

数值诊断的统计方法

蒋大宗 张文修 编著

陕西科学技术出版社

R319

5

3

数值诊断的统计方法

蒋大宗 张文修 编著

132005/21



陕西科学技术出版社

A 895193

内 容 简 介

计算机数值诊断是最近发展起来的生物医学工程的一个重要分支，它是利用统计数据处理技术和电子计算机对于不同的疾病做出定量的诊断。本书紧密结合计算机数值诊断技术，介绍了常用的统计方法，如贝叶斯可能度方法，回归分析与逐步回归方法，判别分析与逐步判别方法，聚类分析、方差分析、信息分析、模糊判决、数量化理论以及某些模拟医生经验的简易方法等，并以肝脏病和心脏病鉴别诊断的实例说明如何利用这些方法选择症状资料和病例，建立鉴别诊断的数学模型。本书深入浅出，是一本系统地介绍数值诊断的著作，适宜于作为生物医学工作者以及医务人员进行计算机数值诊断的科研和教学的参考资料。也可作为医学院学生的教学参考书。

数值诊断的统计方法

蒋大宗 张文修 编著

陕西科学技术出版社出版

(西安北大街 131 号)

陕西省新华书店发行 西安交通大学印刷厂印刷

开本 787×1092 1/32 印张 8.625 字数 180,000

1981 年 10 月第 1 版 1981 年 10 月第 1 次印刷

印数 1—3,500

统一书号: 13202·30 定价: 0.90 元

前 言

计算机数值诊断是利用统计数据处理技术和电子计算机对疾病作出定量的诊断。由于统计中多元分析的发展和电子计算机的广泛使用,使计算机数值诊断迅速发展为生物医学工程的一个重要分支。最近十年来,我国在医学数值诊断方面做了不少工作,例如肝脏病、胃病、心脏病、肺病以及肝病辩证施治的计算机实现都取得了较好的效果。

目前,计算机数值诊断总的来说还处于研究阶段,所得到的数值计算的具体诊断模型,还需要和医生的临床诊断相配合,作为一种辅助诊断的手段。但是,随着计算机诊断设备的完善化以及计算机诊断软件系统研究的逐步深入,数值诊断将会在疾病诊断中发挥越来越大的作用。

本书比较系统地介绍了建立计算机数值诊断模型中经常用到的一种重要方法,即统计方法。为了易于掌握这些方法,书中始终以肝脏病鉴别诊断为具体实例。可供从事计算机数值诊断工作的同志学习、参考。

第四军医大学第一附属医院内科马挺光副教授详细审阅了原稿,提出了许多修改意见。另外,我们与西安医学院第二附属医院内科黄宗心副教授、罗金燕同志,第四军医大学陈国经、张福臣等同志以及西安交通大学生物医学电子工程教研室马春排、蒙建国同志长期进行肝脏病和心脏病鉴别诊断的合作,无疑对本书的完成有着密切的关系。在此一并表示感谢。

蒋大宗 张文修

目 录

第一章 绪论

§ 1.1 计算机数值诊断的意义	1
§ 1.2 计算机数值诊断的方法	4
§ 1.3 计算机数值诊断的前景	7

第二章 医学资料的汇集与整理

§ 2.1 计算机数值诊断所需要的基本资料	11
§ 2.2 病例选择的方法和要求	18
§ 2.3 病例分析中两个重要的特征数	22
§ 2.4 两组病例的相关性特征	33
§ 2.5 频率的稳定性	38

第三章 贝叶斯定理和最大可能度法

§ 3.1 事件与概率	42
§ 3.2 条件概率与贝叶斯定理	47
§ 3.3 利用贝叶斯公式作鉴别诊断的方法	56
§ 3.4 最大可能度法与评分法	65
§ 3.5 贝叶斯公式在逐次问诊中的应用	72

第四章 回归分析与逐步回归分析方法

§ 4.1 回归分析的意义	79
§ 4.2 一元线性回归分析方法	80
§ 4.3 多元线性回归分析方法	90

§ 4.4 线性方程组的解法·····	101
§ 4.5 逐步回归分析方法·····	111
§ 4.6 用逐步回归分析方法作肝脏病的鉴别 诊断·····	123

