文章编号：2095-2163(2020)04-0081-05 中图分类号：TP391.7 文献标志码：A

基于数据挖掘的 CRC 肠道菌群营养干预可行性分析

成雨风，贺 松，刘 燕，黄诗懿
（贵州大学，贵阳 550025）

摘 要：随着智能技术与医疗健康领域融合的加深，正在不断提升着医疗服务水平。科学研究已经证明：结直肠癌的发生与肠道菌群存在密切关系。将人工智能运用在结直肠癌(CRC)肠道菌群营养干预上，可以帮助优化资源分配，提高医疗各环节的效率，提升诊疗效果。本文以五种常见肠道菌群为基础，结合数据挖掘的 K-Means 和 Apriori 算法，分析了基于数据挖掘的 CRC 肠道菌群营养干预的可行性。

关键词：结直肠癌(CRC)；人工智能 ；肠道菌群；数据挖掘；可行性

Feasibility analysis of CRC intestinal flora nutrition intervention based on data Mining

CHENG Yufeng, HE Song, LIU Yan, HUANG Shiyi
(Guizhou University, Guiyang 550025, China)

【Abstract】 With the deepening of the integration of intelligent technology and medical and health care, the level of medical services is constantly improving. Scientific research has proved that the occurrence of colorectal cancer is closely related to intestinal flora. The application of artificial intelligence in the nutritional intervention of colorectal cancer (CRC) intestinal flora can help optimize the allocation of resources, improve the efficiency of all aspects of medical treatment and improve the effect of diagnosis and treatment. Based on five common intestinal flora and combined with k-means and Apriori algorithms of data mining, this paper analyzed the feasibility of nutrition intervention for CRC intestinal flora based on data mining.

【Key words】 CRC; Artificial intelligence; Intestinal flora; Data mining; feasibility

0 引 言

近年来，随着人工智能领域的飞速发展，人工智能的应用场景越来越丰富，语音交互、计算机视觉、认知计算技术逐渐成熟。人工智能技术逐渐影响医疗行业的发展，人工智能技术与医疗卫生领域的集成持续增长，成为提高卫生保健水平的重要因素[1]。数据挖掘是通过算法查找隐藏在大量数据中信息的过程，通过数据挖掘可以得到一些不平凡的、新奇的、有价值的信息。如本文研究的，通过挖掘过往 CRC 病人的病例，运用 K - Means 算法和 Apriori 算法得出患者体内肠道菌群的内在联系，从而判断基于数据挖掘的 CRC 肠道菌群营养干预的可行性。

1 营养干预

营养干预是对人们营养上存在的问题进行相应改进的对策，也叫营养调理。简单来说，就是通过食用相关食物或者营养药物让人们的身体状况向健康的方向发展。

近年来，科学研究已经证明：由于肠道菌群是人体肠道的正常微生物，因此 CRC 的发生与肠道菌群存在密切的关系。据研究表明，大约有 10 万亿个细菌在人体肠道中寄生，影响人们的体重和消化能力，可以抵抗感染和自身免疫性疾病的风险，并在人类治疗过程中控制癌症的应激反应。人体健康与肠道中益生菌的结构密切相关。在肠道菌群的长期进化中，根据个体适应和自然选择，它们在不同菌群之间、菌群与宿主之间、菌群与宿主之间以及“环境平衡”中始终是动态的。正常情况下，由于相互依存和相互制约系统的形成，解剖结构相对稳定并且对宿主无致病性[2]。一些研究中得出，健康人肠道中健康细菌的比例已达 70%，普通人已降至 25%，便秘人群已减少至 15%，结直肠癌患者肠道中的益生菌比率仅为 10%。

据了解，现已有医院将肠道菌群的调节，作为结直肠癌治疗的主要手段，并研发了很多相关药物。如：双歧杆菌三联活菌肠溶胶囊、马来酸曲美布汀、丽珠肠乐、整肠生等药物，都是通过补充病人体内某种肠道菌群使病人体内的肠道菌群平衡，达到治疗

