1996 年 12 月
第 26 卷第 6 期

西北大学学报(自然科学版)
Journal of Northwest University(Natural Science Edition)

Dec. 1996
Vol. 26 No. 6

③ 471-476

一类统计决策优化模型的建立及其应用研究*

O212.1

肖筱南

(西安石油学院数理研究所,710065,西安;52 岁,男,副教授)

摘 要 借助于统计信息论,在信息熵的基础上给出了一种统计分析优化的信息方法,并通过对一个疾病诊断统计信息优化模型的建立与应用实例分析,探索了统计分析优化的新方法,为进一步充实与完善现行统计方法、深化统计分析改革提出了一种新途径。

关键词 统计决策;优化模型;信息熵;信息方法
分类号 O213.1

统计分析

1 传统统计分析方法的困惑与出路

统计分析是统计预决策理论研究的一个重要组成部分,它在社会与自然科学群的各个决策领域,如现代化企业管理与方案优选、投资决策、效益评估、生物医学、人口与保险等多方面都有着极为广泛的应用。然而,无论在理论与方法应用上,统计分析研究目前尚不够成熟,尤其是在方法研究上,还有待于进一步充实与完善。

传统统计分析方法,是在系统整理与综合评价各种统计调查资料的基础上,根据统计任务的要求,运用对比分析与动态比较、综合平衡与相关分析、抽样推断与随机分析等方法对被研究的对象进行数量分析,以揭示事物内部的矛盾与规律性,从而得到科学决策的优化方案的。然而,为了适应社会主义经济体制改革与统计改革的需要,进一步扩充与完善现行统计分析方法已刻不容缓。可是,由于传统统计分析方法的局限,特别是在如何充分利用统计信息与信息论的观点方法来进行统计分析的问题尚未得到很好解决,因而就很难得到适应当前经济体制改革的更为准确全面的最优决策方案。对这样一个亟待解决的现代化统计分析问题,本文拟运用 Shannon 的"统计信息论",在建立信息熵的基础上给出一种统计分析优化的信息方法,并通过对医学上的一个疾病诊断群的统计分析决策实例,来探索统计分析优化的一种新途径。

2 一种基于信息熵的统计信息方法

信息方法适用于对一切自然与社会系统的研究,揭示了物质运动形态之间的信息特征与联系以及事物运动的本质和矛盾,为科学技术和社会管理决策提供了新的思路与方法。然而,生物医学系统确是一个典型的复杂信息系统,若在对此系统进行统计分析时应用信息方法,无疑将会提供一条统计优化的新途径。

我们知道,运用统计与计算机方法进行疾病诊断有两个基本课题:一是评判"与所考虑的疾病群有关的征候及其表现"的诊断价值,据以筛选征候,精简资料;二是以诊断价值大的征候为基础,进而制定诊断判据(或算法),作为诊断疾病的准绳。目前,较常用的极大似然法、Bayes 分析法以及序贯分析法等数理统计方法,只能提供诊断判据,不能评判征候的诊断价值。为此,本文将提出一个能统筹解决这两方面问题的实用信息方法。

* 收稿日期:1995-11-09

用信息论的观点来看疾病诊断过程，可以把病人(可能患的疾病群)视为信源，医生视为信宿，而每做一项诊断检查所获得的一个征候表现，则视为一次通信。

设疾病群为$\{D_1, D_2, \cdots, D_m\}$，其中 m 种疾病互斥，则由 Shannon 信息论，可定义该疾病群(Z)的熵为

$$H(Z) = -\sum_{i=1}^{m} P(D_i)\log P(D_i) \tag{1}$$

其中 $P(D_i)$为疾病 D_i 的事前概率，$0<P(D_i)<1$，且 $\sum_{i=1}^{m} P(D_i) = 1$。采用自然对数时，熵的单位为 nat。显然，熵 $H(Z)$表示了医生对病人所患疾病诊断的不肯定性。$H(Z)$值越大，不肯定性就越大，反之亦然。

今设独立地做了 r 项诊断检查，$S_1, S_2, \cdots, S_r$，由检查 $S_k(1\leqslant k\leqslant r)$获得 n_k 个互斥的征候表现之一，记作 S_{kj}，此时，医生对病人所患疾病的不肯定性便由 $H(Z)$降为 $H(Z|S_{kj})$：

$$H(Z|S_{kj}) = -\sum_{i=1}^{m} P(D_i|S_{kj})\log P(D_i|S_{kj}) \tag{2}$$

$$k=1, 2, \cdots, r;\ j=1, 2, \cdots, n_k$$

其中 $P(D_i|S_{kj})$是已知征候表现 S_{kj}时，疾病 D_i 的条件概率。若记

$$T(Z, S_{kj}) = H(Z) - H(Z|S_{kj}) \tag{3}$$

则 $T(Z, S_{kj})$表示已知 S_{kj}时，疾病群不肯定性的减少量，即医生从中获得的信息量。显然，$H(Z|S_{kj})$越小，或 $T(Z, S_{kj})$越大(即越接近于 $H(Z)$)，则 S_{kj}的诊断价值就越大。