第五章 判别分析与逐步判别分析方法

§ 5.1 判别分析的意义·····	131
§ 5.2 用训练迭代法寻求判别式·····	135
§ 5.3 费歇意义下的两类间的判别分析·····	141
§ 5.4 贝叶斯意义下多类间判别分析·····	152
§ 5.5 逐步判别分析方法·····	157
§ 5.6 用逐步判别分析做肝脏病的鉴别诊断·····	168
§ 5.7 用逐步判别分析做心脏病的鉴别诊断·····	170

第六章 数量化理论与方法

§ 6.1 数量化理论的意义·····	174
§ 6.2 数量化理论 I ·····	176
§ 6.3 数量化理论 I 在肝脏病鉴别诊断中的 应用·····	184
§ 6.4 数量化理论 II ·····	189

第七章 模糊集合论与数值诊断

§ 7.1 模糊集合的基本概念·····	195
§ 7.2 隶属原则与数值诊断的直接方法·····	199
§ 7.3 贴近原则在数值诊断中的应用·····	204
§ 7.4 模糊变换与综合评判·····	206

§ 7.5 基于模糊积分的数值诊断模型.....	209
--------------------------	-----

第八章 病例分类和症状选择

§ 8.1 病例分类和症状选择的意义.....	213
§ 8.2 相似系数与距离.....	215
§ 8.3 系统聚类法.....	222
§ 8.4 逐步聚类法.....	231
§ 8.5 模糊聚类的分析方法.....	240
§ 8.6 运用模糊聚类分析压缩临床症状资料.....	247
§ 8.7 利用熵函数压缩和选择症状资料.....	252
§ 8.8 运用方差分析选择症状变量.....	257

附表一 F 分布(一)

附表二 F 分布(二)

第一章 绪 论

§ 1.1 计算机数值诊断的意义

计算机数值诊断是最近十几年发展起来的生物医学工程的一个重要分支。它是利用数学的信息理论、数据处理技术以及电子计算机这个强有力的工具，对病人的症状、体征以及检验结果进行数学加工分析，做出定量（计量或数值）诊断。

计算机的数值诊断与医生的临床诊断不同，临床诊断是医生根据病人的症状、体征和检验的结果等以自己的经验做出最可能的判断，诊断的准确性与医生的学识、专业、经验有着直接的关系。而计算机的数值诊断不同，它仅依赖于收集使用的各个医院的诊断病例和对这些资料处理的方式。历史确诊的病例越多，临床资料越详细，处理分析得当，就可能得到比较好的诊断结果，它要比一位医生只凭个人经验考虑的全面的多。计算机数值诊断是将医生的个人经验以及根据经验诊断的实际，提高到科学的理论体系。

所谓诊断是根据病人身体的各种状态，或者某些局部状态作出判断的行为。由于病人体内蕴藏着各种复杂的器官内部变化，所以表现这种体内各种器官变化的状态变量非常多，而且这些状态变量是用目前医学的手段尚比较难于直接观测到的。因此，完全根据疾病的机理来作出诊断一般来说

难于实现。医学上比较容易做到的是在一定程度上反映这种内部状态而表现于病人的体征和检验的结果。目前，一般的诊断还是凭借这些临床表现所构成的临床资料去判断各器官的状态，根据可能直接观测的临床表现的数值，推断不能直接观测的内部器官状态变量的值。

身体内部器官的状态变量可以处于两种状态，一是健康状态，即正常状态；一是疾病状态，即不正常状态。所谓病名是对某种特定的疾病状态给予的名称。疾病的命名反映了内部器官所处的状态，但它并不就是内部器官状态本身。比如慢性肝炎、肝硬化、肝癌，或者胃炎、胃溃疡、胃癌反映了肝部或者胃部处于疾病状态，但它们究竟处于一种什么状态，病名并没有给出这种细节。事实上，对一般的治疗来讲、确定病名已为治疗创造了基本的前提条件。所以实际的诊断是根据病人的临床表现确定病名，还不是确定病人的内部器官状态变量本身。正确的确定病名，是诊断的基本要求。