作者简介：成雨风(1994-)，男，硕士研究生，主要研究方向：医疗大数据的应用。
通讯作者：成雨风 Email：674583622@qq.com

收稿日期：2020-02-08

结直肠癌的效果。但这些手段也存在一些弊端,药物容易引起抗药性,对后期治疗产生相应干扰。如果采用营养干预,间接使肠道菌群平衡就能避免这些问题,并且营养干预比较注重食物疗法,在饮食的同时进行治疗,从而达到最理想的状态。

2 数据加工与清洗

数据挖掘是通过算法,查找隐藏在大量数据中的信息过程。通过数据挖掘可以得到一些不平凡的、新奇的、有价值的信息[3]。目前在人工智能领域,数据挖掘的常用算法有十余种,本文通过 K - Means 算法和 Apriori 算法,分析了 CRC 肠道菌群营养干预的可行性。

在进行数据挖掘之前,需要将数据转换为适合进一步分析和处理的数据形式。需要进行数据加工和数据清洗等前期工作。

数据加工的前提是数据提取,数据提取是一个涉及从各种来源检索数据的过程。而本文课题的数据均采用已有的电子病例,数据已经整理好,不需要多方面检索数据,只需将数据进行简单的数据加工即可。本文最主要的字段匹配,是将原数据表中缺少的字段,从其它数据表中有效地匹配过来。其中最常用的字段匹配是利用 VLOOKUP 函数:在表格的首列查找指定的数据,并返回指定的数据所在行中的指定列处的单元格内容。VLOOKUP 函数形式如下:

VLOOKUP(lookup_value,table_array,col_index_num,range_lookup)。

数据清洗是对删除、更正、不完整、格式有误等数据的处理。数据清洗的目的不仅仅更正错误,同时完成来自各个单独信息系统不同数据间的一致性,让原始数据可信且可用[4]。数据清洗的过程就是数据流动的过程,从不同异构数据源流向统一的目标数据。其间,数据的抽取、清洗、转换和装载形成串行或并行的过程,把数据清洗的流程模块化、清洗工具组件化、清洗过程智能化。整个过程中主要用到的是 SOL Server 数据库[5]。数据清洗中最主要的是数据类型转换,例如图 1 中,是将一段 varbinary 类型的数据转换为 varchar 类型的常规代码:

3 算法应用及分析

3.1 K-Means++算法

3.1.1 算法介绍

K - means 算法是一个聚类算法,目的是根据对象之间的相似性,将 n 个对象收集到 k 个不同的指定簇中,每个对象仅属于落在距聚类中心最小距离的类簇中[6]。虽然 K - means 算法已经提出来很久,但是它依然存在一些缺陷。如:必须预先给出聚类中心的数量 k,但实际上很难估计 k 的选择。K - Means 算法需要人为地确定初始分组中心,不同的初始分组中心可以导致完全不同的聚类结果。

```
IF OBJECT_ID ('dbo.varbin2hexstr') IS NOT NULL
    DROP FUNCTION dbo.varbin2hexstr
GO
CREATE function varbin2hexstr(
@bin varbinary(8000)
)returns varchar(8000)
as
begin
    declare @re varchar(8000),@i int
    select @re='',@i=datalength(@bin)
    while @i>0
        select
@re=substring('0123456789ABCDEF',substring(@bin,@i,1)/16+1,1)
+substring('0123456789ABCDEF',substring(@bin,@i,1)%16+1,1)
                +@re
               ,@i=@i-1
  -- return('0x'+@re)
    return @re
end
GO
```

