第34卷 第15期 计 算 机 工 程 2008年8月
Vol.34 No.15 Computer Engineering August 2008
·软件技术与数据库· 文章编号：1000—3428(2008)15—0075—02 文献标识码：A 中图分类号：TP301.6

# 基于云模型的肿瘤标志物联合检测

林 颖[1]，郭 锋[1]，李绍滋[1]，林端宜[2]
(1. 厦门大学智能科学与技术系，厦门 361005；2. 福建中医学院中西医结合研究院，福州 350108)

**摘 要：**医学发现，将具有相关性的肿瘤指标进行联合检测能提高癌症的阳性检出率。基于云模型的关联规则挖掘通过将属性的定义域模糊化，从而达到更好的挖掘效果。该文提出一种基于云模型的检测组合发现方法，在大量的体检报告中进行关联挖掘，找出与癌症相关性最大的肿瘤指标。实验表明，挖掘出的前10个检测组合中，有80%符合目前的医学常识。

**关键词：**数据挖掘；关联规则；云模型；肿瘤标志物；联合检测

# Combined Detection of Tumor Marker Based on Cloud Model

LIN Ying[1], GUO Feng[1], LI Shao-zi[1], LIN Duan-yi[2]
(1. Cognitive Science and Technology Department, Xiamen University, Xiamen 361005;
2. Academy of Integrative Medicine, Fujian College of Traditional Chinese Medicine, Fuzhou 350108)

**【Abstract】** In the field of tumor marker, detecting some markers together can improve the successful detection score. The cloud model has great advantages in interval definition of numerical attribute. By fuzzing the boundary of attribute and membership degree, the cloud model can express the real world well. This paper presents a combined markers detection discovery method to mine association and find out most valuable tumor markers in examination dataset. The experiments show 80% of the top 10 combined markers are fit for iatrical knowledge.

**【Key words】** data mining; association rule; cloud model; tumor marker; combined detection

## 1 概述

自 R.Agrawal 等人[1]提出了有关关联规则的挖掘算法以及相应算法以来，有关关联规则挖掘的研究一直是数据挖掘领域的热点。由于实际应用中很多属性都是数量型的，如年龄、收入等，因此需要将这些关联的研究扩展到数量属性上来。在传统的研究中有很多区间划分的方法，但这些方法都太硬，容易将边界元素排斥在区间外。

针对上面硬划分的问题，文献[2]提出了模糊集的概念，将数量型属性的定义域进行模糊化，从而软化了边界。但这种方法的缺点在于其隶属函数的实质及具体确定方法一直没有得到根本解决，隶属函数一旦被规定成精确的数值表达后，在概念定义、不确定性推理等过程中，就不再有丝毫模糊性。

为了真正解决区间硬划分的问题，文献[3]提出了基于云模型的概念划分方法，并将其应用在关联规则挖掘中的支持度、可信度以及相关性的计算。文献[4]则根据挖掘出的关联规则对预测进行了研究。云模型在模糊化区间划分的基础上，还将隶属函数进行模糊化。他们的研究表明云模型不仅可以解决硬划分带来的问题，而且将不确定推理等过程中的模糊性与随机性集成到了一起，较好地解决了以上的问题。

目前在肿瘤标志物的检测中，文献[5]发现将多项标志物进行联合检测可以利用不同标志物之间的互补作用极大提高肿瘤的阳性检出率。因此，发现新的联合检测组合是十分重要的工作。

## 2 基于云模型的关联规则

设 $T=\{t_1,t_2,\cdots,t_n\}$为一张数据表，其中，$t_j$ 表示 $T$ 的第 $j$ 个元组或称第 $j$ 条记录；$I=\{i_1,i_2,\cdots,i_m\}$表示 $T$ 的属性集或称字段集；$t_j[i_k]$表示第 $j$ 条记录在属性 $i_k$ 上的值。

设属性 $i_k$ 的定义域可以划分为多个基于云模型的概念：$F_{ik}=\{f^1{}_{ik},f^2{}_{ik},\cdots,f^l{}_{ik}\}$，其中，$f^j{}_{ik}$表示属性 $i_k$ 的第 $j$ 个概念，而每个概念都可以用数字特征 $Ex^j{}_{ik},En^j{}_{ik},He^j{}_{ik}$ 来表示。

设 $X=\{x_1,x_2,\cdots,x_p\}$和 $Y=\{y_1,y_2,\cdots,y_q\}$是 $I$ 的子集，且 $A=\{f_{x1},f_{x2},\cdots,f_{xp}\}$, $B=\{f_{y1},f_{y2},\cdots,f_{yq}\}$。其中，$f_{xi}$与 $f_{yj}$分别是属性 $x_i$ 和 $y_j$ 在论域上的概念。则现在的规则可以抽象为“如果 $X$ 是 $A$，则 $Y$ 是 $B$”。