对于一个医生来讲，确定一个病人的病名，首先是从病人获取临床表现，比如是否有发热、是否有胃痛、是否有腹泻、是否有血压增高等等。根据这些临床表现逐步缩小诊断范围，再通过有目的化验指标，经过头脑的综合分析进行判断后给出诊断病名。计算机数值诊断过程与医生的临床诊断过程相仿，也是接受病人的临床表现的基础上将临床表现数值化，综合分析得出病名。这时综合分析的工具不是某一个人的头脑，而是电子计算机。综合分析的依据不是某一个人凭个人经验，而是以往各家的大量经验和医院的确诊病例。这就排除了医生在诊断过程中的主观因素，而更接近于客观的本质。

电子计算机能够作为对临床表现进行加工分析的工具，除了要记忆大量的已确诊的病例资料以外，还要求它有对这些数据进行加工的方法。由于临床表现与身体各器官状态变量关系的复杂性，探究这种理论模型无疑是一件困难的事情。而计算机数值诊断的统计方法却弥补了这一缺陷，给出了处理这种问题的一般工具。当然，利用统计方法构造诊断模型的细节仍然要通过反复试验，但这正是发挥计算机优越性的地方。电子计算机能够对不断接收的病例资料进行学习，不断修改诊断的模型，使它更符合实际。

这样，采用电子计算机的数值诊断，由于采用了大量的过去病例资料就能够提高诊断的准确性；由于综合了各家医生的诊断经验，而排除了个人诊断的局限性；由于电子计算机有极快的计算速度，大大提高了诊断的效率；由于采用了数学模型，可以预测人体健康的变化趋势。同时，由于电子计算机能够定量的分析各种临床表现，比如对心电图的频谱分析可以利用心电图的频谱与心脏失常的关系预测心脏的疾病状态，使得临床表现更好地发挥在诊断过程中的作用。这样，电子计算机就可以通过同样的临床表现获取比较多的信息，尽可能避免或减少那些手续繁琐、费用较高甚至于使病人痛苦的诊断手段而得到同样的诊断效果。

电子计算机数值诊断依赖于它记忆的资料，诊断的方法以及为实现这些方法的计算程序，这些都是人赋予它的功能。因此它不能比这些做的更多。如果人赋予它的计算程序没有诊断胃病的功能，它就不可能诊断各种胃病，另外不正确的资料的输入也会带来不正确的诊断结果，不正确的诊断方法的选择也会大大降低诊断的可靠性。因此，医生使用计算机进行

数值诊断并不意味着工作简单，他还要学会怎样与计算机发生联系，怎样从计算机获取信息以及不断改进计算机的诊断结构和推理方式，不断提高电子计算机数值诊断的效果。

尽管如此，由于电子计算机的数值诊断使医生个人的定性诊断进入到电子计算机的综合定量诊断，对诊断的可靠性和有效性将会发生深远的影响。迅速的发展这样一门与人体健康有着直接关系的新学科，显然是为人们所关注的问题。

§ 1.2 计算机数值诊断的方法

计算机数值诊断的基本内容是定量的得到临床资料的数据集合与不同的病名诊断之间的关系。比如为鉴别诊断慢性肝炎、肝硬化和肝癌，我们需要病人的临床资料，像年龄、性别、病程、上腹包块、胸水、肝大肋下、肝剑突下、脾大等。但是怎样确定不同的肝病与这些临床表现之间的关系呢？从目前来讲，主要涉及到判别分析，多元回归分析，聚类分析、序贯分析等一系列的统计方法。

建立计算机数值诊断模型采用统计方法还是比较符合实际的。这是因为我们所得到的临床表现都是和某种疾病的内部状态变量相关的外部表现，它是疾病所表现出的现象，并不是导致某种疾病发生的真正原因。疾病往往是很复杂的，一种疾病可以导致多种症状，同样的疾病可能会出现不同的症状，而不同的疾病也可能发生同样的症状。显然根据某一种临床表现进行诊断是不会成功的，它需要一个症状群。比如发热可能是感冒，也可能是炎症或者其它疾病。腹痛也很复杂。不能只根据发热或腹痛等一两种症状就下结论。一般说来，对不同的疾病如果有某些临床表现是相同的，也一定