图 1 varbinary 类型的数据转换为 varchar 类型数据常规代码

Fig. 1 Varbinary type data converted to varchar type data conventional code

通常情况下,在 CRC 病人体内乳酸杆菌产生病变,数目已经发生变化,有可能比正常值高或低。但 CRC 患者因年龄性别的不同,正常值都会有变化。所以,无法确定分类 k 值到底取多少。因此,需要运用一种新的基于数据分布选取初始聚类中心的 K - Means ++算法。K - *Means* ++算法的基本思想:使初始的聚类中心之间的相互欧氏距离尽可能的远,通过这种定义,可用该算法求解 k 值[7]。

K - *Means* ++ 算法流程:

输入:样本集 $D=\{x_1,x_2,x_3,\ldots,x_m\}$;聚类簇数 k。

Step 1 从样本集中随机选取一个样本作为第一个初始聚类中心 c_1;

Step 2 对数据集中的每个点 x_i,计算它与当前 c_1 聚类中心之间的最短距离(即与最近的一个聚类中心的距离),用 $D(x)$ 表示;

Step 3 把数据集中每个点与其距离最近的聚类中心点之间的距离相加,其和用 sum($D(x)$) 表示;

Step 4 在 0 到 sum($D(x)$) 之间取随机值 Random。Random = Random - $D(x)$,直到 Random ≤0,此时的点就是第二个聚类中心;

Step 5 重复 Step2 至 Step4,直到选出所有的 k

个聚类中心；

Step 6　对于数据集中的每个样本 x_i，分别计算它们到 k 个聚类中心的欧式距离，并将其分到距离最小的聚类中心对应的簇中；

Step 7　针对每个簇 c_i，更新聚类中心 $c_i = \frac{1}{|c_i|}\sum_{x \in c_i} x$（即计算该类样本的质心），$|c_i|$ 为该类样本个数；

Step 8　Until 聚类中心不再变化，误差平方和准则函数收敛。

输出：k 个簇，满足误差平方和准则函数收敛。

通过分析上述流程可见，改进的 K - Means ++ 算法在一定程度上降低了传统 K - Means 算法对初始值的依赖，降低了算法的不稳定性，提高了算法效率，减少了算法开销。

3.1.2　算法应用及分析

将筛选过的 300 份含有大肠埃希菌详细检测数据的病例通过上述数据清洗办法，集成数据集，然后将数据集导入进 K - Means ++ 算法。大肠埃希菌在人体内的标准值是 DNA(G+C)mol%为 48-59，以 54 为 c_1 的纵坐标，通过 K - Means ++ 算法应用计算 k 聚类中心个数，得到 k = 3，因为数据比较规范，通过 2 次迭代就得出稳定的纵坐标 y_{c_2} = 23，y_{c_3} = 71，可以看出 $y_{c_2} < 48$，$y_{c_3} > 59$，符合本文设想的数值。通过病例分析可知，当 DNA(G+C)mol%<48 时人们会有其他并发症产生、DNA(G+C)mol%>59 的时候人们的结直肠会发生病变，严重的会造成 CRC。

通过结果发现，个别病人的数据游离于 3 个聚类簇之外，距离 c_i 很远。通过分析，这些患者均是年龄较大或较小，这就是 K - Means 算法存在的不足。由于样本数目少，年龄分化低，所以无法更多的体现 K - Means ++ 算法的优势。

如果运用更多的样本，细化每个年龄段不同性别的人群，将常规的 5 种菌群所有簇类都计算出来，建立数据库，当有病人诊断时，导入 PCR 检测数据（人体内肠道菌群检测方式）智能模型，可直接对比病人的检查数据，提示医生该病人的对比情况，判断菌群是否病变。

3.2　Apriori 关联分析算法

3.2.1　算法介绍

CRC 患者体内并不是单一的肠道菌群病变，而是多种病变结合，并且每种关系间的营养干预方案也不一样，需要通过关联规则挖掘出内部的所有关系。

关联规则是指形如 $X \rightarrow Y$ 的表达式，且 $X \cap Y = \phi$。关联规则的强度可以用支持度（support）和置信度（confidence）度量。其中，支持度（s）指 X 与 Y 同时出现的事务在 T（所有事物的集合）中的比例，确定规则可用于给定数据集的频繁程度；置信度（c）指 X 与 Y 同时出现的事务在 X 出现的事务中的比例，衡量 Y 在包含 X 的事务中出现的频繁程度：

