

# 医院临床数据分析智能分类处理技术研究

廖多扬

(成都医学院第一附属医院 信息管理部, 成都 610500)

**摘要:** 针对医院临床数据数量庞大, 数据之间关联性大, 容易出现数据提取不准确等问题, 提出基于模糊分类处理技术的医院临床数据智能分类方法; 通过对临床运营各项指标的说明, 根据指标分析数据的特性; 对医院临床数据进行检索, 将检索出来的数据进行提取, 根据数据的特点, 使用模糊分类的技术对数据进行处理, 完成临床数据的智能分类; 实验结果表明, 所提方法对临床数据的分类效果远远优于传统方法, 满足了医院对数据处理的要求, 为未来医院大量的数据分类处理奠定了坚实的基础。

**关键词:** 医院; 临床数据; 智能分类; 处理技术

## Intelligent Classification and Processing of Clinical Data Analysis in Hospital

Liao Duoyang

(Information Management Department, First Affiliated Hospital of Chengdu Medical College, Chengdu 610500, China)

**Abstract:** In view of the large amount of clinical data in hospital, the large correlation between data and the inaccuracy of data extraction, a method of intelligent classification of hospital clinical data based on fuzzy classification processing technology is proposed. According to the clinical operation indicators, according to the analysis of the characteristics of the data index; retrieval of hospital clinical data, will retrieve the data were extracted according to the characteristics of data used for data processing of fuzzy classification technology, complete clinical data of intelligent classification. The experimental results show that the proposed method classification results for clinical data is far superior to the traditional method, meets the requirement of data processing for the future hospital, a large amount of data classification has laid a solid foundation.

**Keywords:** hospital; clinical data; intelligent classification; treatment technique

## 0 引言

现如今医院的数据库聚集了大量复杂的数据结构与多种信息, 急需分类处理技术对临床的数据进行处理<sup>[1-2]</sup>。但是传统的分类处理技术缺乏结构化和规整化的分类特性, 其分类处理的准确率也相对较低, 存在速度慢、使用过程繁琐、效率低等问题, 不能满足医院对数据分类处理的需求<sup>[3-4]</sup>。文献 [5] 中提出了一种基于 VSM 模型的数据分类处理技术, 该技术能够利用模型进行数据的统计, 并在此基础上对数据特征进行简单明了的描述, 进而进行分类处理。虽然该方法对数据处理的准确率较高, 但是指定的过程比较困难、推广性比较差, 很难实现大规模的推广应用; 文献 [6] 中提出了一种离散化的数据分类处理技术, 该技术能够提高算法的精准度, 具有较强的伸缩性。虽然该方法能够进行大规模的数据分类, 但是进行数据存取与预处理的过程中效率较低; 文献 [7] 中提出了一种概念漂移的数据分类处理技术, 该技术能够将分类的误差作为概念漂移的重要指标, 其分类的模型可根据指标调整构建的大小, 具有较大的灵活性。虽然该方法能够方便的对误差进行检验, 但是降低了分类处理的泛化错误率。

针对上述的观点, 我提出了一种模糊分类处理技术对医院临床数据进行分析。首先对医院临床运营的各种数据指标进行了说明, 根据指标分析了数据的特性; 然后对临床的数据使用了模糊的分类处理技术; 最后进行了实验。实验结果证明, 该技术的使用能够改善传统的分类的方法, 提高数据分类的高效

性与准确性, 是对数据处理的重要手段, 也是未来我国医院大量数据分类处理必然的发展趋势。

## 1 医院临床数据分析智能分类原理

### 1.1 临床数据分析分类特性获取

根据医院业务的流程, 建立临床运营的指标数据, 并分别从临床的效率、临床的负荷等方面对临床的数据进行分析, 如表 1 所示。

表 1 临床运营的数据指标

临床门诊的预约情况	入住的病人人数	临床门诊护士的排班
进行门诊的人数	平均住院的病人人数	护士对病人的注射次数
临床医护人员的平均工作量	参加门诊的次数	护士对病人的输液次数
临床医生日均治疗的人数	进行抢救的病人人数	护士对病人的换药次数
参加会诊的人数	平均抢救重病病人的次数	护士对病人的护理次数
日均参加门诊的医生数量	手术例数	护士对病人的抽血的次数
实际占用床位的数量	手术的时长	参加重病病人护理的人次
平均开放的床位数	住院的病床以及人数	发药的次数

由表 1 可知: 进行数据处理的时候应该根据最基本的原理来处理数据集  $Z = \{z_1, z_2, \dots, z_n\}$ , 并按照相关的规则, 将数据划分到预定义的数据类别  $X = \{x_1, x_2, \dots, x_m\}$  中, 具体处理流程如图 1 所示。

由图 1 可知: 对医院临床数据分析智能分类处理的过程中

收稿日期: 2017-11-20; 修回日期: 2017-12-19。

作者简介: 廖多扬 (1986-), 男, 四川成都人, 大学, 工程师, 主要从事医院信息化建设方向的研究。

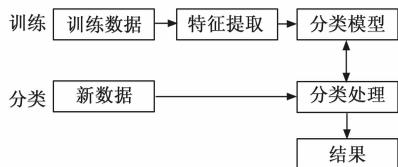


图 1 医院临床数据分析智能分类处理流程

主要分为了两个部分，分别是：训练的阶段、学习的阶段。训练的阶段是需要大量的人工来标记医院临床数据，并使用联机进行分析；学习的阶段是对整个训练的过程进行监督，需要对数据进行向量化<sup>[8]</sup>，然后用向量的特征组来表示。