若 S_{kj}出现的概率为 $P(S_{kj})$，则医生通过检查 S_k 得到对病人所患疾病的平均不肯定性为

$$H(Z|S_k) = \sum_{j=1}^{n_k} P(S_{kj})H(Z|S_{kj}),$$

$$k = 1, 2, \cdots, r \tag{4}$$

而医生从中获得的平均信息量为

$$T(Z, S_k) = H(Z) - H(Z|S_k), \tag{5}$$

$$k = 1, 2, \cdots, r。$$

显然，$H(Z|S_k)$越小，或 $T(Z, S_k)$越大，则 S_k(即第 k 个征候)的诊断价值就越大。

这样，我们就可以利用(2)或(3)式来评判各征候表现的诊断价值，利用(4)或(5)式来评判各项检查(各个征候)的诊断价值。

经过诊断价值的评判，从中挑选出诊断价值大的若干个征候 $S_k(1\leqslant k\leqslant r)$以后，就可在这些重要征候的基础上制定诊断判据。为此，我们把(2)式改写为

$$H(Z|S_{kj}) = \sum_{i=1}^{m} H_i(S_{kj}),$$

其中 $H_i(S_{kj}) = -P(D_i|S_{kj})\log(D_i|S_{kj}),$ (6)

$$i = 1, 2, \cdots, m$$

表示已知征候表现 S_{kj}时，在总的不肯定性 $H(Z|S_{kj})$中有关疾病 D_i 的部分不肯定性。若各项检查是独立的，则经 r 项检查后，有关疾病 D_i 的部分不肯定性之和为

$$H_i = \sum_{k=1}^{r} H_i(S_{kj}), i = 1, 2, \cdots, m, \tag{7}$$

若 $\{H = \min H_i, i = 1, 2, \cdots, m\}$ (8)

则断言病人患疾病 $D_i(1\leqslant i\leqslant m)$。

3 应用实例分析

3.1 应用分析

现将上述方法应用于急性肠梗阻的鉴别诊断。

急性肠梗阻临床上分为绞窄性梗阻和单纯性梗阻两种类型。作者从某医院的临床统计资料，得到有关鉴别诊断急性肠梗阻的 17 个征候共 39 个征候表现如表 1 所示：

表 1　546 例肠梗阻原始统计资料

Tab. 1　The Ariginal Statistio Data of 546 Ileus Examples

征　候	表　现	单纯性		绞窄性	
		例　数	小　计	例　数	小　计
S_1(发　病)	急	69	263	174	282
	缓	194		108	
S_2(腹痛性质)	阵发性	180	240	144	273
	持续性	60		129	
S_3(腹痛强度)	剧	83	258	169	279
	非剧	175		110	
S_4(初始呕吐)	有	151	224	197	255
	无	73		58	
S_5(腹　胀)	有	173	238	208	264
	无	65		56	
S_6(既往开腹史)	有	181	264	134	282
	无	83		148	
S_7(肠鸣音)	亢进	131	213	98	239
	不亢进	56		44	
	减弱或消失	26		97	
S_8(腹部压痛)	重或中	85	229	181	259
	轻或无	144		78	
S_9(肌紧张)	有	35	190	98	194
	无	155		96	
S_{10}(反跳痛)	有	38	131	98	185
	无	93		87	
S_{11}(腹部肿块)	有	20	129	45	130
	无	109		85	
S_{12}(脉搏，次/min)	≤85	115	177	75	177
	85～105	44		58	
	≥105	18		44	
S_{13}(血压 Pa)	≤11 997	5	157	25	230
	11 997～15 996	103		147	
	≥15 996	49		58	
S_{14}(体温，℃)	≤37	95	217	45	120
	>37	122		75	
S_{15}(白细胞总数，千)	≤9	118	171	64	165
	9～13	37		48	
	≥13	16		53	
S_{16}(中性白细胞，%)	≤65	26	166	8	161
	65～90	133		121	
	≥90	7		32	
S_{17}(肠内液平面)	有	80	103	74	89
	无	23		15	

现根据表 1 所列 546 例病案资料(其中单纯性梗阻 $D_1$264 例，绞窄性梗阻 $D_2$282 例)依(1)～(5)式计算结果见表 2：

表 2 有关肠梗阻鉴别诊断征候的信息评价(nat)

Tab. 2 The Information Evaluation about Symptom of the Ileus Discernment and Diagnosis(nat)