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N},$$

$$\sigma(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

通过定义可知，由关联规则作出的推论并不必然蕴涵因果关系[8]。

3.2.2　改进的 Apriori 算法

Apriori 是关联规则中最常用的一种算法。在 Apriori 算法中，最小支持度和最小置信度是该算法最重要的分析阈值和约束条件。但是，在处理医学数据时只有这两个条件是不够的。如需要更加精准的分析 CRC 病人的肠道菌群数据关联性，Apriori 算法还存在一些需要改进的地方[9]。

改进的 Apriori 算法依然是以 Apriori 算法的分段思想基础。首先在输入的事务数据集中寻找频繁项集，部分代码如下：

```
for (i=1;i <=n;i++) do begin
item[ i] .id=i;
item[ i] .frq=0;
  item[ i] .app=APP_BOTH;
  end;
  iset.init( );
  itemsets=iset .frequent(1);
  k=2;
  while( k<=maxlen and itemsets>=k) do begin
  itemsets=iset .candidates(k);
  if( itemsets>0)
  itemsets=iset .frequent(k)
  k=k+1 ;
end;
```

其中，集合 iset 选择树结构（抽象数据类型）存储，init 初始化数据结构，frequent(k) 产生 Lk，itemsets 返回项集中项的数量，maxlen 是关联规模的限制。

使用上述方式找到的频繁项集产生期望的规则如下：

For each frequentitemset lk in Lkwith k>=2

For eachitemset liin lk

If(supp(lk)>=minconf * supp(li)and(item[j] .appAPP_HEAD li)

and(item[j] .appAPP_BODY(lk -li)))

output rule li=>(lk-li)

With conf =supp(lk)/ supp(li)

And supp=supp(lk).

改进算法解决了两个问题:

(1)通过扩展项的属性、添加项出现位置的约束标记,解决了项的位置问题;

(2)通过设置关联最大长度 maxlen 限制了关联规模[10]。

3.3 实验结果与分析

实验中关于 CRC 病人的简单数据集,设有 7 个属性(p = 7)、5 个病人(n = 5)。表 1 是原始医疗数据,表 2 是把所有原始属性一一对应映射成项的映射表(索引表),表 3 是转换后的 CRC 事务数据表。

表 1 CRC 原始医疗数据

Tab. 1 CRC original medical data

Gender	Age	双歧杆菌	乳酸杆菌	大肠埃希菌	粪肠球菌	屎肠球菌
M	33	1(病变)	1(病变)	0(正常)	0(正常)	0(正常)
F	42	1(病变)	0(正常)	1(病变)	1(病变)	1(病变)
F	55	0(正常)	1(病变)	1(病变)	1(病变)	0(正常)
F	43	1(病变)	1(病变)	1(病变)	1(病变)	0(正常)
M	66	1(病变)	1(病变)	1(病变)	0(正常)	0(正常)

