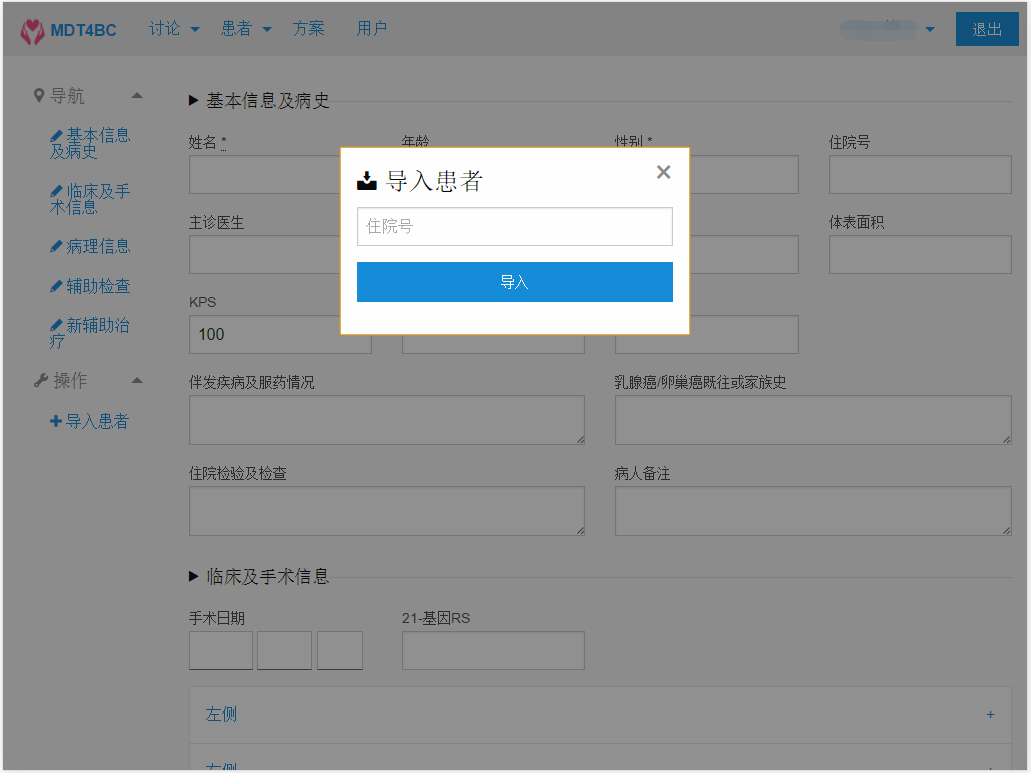


图6-2 从外部数据库导入患者信息示例

Fig.6-2 Screenshot of importing patient information from external database



图6-3 MDT决策过程中投票统计结果示例

Fig.6-3 Screenshot of voting statistics during MDT decision making



图6-4 不同方法的治疗方案智能推荐结果示例

Fig.6-4 Screenshot of different kinds of treatment plan recommendation

# 总结与展望

## 主要工作与创新点

本文的研究旨在利用机器学习领域的相关技术，开发一个基于病例的智能诊断支持系统，用以提高医疗临床诊断的效率、为医学研究提供数据分析支持，从而对医学的发展起到积极的促进作用。

本文相关研究构建了一个标准化的高质量病例数据库，为国内医院的乳腺癌研究提供高质量的数据和服务，为医学研究提供必要的数据支持。

在病例数据库的支持下，本文结合机器学习相关算法，构建了一个基于病例数据的智能诊断支持系统，并将其实际应用到多学科讨论的过程中，为制定完善的综合治疗方案提供支持。在试运行阶段，根据系统产生的数据，本文对推荐结果和实际的讨论记录进行分析，研究了智能诊断系统中，推荐方案的准确度、医生决策的分歧度等统计指标和对多学科讨论过程之间的关系。

为了进一步提高推荐方案的准确度，本文对概念漂移现象进行了详细的介绍，提出了一种适用于概念漂移的推荐算法，并在模拟数据集和真实数据集上做了相应的实验，进行了结果分析和比较。

## 后续研究工作

基于病例的智能诊断支持系统上线运行至今，我们已经获得了一定规模的统计数据，并且在本文中进行了分析和讨论。随着系统的运行，将来将会积累更多的数据以供一些有意义的研究使用。这些研究包括且不限于：