征候	表现	$H(Z\|S_{kj})$	$T(Z,S_{kj})$	$H(Z\|S_k)$	$T(Z,S_k)$
S_1(发病)	急	0.597 4	0.095 2	0.627 4	0.065 2
	缓	0.651 5	0.041 1		
S_2(腹痛性质)	阵发性	0.683 0	0.009 6		
	持续性	0.635 1	0.507 5	0.665 5	0.027 1
S_3(腹痛强度)	剧	0.635 6	0.057 0		
	非剧	0.665 5	0.027 1	0.651 5	0.041 1
S_4(初始呕吐)	有	0.688 1	0.004 5		
	无	0.682 5	0.010 1	0.686 6	0.006 0
S_5(腹胀)	有	0.690 5	0.002 1		
	无	0.688 8	0.003 8	0.690 1	0.002 5
S_6(既往开腹史)	有	0.682 0	0.010 6		
	无	0.653 0	0.039 6	0.669 6	0.022 9
S_7(肠鸣音)	亢进	0.678 9	0.013 7		
	不亢进	0.682 7	0.009 9	0.638 8	0.053 8
	减弱或消失	0.526 5	0.166 1		
S_8(腹部压痛)	重或中	0.653 7	0.056 9		
	轻或无	0.640 0	0.052 6	0.637 7	0.054 9
S_9(肌紧张)	有	0.567 2	0.125 4		
	无	0.670 2	0.022 4	0.634 2	0.058 4
S_{10}(反跳痛)	有	0.640 3	0.052 3		
	无	0.678 4	0.014 2	0.662 6	0.030 0
S_{11}(腹部肿块)	有	0.606 0	0.857 0		
	无	0.688 6	0.004 0	0.667 9	0.024 7
S_{12}(脉搏,次/min)	≤85	0.677 1	0.015 5		
	85～105	0.678 7	0.013 9	0.662 1	0.030 5
	≥105	0.590 0	0.102 6		
S_{13}(血压 Pa)	≤11 997	0.520 8	0.171 8		
	11 997～15 996	0.692 9	−0.000 3	0.679 9	0.012 7
	≥15 996	0.690 4	0.002 2		
S_{14}(体温,℃)	≤37	0.692 2	0.000 4		
	>37	0.689 5	0.003 1	0.690 6	0.002 0
S_{15}(白细胞总数,千)	≤9	0.661 6	0.031 0		
	9～13	0.677 0	0.015 6	0.635 5	0.057 1
	≥13	0.519 6	0.173 0		
S_{16}(中性白细胞,%)	≤65	0.565 7	0.126 9		
	65～90	0.693 1	−0.000 5	0.650 1	0.042 5
	≥90	0.449 0	0.243 6		
S_{17}(肠内液平面)	有	0.690 9	0.001 7		
	无	0.687 4	0.002 5	0.690 2	0.002 4

$H(Z)=0.692\ 6$(nat)

由表 2,可依 $T(Z,S_k)$ 之值评判各征候的诊断价值,得到自大而小的排列次序如下:

$S_1,S_9,S_{15},S_8,S_7,S_{16},S_3,S_{12},S_{10},S_2,S_{11},S_6,S_{13},S_4,S_5,S_{17},S_{14}$。

若将 $T(Z,S_k)<0.01$(nat)的征候 S_4,S_5,S_{17},S_{14} 减掉,则可在余下的 13 个重要征候的基础上建立诊断判据。为此,先对 13 个征候依(7)式算得 $H_1(S_{kj})$ 和 $H_2(S_{Kj})$ 的结果于表 3 所示:

然后,再由待诊病人相应的一组 13 个征候表现,按(8)式计算 H_1 和 H_2,若 $H_1<H_2$,则诊断病人患疾病 D_1(单纯性梗阻);若 $H_1>H_2$,则诊断为患疾病 D_2(绞窄性梗阻)。

表3　供鉴别诊断用的 H 值

Tab. 3　The H Numerical Value Applying for Discernment and Diagnosis

重要征候	表现	$H_1(S_{kj})$	$H_2(S_{kj})$	$\Delta H(S_{kj})$
发病(S_1)	急	0.358	0.240	0.118
	缓	0.248	0.368	−0.084
肌紧张(S_6)	有	0.348	0.219	0.129
	无	0.303	0.367	−0.064
白细胞总数(千,S_{10})	≤9	0.294	0.368	−0.074
	9～13	0.366	0.311	0.055
	≥13	0.330	0.189	0.141
腹部压痛(S_8)	重或中	0.366	0.270	0.096
	轻或无	0.273	0.367	−0.094
肠鸣音(S_7)	亢进	0.314	0.365	−0.050
	不亢进	0.320	0.363	−0.043
	减弱或消失	0.333	0.194	0.139
中性白细胞(%,S_{14})	≤65	0.218	0.348	−0.130
	65～90	0.347	0.347	0.000
	≥90	0.298	0.151	0.147
腹痛强度(S_3)	剧	0.366	0.270	0.096
	非剧	0.298	0.368	−0.070
脉搏(次/min,S_{12})	≤85	0.312	0.365	−0.053
	85～105	0.365	0.314	0.051
	≥105	0.356	0.235	0.121
反跳痛(S_{13})	有	0.367	0.274	0.093
	无	0.313	0.365	−0.052
腹痛性质(S_2)	阵发性	0.320	0.363	−0.043
	持续性	0.366	0.269	0.097
腹部肿块(S_{11})	有	0.360	0.246	0.114
	无	0.330	0.359	−0.029
既往开腹史(S_4)	有	0.318	0.364	−0.046
	无	0.368	0.285	0.083
血压(S_{11},Pa)	≤11 997	0.330	0.190	0.140
	11 997～15 996	0.350	0.343	0.007
	≥15 996	0.334	0.357	−0.023