会有某些临床表现是不相同的，只要症状群的症状选择的足够多，就能够根据症状群的整体区别作出正确的诊断。但事实上，症状群的症状数目又受到一定的限制，特别是诊断手段的限制，只能是根据可以计测的症状进行诊断。由于症状表现的现象和身体内部的状态变量之间存在着联系，使我们能够通过这些现象去反映内部状态变量。寻求这种联系不是从原因上去研究现象与本质的关系，而是通过大量的现象寻求它的统计规律性。也就是说，利用已往的大量病例，它们是已确诊的病例，从这些已存在的症状群与病名的现实资料出发，确定症状群与病名之间的统计规律性。这样，计算机数值诊断的信息处理过程如果采用统计方法，那么诊断的结果势必带有某种统计性质。它告诉我们某种疾病以多大的可能性发生，以及做出这种判断所可能犯的误差。但是医生诊断同样会发生某种错误判断，这是不可避免的，因为临床表现并不能完全反映人的内部器官变化情况。计算机数值诊断的发展可期望提高诊断的准确性。

由于计算机采用二进制的数字进行逻辑运算和推理，所以输入的症状须做数量化的处理。有些症状本身就是数量化的，比如体温、血压、白细胞计数；而有些症状则是定性的，比如性别有男和女之别，上腹包块是有与无之别，一般的化验指标则是阳性与阴性之别。这些定性指标必须做数量化的处理。

计算机能够诊断，必须存储一组与医生诊断知识相一致的 symptom 群和疾病关系的表格，以便根据新的病患者输入的症状群查表进行判断。或者存储 symptom 群和疾病的数学模型，使得能够根据输入症状信息进行计算和推理做出正确的判断。

所有这些逻辑表格与数学模型都必须是已经通过计算机对大量病例处理得到的经得起检验的结果。这是一种根据症状群进行综合分析一次得出结论的处理方式。

像医生一样，计算机也可采用问诊的方式。输入一种症状，进行一种判断，或对进一步诊断所需的症状进行提问，根据提问要求再输入新的症状。这样在问诊与分析之间循环进行，序贯的对症状进行分析，可期望用较少的症状获取数值诊断较高的符合率。问诊系统同样是由计算机首先处理好而存储到计算机里去的逻辑结构。

进一步，我们可以直接将大量的典型病例，以及对这些病例进行数据处理的多种方式全部存储入计算机中，对病患者输入的症状信息，自动地检索已存储的病历和选择适当的诊断结构和处理方式，进行自动诊断，并通过电传打字机打印标志或显示标志使计算机与人相互交换信息，实现人机对话。

所以，一个计算机的诊断系统，除了计算机的核心部分存储器、运算器、控制器以外，还需要实现人机对话的一套设备和手段，以及使计测工作标准化的医学工程的检测设备。计算机通过标准化的计测设备收集不同时刻，不同医院的信息资料，并通过比较、评价和压缩的信息处理，保留那些对诊断重要的症状和资料，不断地排除那些对诊断不太重要的或者其重要性已为新的症状和资料所代替的旧的症状和资料，并使它们加入到诊断系统中去，不断地提高诊断的可靠性和有效性。

医生诊断首先需要学习，在不断的诊断过程中积累经验。计算机数值诊断也有个不断学习和积累的过程，本质上

依赖于对事实的观测和分类。为了诊断的目的，必须有计划的观测某种病患的某些症状。要观测全部想像得到的症状是不可能的，只能是通过某些特殊病例去建立比较一般的原则。这是一种不完全肯定的推断过程。这种过程所发生的错误在诊断的普查工作中并不存在什么问题，作为统计方法的原则，不会影响到总体性质。但对具体人来讲，偶然的诊断错误会带来不应低估的损失。不断改进计算机诊断方法显然是一个重大课题。

§ 1.3 计算机数值诊断的前景

根据某病患者的症状表现综合分析得到病名的推断是计算机的静态诊断。静态诊断是只考虑病患者的目前状态，而不顾及他的历史状态和发展趋势。同时所用诊断方法的统计性质，使这些诊断模型都存在着一定局限性。而改进这种诊断模型的基本途径是研究动态诊断，即将诊断看作是诊断和治疗的一个完整的体系，并在不断诊断和治疗的过程中实现计算机诊断。