### 分歧度预测

在标准的多学科讨论流程中，每一个用户先后都会进行两次投票。为了评估每次投票中，参与投票的用户之间意见的差异性，可以借用表示混乱程度的熵（Entropy）的概念，用来表示分歧度（Divergence）。

对于一次投票，所有投票的列表记作，用表示该投票选择的方案，其中。投票中出现的不同方案的集合为，用表示每一种方案，其中。显然，V中的元素可以重复，而C中的不能重复。对于某一种方案在本次投票中出现的频率，我们使用来表示：

|  |  |  |
| --- | --- | --- |
|  |  | (7‑1) |

一组投票的结果的混乱程度用Entropy表示如下：

|  |  |  |
| --- | --- | --- |
|  |  | (7‑2) |

考虑Entropy的最值情况：当所有投票结果一致的时候，达到最小值；当所有投票均不同之时，达到最大值。因此，我们将Entropy做归一化处理，使用值域为的Divergence来表示分歧度：

|  |  |  |
| --- | --- | --- |
|  |  | (7‑3) |

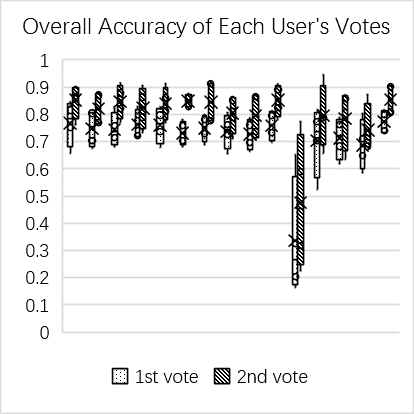


图7-1 用户投票的分歧度和准确度

Fig.7-1 Divergence and accuracy of user’s votes

对于不同类型的治疗方案，我们分别统计了两次投票的分歧度。此外，对于频繁使用该智能诊断支持系统的用户，我们还对两次投票的准确率进行了计算。用户投票的分歧度和准确度如图7-1所示，可以看到经过了集体讨论的中间环节，决议投票的分歧度明显低于初次投票。对于每一位用户而言，虽然他们各自投票的准确率存在差异，但是对比各自的两次投票的准确率，决议投票显然比初次投票更为准确了。

在多学科讨论的过程中，一般会轮流讨论多个病例。由于参与者的体力、注意力等因素在不同的时间是不一样的（例如有的人在会议的前半小时注意力较为集中），病例被讨论的先后顺序和分配到该病例的医疗资源便产生了关系。从投票的环节中统计数据，我们发现了一些规律：对于一些显而易见的病例，投票结果一致性很强；而对于一些情况特殊的病例，投票结果往往分歧较大。为了充分提高效率，在分歧度较大病例中分配更多的医疗资源（讨论时间）是一种值得研究的策略。这样一来，预测多学科讨论中分歧较大的病例，以合理分配更多的医疗资源（如优先讨论），也是一个值得研究的方向。

### 用户行为分析

从统计数据中我们发现参与者在第二次决议投票时，准确率和一致性都比初次投票的高一些。当然，由于最终的综合诊疗方案的制定是参考投票结果的，因而我们无法断言其中的因果关系。但是值得研究的是，系统的推荐结果是否会对参与者的行为造成影响，造成了什么影响。目前在多学科讨论的流程中，参与者可以随时查看系统对当前病例的治疗方案推荐详情。如果限制推荐方案的展示时机，例如在初次投票之前、两次投票之间等，观察两种情况下，初次投票和决议投票的变化，也许是研究推荐结果对用户行为影响的一个很好的切入点。