在进行关联规则挖掘时，需要计算规则的支持率、可信度以及兴趣度，从而排除不好的规则，保留合适的规则。本文关联规则的支持率计算公式如下所示：

$$S_{X\text{ is }A\Rightarrow Y\text{ is }B}=\frac{\sum_{i=1}^{n}[\prod_{j=1}^{p}\mu f_{xj}(t_i[x_j])\cdot\prod_{k=1}^{q}\mu f_{yk}(t_i[y_k])]}{n} \quad (1)$$

关联规则的可信度计算公式如下所示：

$$C_{X\text{ is }A\Rightarrow Y\text{ is }B}=\frac{\sum_{i=1}^{n}[\prod_{j=1}^{p}\mu f_{xj}(t_i[x_j])\cdot\prod_{k=1}^{q}\mu f_{yk}(t_i[y_k])]}{\sum_{i=1}^{n}\prod_{j=1}^{p}\mu f_{xj}(t_i[x_j])} \quad (2)$$

然而可信度并不能真实反映规则的有效性。在部分情况下，规则的可信度超过了阈值，但规则中的结论如“$Y$ is $B$”本身的发生概率很大，甚至超过规则的可信度，则规则的前提与结论就形成了负相关，即“$X$ is $A$”将影响“$Y$ is $B$”的出现。因此，使用了兴趣度对规则进行度量，其计算公式如下所示：

**基金项目：**福建省科技厅社会发展基金资助重点项目(2006Y0016)
**作者简介：**林 颖(1979－)，女，助教、硕士，主研方向：自然语言处理，数据挖掘；郭 锋，助教、博士研究生；李绍滋，教授、博士生导师；林端宜，研究员
**收稿日期：**2007-09-06 **E-mail：**betop@xmu.edu.cn

$$I_{X\text{ is }A\Rightarrow Y\text{ is }B}=\frac{C_{X\text{ is }A\Rightarrow Y\text{ is }B}-S_{Y\text{ is }B}}{\max(C_{X\text{ is }A\Rightarrow Y\text{ is }B},S_{Y\text{ is }B})} \tag{3}$$

兴趣度的取值范围为[-1,1]，0 表示前提与结论无关，-1 表示负相关，1 表示正相关。根据云模型的理论，在对定义域中的任意 $x$ 计算隶属度时都将使用相应的 $X$ 条件云发生器得到其隶属度，即 $x$ 与隶属度的映射不是唯一的，这可以使计算结果更具有随机性。问题在于，在式(3)中对隶属度函数进行计算时将产生不同的值，不但使得兴趣度的绝对值有可能大于 1，还使得接下来应用的 CloudApriori 算法的前提不成立。因此规定，当式(3)被计算过一次后，接下来的值都采用第 1 次计算的结果。

## 3 基于云模型的肿瘤标志物关联规则挖掘

### 3.1 概念定义与数字特征

肿瘤标志物检测的数据如表 1 所示，每一条记录表示一个体检结果，体检的项目为肿瘤标志物中常见的项目，由于不是每个人都会检测所有项目，因此未检测的项目用 NULL 表示。

**表 1 部分检测数据**

| 姓名 | 标本日期 | 年龄 | 性别 | AFP | CEA | CA199 | … |
|---|---|---|---|---|---|---|---|
| 张三 | 2006.9.3 | 43 | 男 | 3.30 | 0.86 | 8.00 | |
| 李四 | 2006.9.4 | 42 | 男 | 阴性 | 阴性 | NULL | |
| 陈秀 | 2004.3.21 | 70 | 女 | 186.50 | 4.95 | 108.30 | |
| | | | …… | | | | |

对于每个检测的项目都有一个参考值的范围，如 AFP 的参考值为 0.0~20.0。对于每个项目设置 4 个概念，分别为阴性、正常、高和很高。以 AFP 为例，4 个概念的数字特征设置如表 2 所示。

**表 2 AFP 的 4 个概念以及相应的数字特征**

| 概念 | *Ex* | *En* | *He* |
|---|---|---|---|
| 阴性 | NULL | NULL | NULL |
| 正常 | 10 | 6.0 | 1.0 |
| 高 | 25 | 1.3 | 0.1 |
| 很高 | 100 | 23.0 | 3.0 |

在表 2 中，“正常”和“高”这 2 个概念使用正态云模型。“很高”使用半云模型进行描述。“阴性”表示无法找到肿瘤标志物，因此没有相应的检测数值而无法采用云的方式进行映射，规定这时直接映射到“阴性”这个概念，隶属度在 1 左右随机小幅度波动。在本实验中需要找到与肿瘤标志物阳性有关的其他指标，因此挖掘关联规则的结论都是阳性肿瘤标志物，且仅设置一个属性，如 $Y$={AFP},$B$={高}，或 $Y$={AFP}, $B$={很高}。