表 2 映射表 1

Tab. 2 Mapping table1

M	F	age<45	45≤age	双歧杆菌		乳酸杆菌	
1	2	3	4	1(病变) 5	0(正常) 6	1(病变) 7	0(正常) 8

表 3 映射表 2

Tab. 3 Mapping table2

大肠埃希菌		粪肠球菌		屎肠球菌	
1(病变)	0(正常)	1(病变)	0(正常)	1(病变)	0(正常)
9	10	11	12	13	14

表 4 将 CRC 医疗数据据映射成项

Tab. 4 Map CRC medical data into items

A_1	A_2	A_3	A_4	A_5	A_6	A_7
1	3	5	7	10	12	14
2	3	5	8	9	11	13
2	4	6	7	9	11	14
2	3	5	7	9	11	14
1	4	5	7	9	12	14

通过执行上述分配过程,原始数据中的每个元组(记录的一行)都映射到包含元素的事务元组。结果表和原始表的列数相同,但是每列都只包含整数值,该值与映射过程的索引一致。

实验数据使用表 1 提供的关于 CRC 诊断的数据集,数据集包含 5 个病人,7 个属性,其中两个为类别属性,5 个为数值属性。

对于关联规则挖掘程序的参数设置方法如下:由于选用的数据集较小,但希望发现两个或更多病人的关联,故最小支持度设置为 20%,设最小置信度为 80%。根据研究需要对项进行限制,得到实验需要的规则:

格式:[1,2,3,4]BODY

[5,6,7,8,9,10,11,12,13,14]BOTH

其中,BODY、BOTH 分别代表限制项出现的左部、左右部位置标记。即 CRC 病人的性别、年龄项(1,2,3,4)限制出现在规则的左部; CRC 病人肠道菌群指数项(5,6,7,8,9,10,11,12,13,14)出现在规则的左右部均可。

根据实验研究要求,设左部最大关联长度 body maxlen=3,后续研究可以根据规则需求设置不同的 maxlen 值。

通过实验,可以得到 5 个有关病人肠道菌群病变的项(5,7,9,11,13)出现在规则右部的所有规则。例如: 235 → 711(40%,50%) 规则表示 $Gender = F and age < 45$ *and* 双歧杆菌病变 → 乳酸杆菌病变 *and* 粪肠球菌病变 具有 40%的支持度和 100%的置信度,即:所研究病人有 40%(支持度)在 45 岁以下且双歧杆菌病变的男性,其肠道菌群中乳酸杆菌和粪肠球菌都产生病变,在 45 岁以下且双歧杆菌病变的男性中,因乳酸杆菌和粪肠球菌都产生病变导致 CRC 的可能性为 100%(置信度)。

由实验结果可见,若设 $X \rightarrow Y$ 是一个合法规则,则添加右部一项限制条件后,会有 $O(2^{|x|+|y|-1})$ 个无用规则被删除,达到了改进 Apriori 算法的目的。利用实际的诊断数据集可以验证已用的实际诊

断规则，但还需进一步考虑项的分组（菌群病变组合）问题，以及添加营养方案的项关联，这样将提高挖掘关联规则的性能，得到更准确的医疗规则。

通过深度学习，将已经完成的部分结合处理，通过不断的 AI 学习以及医学专家的修正，使病变规则和营养干预方案之间的联系更加合理，达到预设的地步。简易的、完善的干预模型可以通过下述程序体现：

```
SELECT
    programNumber,
intervention program,
FROM
nutrition intervention program library
WHERE
gender = ´M´ and age = ´33´ and bacillus bifida = ´53´;
```

上述代码表示：在营养干预方案库中查找 33 岁男性且检测出双歧杆菌数目只有 53 个（每克粪便）的营养干预方案。在实际过程中，若需要添加其他限制规则，都可以通过深度学习不断添加限制，继续完善。

4　结束语

本文通过研究 K - Means 算法、Apriori 算法在医疗数据上应用，发现相关算法在研究应用上的不足，提出了相应的 K - Means ++ 算法以及改进的 Apriori 算法，达到了对 CRC 疑似病人体内肠道菌群的分类，并发现其内部关联规则。从多方面验证了基于数据挖掘的 CRC 肠道菌群营养干预切实可行。本论文的研究还存在一些完善和探索的地方。如，后续需要完善所有规则，就需运用大型医疗数据集进行测试，同时要经过医疗专家考虑项的分组问题，从而进一步改进算法。随着 AI 技术的不断发展，以及后续的进一步研究，相信在不久的将来，这项技术一定会在临床治疗当中得以应用。

参考文献

[1] 马玲. 人工智能技术在医疗健康领域大显身手[J]. 科技中国，2019，12：79-81.