参 考 文 献

1. Plsek P E, Greenhalgh T. The challenge of complexity in health care[J]. British Medical Journal, 2001, 323(7313): 625.
2. Power D J. Decision support systems: a historical overview[M]//Handbook on Decision Support Systems 1. Springer Berlin Heidelberg, 2008: 121-140.
3. Turoff M, Hiltz S. Computer support for group versus individual decisions[J]. IEEE Transactions on communications, 1982, 30(1): 82-91.
4. Codd E F, Codd S B, Salley C T. Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate[J]. Codd and Date, 1993, 32.
5. Bruce G, Buchanan B G, Shortliffe E D. Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project[J]. 1984.
6. Gonzalez A J, Dankel D D. The engineering of knowledge-based systems: theory and practice[M]. 1993.
7. Ledley R S. Computer aids to medical diagnosis[J]. Jama, 1966, 196(11): 933-943.
8. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective[J]. Artificial Intelligence in medicine, 2001, 23(1): 89-109.
9. 胡育. 基于病历信息的智能诊断技术研究 [D]. 电子科技大学,2015.
10. 顾东晓. 基于案例库的诊疗决策支持技术研究 [D]. 合肥工业大学, 2011.
11. Balakrishnan K, Honavar V. Intelligent diagnosis systems[J]. Journal of Intelligent Systems, 1998, 8(3-4): 239-290.
12. Turkoglu I, Arslan A, Ilkay E. An intelligent system for diagnosis of the heart valve diseases with wavelet packet neural networks[J]. Computers in Biology and Medicine, 2003, 33(4): 319-331.
13. Ligeza A. Logical foundations for rule-based systems[M]. Heidelberg: Springer, 2006.
14. Buchanan B, Sutherland G, Feigenbaum E A. Heuristic DENDRAL: a program for generating explanatory hypotheses in organic chemistry[M]. Stanford University, 1968.
15. Shortliffe E H. Computer-based medical consultations[J]. 1976.
16. McDermott J. R1: A rule-based configurer of computer systems[J]. Artificial intelligence, 1982, 19(1): 39-88.
17. Mitchell T M. Machine learning. 1997[J]. Burr Ridge, IL: McGraw Hill, 1997, 45: 37.
18. Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1(1): 81-106.
19. McGarry K, Wermter S, MacIntyre J, et al. Hybrid neural systems: from simple coupling to fully integrated neural networks[J]. Neural Computing Surveys, 1999, 2(1): 62-93.
20. Cercone N, An A, Chan C. Rule-induction and case-based reasoning: hybrid architectures appear advantageous[J]. IEEE Transactions on Knowledge and Data Engineering, 1999, 11(1): 166-174.
21. Forgy C L. Rete: A fast algorithm for the many pattern/many object pattern match problem[J]. Artificial intelligence, 1982, 19(1): 17-37.
22. Schank R C, Abelson R. Scripts, goals, plans, and understanding[J]. 1977.
23. Schank R C. Dynamic memory: A theory of learning in people and computers[J]. 1982.
24. Lebowitz M. Generalization from natural language text[J]. Cognitive Science, 1983, 7(1): 1-40.
25. Lebowitz M. Memory-based parsing[J]. Artificial Intelligence, 1983, 21(4): 363-404.
26. Kolodner J L. Maintaining organization in a dynamic long‐term memory[J]. Cognitive science, 1983, 7(4): 243-280.
27. Aamodt A, Plaza E. Case-based reasoning: Foundational issues, methodological variations, and system approaches[J]. AI communications, 1994, 7(1): 39-59.
28. Kolodner J L, Wills L M. Case-based creative design[J]. AISB QUARTERLY, 1993: 50-50.
29. Jarmulak J, Kerckhoffs E J H, van’t Veen P P. Case-based reasoning for interpretation of data from non-destructive testing[J]. Engineering Applications of Artificial Intelligence, 2001, 14(4): 401-417.
30. Sabater J, Arcos J L, López de Mántaras R. Using rules to support case-based reasoning for harmonizing melodies[C]//Multimodal Reasoning: Papers from the 1998 AAAI Spring Symposium. 1998: 147-151.
31. Marling C R, Petot G J, Sterling L S. Integrating case-based and rule-based reasoning to meet multiple design constraints[J]. Computational Intelligence, 1999, 15(3): 308-332.
32. Mitra R, Basak J. Methods of case adaptation: A survey[J]. International journal of intelligent systems, 2005, 20(6): 627-645.
33. Kinley A. Learning to improve case adaptation[D]. Indiana UniverSity Bloomington, 2001.
34. Altman N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175-185.
35. Siegel S. Nonparametric statistics for the behavioral sciences[J]. 1956.
36. Richardson L, Ruby S. RESTful web services[M]. " O'Reilly Media, Inc.", 2008.
37. Jula A, Sundararajan E, Othman Z. Cloud computing service composition: A systematic literature review[J]. Expert Systems with Applications, 2014, 41(8): 3809-3824.
38. Lin C, Minglu L, Jian C. ECA rule-based workflow modeling and implementation for service composition[J]. IEICE transactions on information and systems, 2006, 89(2): 624-630.
39. Michelson B M. Event-driven architecture overview[J]. Patricia Seybold Group, 2006, 2.
40. Gradishar W J, Anderson B O, Balassanian R, et al. Breast cancer version 2.2015[J]. Journal of the National Comprehensive Cancer Network, 2015, 13(4): 448-475.
41. Boriah S, Chandola V, Kumar V. Similarity measures for categorical data: A comparative evaluation[J]. red, 2008, 30(2): 3.
42. Mahalanobis P C. On the generalized distance in statistics[J]. Proceedings of the National Institute of Sciences (Calcutta), 1936, 2: 49-55.
43. Stanfill C, Waltz D. Toward memory-based reasoning[J]. Communications of the ACM, 1986, 29(12): 1213-1228.
44. Tsymbal A. The problem of concept drift: definitions and related work[J]. Computer Science Department, Trinity College Dublin, 2004, 106.
45. Widmer G, Kubat M. Effective learning in dynamic environments by explicit context tracking[C]//European Conference on Machine Learning. Springer Berlin Heidelberg, 1993: 227-243.
46. De Houwer J, Hermans D. Automatic affective processing[J]. Cognition & Emotion, 2001, 15(2): 113-114.
47. Gama J, Žliobaitė I, Bifet A, et al. A survey on concept drift adaptation[J]. ACM Computing Surveys (CSUR), 2014, 46(4): 44.
48. Widmer G, Kubat M. Learning in the presence of concept drift and hidden contexts[J]. Machine learning, 1996, 23(1): 69-101.
49. Koychev I. Tracking changing user interests through prior-learning of context[C]//International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems. Springer Berlin Heidelberg, 2002: 223-232.
50. Klinkenberg R. Learning drifting concepts: Example selection vs. example weighting[J]. Intelligent Data Analysis, 2004, 8(3): 281-300.
51. Bach S H, Maloof M A. Paired learners for concept drift[C]//2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008: 23-32.