这里所指的计算机动态诊断，不仅仅是只提供病名，还将包括其它因素，特别是对医疗行为有决定意义的具有科学根据的信息。所谓治疗也不仅仅是只根据某个病名所提供的医疗手段，而是以病名为基础的所有信息得到的最能使人体病患状态恢复健康的行为。“诊断可能”就是能够从症状群推定内部状态。而“治疗可能”就是能使病患状态恢复到健康状态。诊断和治疗构成整个诊断过程中两个相互依存的方面。

从症状群到病名的确定以判定身体内部状态变量需要建

立数学模型和逻辑结构，同样也需要从病名和症状群以及根据症状群所预见的其它信息的基础上建立治疗的数学模型和逻辑结构，通过治疗手段改变身体内部不健康状态变量。体内状态变量的变化又引起新的诊断信息和新的治疗手段，在诊断和治疗的动态过程中，使身体内部的状态变量完全恢复到健康状态。

以诊断和治疗为支柱的统一的诊断体系使整个计算机诊断系统发生着深刻的变化。

首先计算机需要一个好的编组系统和把医学信息和治疗信息较好汇存的知识库。要研究信息资料的组织方式和存储使用方式，以便使计算机在接受新的医疗资料时能够比较迅速地查阅知识库，并把它处理后放入到知识库中去。同时对不同的症状群的类型组合能够便于查寻，以便找到诊断的最佳数学结构。一个好的编组系统是计算机动态诊断的基础。只有病例积累、医学资料、医学信息的利用和改进，才能使计算机数值诊断具有明显的效果。

第二，除了医学信息与治疗信息资料之外，还要求有灵活应用这些信息资料的方法，也即要具备各种理论模型存储的方法辞典。比如在诊断过程中某一断面上的多元回归模型、判别分析模型、聚类分析模型、最大可能度模型；在系统诊断过程中的序贯模型、马尔柯夫模型、控制理论模型；在症状选择上的信息计算模型、模糊聚类模型、主成分分析模型、方差分析模型；在症状分析中的频谱分析模型、时间序列分析模型；以及为实现这些模型的计算方法，比如线性代数方程组的解法，矩阵特征值与特征向量的计算、对称矩阵的分解、快速富里哀变换的技术等。把计算机的数值诊断

视为一个过程以及控制理论的应用，使计算机诊断提高到一个新的水平。

第三，进一步改进计测技术和诊断设备，使计测工作标准化，增加诊断手段使与体内状态变量密切相关的诊断手段的产生和改进，以及使医生与计算机实现人机对话的设备的完善，对于动态诊断，提高诊断的可靠性都是非常必要的，也是计算机诊断迅速发展的一个重要方面。

第四，一个完整的计算机诊断系统要求形成一个计算机网络，每一台计算机能够通过终端设备和医生以及其它医院进行通讯、接收、传递和计算所需要的医学信息。同时所有这种计算机诊断系统能够和一个为从事各种医学研究和社会健康控制调查提供所需数据的特殊的计算机进行对话和通讯。这样的计算机网络系统将直接地服务于医院、医生、病患者以及医学研究和管理者。特别像我国有着幅员广大的农村，建立计算机数值诊断的网络系统，使县城和公社的医疗机构，通过终端设备直接和计算机中心发生联系，对于农村的医疗卫生和保健工作就更有意义。

计测技术的发达、诊断理论的推进、电子计算机系统趋于完善是使计算机数值诊断不断得到发展的几个重要方面。在帮助延长人的寿命、增进人类健康、提高生产的效率方面、建立完整的动态的计算机诊断系统和管理系统的作用是不能低估的。

当然，计算机数值诊断的发展密切地联系着人类医学信息的巨大来源。医学工作者对人的内部状态变量研究的越深刻，对体内的状态变量与体外的症状表现关系的研究越是深刻，新的诊断手段的不断引进将直接影响到计算机数值诊断

的发展。计算机的数值诊断作为生物医学工程的一个重要方面，将随着整个生物医学工程学科的发展而发展。

计算机数值诊断的研究已经不是某一些人的兴趣和爱好，而已逐步成为诊断学本身赋予科学根据的不可缺少的一门科学。