[2] 武庆斌，郑跃杰，黄永. 儿童肠道菌群——基础与临床[M]. 北京：科学出版社，2012.

[3] 狄晓娇. 关联分析在学生成绩数据挖掘的应用[J]. 电脑知识与技术，2018，12(34)：246-247，265.

[4] 毛云鹏，龙虎，邓韧，等. 数据清洗在医疗大数据分析中的应用[J]. 中国数字医学，2017，12(6)：49-52.

[5] 张好军. Web 数据集成中数据清洗的关键问题研究[D]. 山东大学，2009.

[6] 姚登举，詹晓娟，张晓晶. 一种加权 K-均值基因聚类算法[J]. 哈尔滨理工大学学报，2017，04：112-116，123.

[7] 王继博，杨蕾. 基于 K-Means 聚类的交通违法行为与事故关联关系研究[J]. 交通建设与管理，2019(05)：92-95.

[8] HAN Jiawei, KAMBER M. Data Mining: Concepts and Techniques[M]. China Machine Press, 2001.

[9] 李虹，蔡之华. 关联规则在医疗数据分析中的应用[J]. 微机发展，2013，06(6)：94-97.

[10] 柴华昕，王勇. Apriori 挖掘频繁项目集算法的改进[J]. 计算机工程与应用，2007(24)：158-161，171.

（上接第 80 页）

参考文献

[1] 任艺. 天津市众创空间发展研究[D]. 天津：天津工业大学，2017.

[2] KERA D. NanoSmano lab in Ljubljana: Disruptive prototypes and experimental governance of nanotechnologies in the hackerspace [J]. Journal of science Communication，2012(4)：37-49

[3] DOUGHERTY D. The maker movement [J]. Innovations，2012，(3)：11-14.

[4] CAPDEVILA I. Coworkers，Makers，and Fabbers - Global Local and Internal Dynamics of Innovation inLocalized Communities in Barcelona[D]. University of Montreal，2014.

[5] 郝君超，张瑜. 国内外众创空间现状及模式分析[J]. 科技管理研究，2016，36(18)：21-24.

[6] 臧维，李甜甜，徐磊. 北京市众创空间扶持政策工具挖掘及量化评价研究[J]. 软科学，2018，32(9)：56-61.

[7] 雷良海，贾天明. 上海市众创空间扶持政策研究[J]. 上海经济研究，2017(3)：32-39.

[8] 贾天明，雷良海. 众创空间的内涵、类型及盈利模式研究[J]. 当代经济管理，2017，39(6)：13-18.

[9] 曾建勋. 从创客空间到众创空间[J]. 数字图书馆论坛，2015(6)：1.

[10] 芦亚柯. 我国众创空间的运行模式、制度环境及制度创新策略[J]. 商业经济研究，2017(4)：121-123.

[11] 杨楠. 众创空间的运行机制与发展策略研究[J]. 商学研究，2018，25(6)：18-22.

[12] 曹如中. 创意资源如何转化为经营资本[N]. 解放日报，2017-03-21(010).

[13] 孙荣华，张建民. 基于创业生态系统的众创空间研究：一个研究框架[J]. 科技管理研究，2018，38(1)：244-249.

[14] 曹如中，史健勇，郭华，邱羚. 区域创意产业创新生态系统演进研究：动因、模型与功能划分[J]. 经济地理，2015，35(2)：107-113.

[15] 向永胜，古家军. 基于创业生态系统的新型众创空间构筑研究[J]. 科技进步与对策，2017，34(22)：20-24.

[16] 刘芹良，解学芳. 创新生态系统理论下众创空间生成机理研究[J]. 科技管理研究，2018，38(12)：240-247.

[17] LINDTNER S，LI D. Created in China. The makings of China's hackerspace community[J]. In Interactions，2012(6)：18-20.

[18] WILDAVSKY A. If Planning is Everything，Maybe It's Nothing [J]. Policy Sciences，1973，4(2)：127-153.