致 谢

值此论文完成之际，谨向在研究生学习阶段所有支持和关心过我的人们致以最诚挚的感谢。

特别地，我要对我的导师曹健教授表示由衷的感谢。我与曹老师相识已逾四年，他为人谦和、治学严谨、具有很高的人格魅力。曹老师学识渊博但仍有不断学习的热情，事务繁忙却能合理安排时间，这些品质都给我留下了深刻的印象，一直是我向往和努力的方向。在这些年的学习生涯当中，曹老师教会了我许多，大到如何以严谨的态度来进行科研，如何以思辨的方式来分析问题；小至如何成功地做好一个项目，如何清晰地展示自己的想法。本文的内容涉及了我在研究生阶段的许多工作，在本文相关的研究过程当中，曹老师给予了我莫大的指导和帮助，每当我的研究工作陷入瓶颈的时候，他总能给出一些建设性的意见，发现并提出关键的问题，从而使得相关的研究工作得以绝处逢生。与此相似的经历数不胜数，无法一一列举，而它们无疑都使我受益颇丰。

我要感谢我的室友仲一昕同学。在多年一同求学的过程当中，他能一直包容我的各种不足之处，并在各方面不计回报地对我提供帮助，是一个值得信赖的朋友，他的很多优秀的品质都值得我学习。

同时我还要感谢瑞金医院乳腺中心的陈小松医师，在研究过程中时常遇到一些涉及专业医学知识的问题，陈医师总是第一时间提供了详细的解释，为本文的研究提供了很大的帮助。

感谢实验室的所有前辈们，不论是学术上的讨论，还是生活上的交流，对我的各方面都有很大的帮助，和大家一起相处这么久的时光是我宝贵的财富。

最后，我要感谢我开明的父母和女友，他们思想开明并且尊重我的选择，对我的学业给以毫无保留的支持，生活上默默地给我以无微不至的关心。正是有了这样的条件，我才得以安心学习并提升自我，才能完成现在这些微小的工作。

攻读硕士学位期间已发表或录用的论文

* 1. 第一作者. 面向大数据分析的决策树算法[J]. 计算机科学, 2016, 1.
  2. First Author, A Multi-Disciplinary Medical Treatment Decision Support System with Intelligent Treatment Recommendation [C]//Computer and Communications (ICCC), 2016 IEEE International Conference on. IEEE, 2016: 838-842.