在只有两种疾病的情况下，也可先计算

$$\Delta H(S_{kj}) = H_1(S_{kj}) - H_2(S_{kj}) \tag{9}$$

$$k = 1,2,\cdots,r,\ j = 1,2,\cdots,n_k$$

然后，由待诊病人特有的 r 个(此例 $r = 13$) 征候表现计算

$$\Delta H = \sum_{k=1}^{r} \Delta H(S_{kj}), \tag{10}$$

若 $\Delta H < 0$，则诊断为患病 D_1；若 $\Delta H > 0$，则诊断为患病 D_2。

3.2　实　例

患者张某，男，50岁，工人，住院当天相应(见表3)13个重要征候摘要记录如下：昨天白天照常上班，晚饭后2 h突感脐周疼痛，阵发性，尚可忍受，同时有恶心呕吐。去年12月曾行阑尾切除术。体检：一般情况尚可，血压14 663～9 331 Pa，脉搏80次/min，腹肌软，局部有轻度压痛，未触及肿块，反跳痛(−)，偶可闻及肠鸣音亢进。化验：白细胞6 500，中性85%。

根据上述临床资料，将表3中相应的13个征候表现的 $H_1(S_{kj})$ 和 $H_2(S_{kj})$ 值相加，得

$$H_1 = 4.130\ (\text{nat}),\ H_2 = 4.581\ (\text{nat})$$

或 $\Delta H = -0.451$ (nat)

故初断为单纯性肠梗阻 D_1，此与住院的最后诊断是一致的。

今随机选取13个征候均有记载的30例急性肠梗阻患者做回顾性检验，结果29例判对，1例判错，准确率达96.7%，比对这30例患者用Bayes方法诊断结果(28例判对，2例判错，准确率为93.3%)准确率要高。

4 展 望

随着新时期市场经济的建立与发展，以及改革开放的进一步深入，统计分析对信息的需求将与日俱增，信息作为一种很有价值的资源将日益受到各级决策分析者的高度重视。可以预言，整个统计分析过程将会构成一个完整的信息处理系统，决策系统综合评价的一个重要方面将是对信息价值的全面系统的统计分析。由此可见，对统计信息的采集、加工、管理、分析以及控制利用问题，无论在理论研究与实际应用上，均具有非常重要的现实意义。

本文借助于信息论，提出了统计分析的一种信息方法，既是传统统计分析方法的一个延伸，又是经典信息论的一个扩充。文中在信息熵的基础上所提出的关于疾病诊断的统计信息优化模型，旨在引起有兴趣同行的注意，进而深化讨论，探索、扩展实现统计系统优化的新途径。

参 考 文 献

1 Meyer P L. 概率引论及统计应用. 潘孝瑞等译. 北京：高等教育出版社，1986. 230～299

2 程兴新，曹敏. 统计计算方法. 北京：北京大学出版社，1989. 157～262

3 Richard J L, Morris L M. An Instorduction to Mathematical Statistics and Its Applications. New Jersey: Prentice-Hall, 1986. 352～394

4 Maritz J S. Distribution-Free Statistical Methods. London: Chapman and Hall, 1981. 125～144

5 Kalbfleisch J G. Probability and Statistical Inference. New York: Springer-Verlag, 1985. 198～246

6 Lucian L C. Asymptotic Methods in Statistical Decision Theory. New York: Springer-Verlag, 1988. 264～298

责任编辑 张素敏

To Build an Optimization Model of Statistical Decision and a Study of Its Application

Xiao Xiaonan

(Mathematical Physics Institute, Xi'an University of Petroleum, 710065, Xi'an)

Abstract A kind of information method is given about statistical analysis majorization basing on the comentropy according to the theory of statistical information. And the new method of statistical analysis majorization is studied by analysing applied example and settlement of statistical information majorization model of disease diagnosis. A kind of new way is applied for sufficing and completing the old ststistical method forward, modifying the statistical analysis deeply.

Key words statistical decision; majorized model; entropy of information; information method