### 3.2 规则提取算法 CloudApriori

提取算法 CloudApriori 以 Apriori 算法为基础，并对其进行修改。为了更好地说明算法，先介绍下面一些相关的概念。

$k$ 项集：当规则的前提中包含有 $k$ 个属性以及相应概念，且结论中包含肿瘤标志物中的一个属性且概念为高或很高时，该规则被称为 $k$ 项集；

强 $k$ 项集：支持率大于最小支持率 *minsup* 的 $k$ 项集；

候选 $k$ 选集：支持率可能大于最小阈值的 $k$ 项集；

$R_k$：所有强 $k$ 项集的集合；

$C_k$：所有候选 $k$ 项集的集合。

为了使基于 Apriori 的算法 CloudApriori 正确运行，必须保证 Apriori 算法的要求，即需要给出下面的定理并证明其正确性。

**定理** 任何强项集的非空子集必是强项集。

证明：

设 $R$ 为强项集，根据强项集的定义有

$$S(R)>minsup \tag{4}$$

对于任意 $R$ 的非空子集 $S$，由于在从 $S$ 扩展到 $R$ 时只对前提进行连接运算，结论部分是不变的，因此 $R$ 与 $S$ 在式(1)中 $\prod_{k=1}^{q}\mu f_{yk}(t_i[y_k])$ 的值是相同的，且在区间[0,1]中。

当从 $S$ 扩展到 $R$ 时，式(1)中的 $p$ 不断增加，即属性的数目在增加，同时由于不同属性之间的隶属度使用了连乘积，而隶属度的值在[0,1]的区间中，$R$ 的 $\prod_{j=1}^{p}\mu f_{xj}(t_i[x_j])$ 要小于 $S$ 的 $\prod_{j=1}^{p}\mu f_{xj}(t_i[x_j])$，因此 $R$ 的支持率要小于 $S$ 的支持率。

最后考虑云模型的性质，由于在不同时刻计算同一个值在同一个属性下概念的隶属度时需要调用随机函数，将造成 $R$ 的支持率有可能大于 $S$ 的支持率，这是与定理相违背的。因此规定当公式 $\mu f_{xj}(t_i[x_j])$ 被计算过一次后，接下来的值都采用第 1 次计算的结果，这样就可以避免该值的变动，从而使定理成立。

根据上面的证明，得到下列结果：

$$S(S)>S(R)>minsup \tag{5}$$

所以定理成立，即可以采用 Apriori 算法的思想进行强项集之间的连接运算。

CloudApriori 算法描述如下：

输入：体检数据表 T，最小支持率 minsup，最小兴趣度 minint

输出：满足最小支持率与最小兴趣度的关联规则集合 R

初始化：

对每个肿瘤标志物根据“高”与“很高”这 2 个概念生成结论列表，结论即 3.1 节中的 Y 与 B，要求 Y 中仅有一个属性且 B 中仅有一个概念与其相对应。

将 T 中每个属性与概念的组合形成前提集合，即 3.1 节中的 X 与 A，要求 X 中仅有一个属性且 B 中仅有一个概念，与上一步生成的结论列表组成规则集合，满足 $X\cap Y=\Phi$；接着计算每个规则相应的支持率，并排除支持率不满足 minsup 的规则，最后将规则放入 $R_1$。

下文用递归的方式生成候选项集与强项集。

Call RecurApriori(T,minsup,$R_1$);

Procedure RecurApriori(T,minsup,$R_k$)

(1)将强项集 $R_{k+1}$ 置为空;

(2)对 $R_k$ 中的所有强 k 项集进行 Apriori 连接运算，产生候选 k 项集 $C_k$;

(3)对 $C_k$ 中的所有候选 k 项集计算其支持率，排除小于最小支持率的 k 项集，形成强 k 项集，如果该强 k 项集不属于 $R_{k+1}$，则将其加入 $R_{k+1}$;

(4)如果 $R_{k+1}$ 不为空，则递归 Call RecurApriori(T,minsup,minconf, $R_{k+1}$);

End

对 $\cup_k R_k$ 中的规则按照兴趣度进行降序排列，大于 *minint* 的规则放入 $R$ 中。

## 4 实验结果

### 4.1 数据集与实验环境

实验中共有 11 个肿瘤标志物，体检数据共有 2 037 条，稀疏度为 73.8%，设最大的 $K$ 固定为 5。经过 CloudApriori 算法计算后得到的规则集合按照兴趣度进行降序排列，得到兴趣度最高的 10 条规则。

### 4.2 关联规则挖掘结果

当取 *minsup* 为 0.000 1，*minint* 为 0.8 时，兴趣度最高的 5 条规则如表 3 所示。 （下转第 79 页）