总体来说，医院临床数据分析智能分类处理的特点主要有以下几点：

1) 数据的分类需要先经过训练再进行分析，因此对数据的样本分类的质量具有影响，会产生较大的误差。

2) 分类模型的建立主要依据训练的数据样本而进行构建的，因此这种建立的模型是具有局限性，而在实际的数据样本中，如果系统中没有与之相关的具有反馈学习的功能，那么整个分类处理的性能都会大大地下降。

3) 数据本身具有容量大、储存复杂、属性特征范围广、稀疏性等特点，因此进行数据分类处理的时候仅仅依靠单一的处理模型，很难提高分类处理的准确性，必须采用不同的技术对数据的分类处理加以解决。

针对以上的特点，结合数据分类处理的特点，以及分析的高效性进行一些智能的处理，并从整个处理的过程中来优化分类的处理技术，从而完善医院临床数据分析智能分类的处理。

## 2 医院临床数据智能分类技术

对医院临床数据分析智能分类处理的过程中首先需要对数据进行检索，然后将检索出来的数据进行提取，根据数据的特点，使用模糊分类的技术对数据进行处理。

### 2.1 数据的模糊检索与提取

进行数据检索的时候需要对数据建模，根据医院所处的工作环境以及对数据检索的要求来满足数据分类中的一些逻辑关系，并将检索的数据分成两个集合，分别是：匹配的集合和非匹配的集合。匹配的集合是指医院临床数据进行分类的时候，需要将数据进行匹配，再根据匹配的特征进行检索；非匹配的集合是指医院临床数据进行分类的时候，需要将单个数据进行记录，再根据记录的单个数据特征进行检索。这种匹配和非匹配的集合只能判断出进行检索的数据中是否与查询有关。建立模型的优势能够使检索的过程变得简单，使检索的速度加快。但是由于在匹配的过程中会出现二值的特性<sup>[9]</sup>，因此无法对查询的结果进行相关程序的排序，也就无法区分数据的特征对检索过程是否重要进行权衡。在模型中，假设数据表示为  $A = (a_1, a_2, \dots, a_n)$ ，而用户的检索可以表示为  $B = (b_1, b_2, \dots, b_n)$ ，其中  $a_i$  表示的是第  $n$  个特征点对数据内容的贡献程度，取值范围为  $[0, 1]$ ； $b_i$  表示的是第  $n$  个特征点对数据检索内容的贡献程度，取值范围为  $[0, 1]$ 。由此可得出数据检索的模型：

$$F(A, B) = 1 - \left[ \frac{\sum_{n=1}^i B_n^k (1 - A_n)^k}{\sum_{n=1}^i B_n^k} \right]^{\frac{1}{k}} \quad (1)$$

式 (1) 中， $1 \leq k \leq \infty$ ，取值的范围为  $[2, 6]$ 。对  $a_i$ 、 $b_i$  和  $k$  选取不同的数值，由此获取不同的结果。当  $k$  趋向于  $\infty$ ，并且  $a_i$  的

取值只能为 0 或 1，而  $b_i$  的取值只能为 1。

将医院临床数据分为  $i$  类， $W = W' \cup W''$ 。 $W'$  表示的是训练的数据集， $W''$  表示的是待分类的医院临床数据。使用该方法将待分类的医院临床数据与已经进行分类的数据重心进行比较，从中选取最大程度相似的种类别进行提取。第  $T$  类的向量可以表示为  $T = (t_1, t_2, \dots, t_n)$ ，待分类的医院临床数据可以表示为  $W'' = (\omega_1, \omega_2, \dots, \omega_n)$ ，它们之间的相似度可以表示为：

$$Sim(T, W'') = \frac{1}{\|T\| \|W''\|} \sum_{j=1}^m t_j \omega_j \quad (2)$$

式 (2) 中， $Sim(T, W'')$  表示的是最大时的医院临床数据类别即为所提取的数据类别。

### 2.2 数据的模糊分类处理

将分类描述为：指定的判断集合为  $d = \{\mathfrak{S}_1, \mathfrak{S}_2, \dots, \mathfrak{S}_m\}$  作为每一个  $\mathfrak{S}_i \in d$  建立的推理线路。推理线路利用医院数据的增益来度量核心的数据，即当某个数据的类别隶属度<sup>[10]</sup>远远超过任何一个预定阈值的时候就被认定为核心的数据。核心数据的评价函数为：

$$InfGain(Q) = R(Q) \sum_i R(\mathfrak{S}_i | Q) \log \frac{R(\mathfrak{S}_i | Q)}{R(\mathfrak{S}_i)} + R(\bar{Q}) \sum_i R(\mathfrak{S}_i | \bar{Q}) \log \frac{R(\mathfrak{S}_i | \bar{Q})}{R(\mathfrak{S}_i)} \quad (3)$$

式 (3) 中， $Q$  为一个数据， $R(Q)$  为数据  $Q$  出现时候的概率； $\bar{Q}$  数据出现，那么  $Q$  就不会出现， $R(\mathfrak{S}_i)$  为第  $i$  个  $\mathfrak{S}_i$  出现时候的概率， $R(\frac{\mathfrak{S}_i}{Q})$  为数据  $Q$  出现的时候， $\mathfrak{S}_i$  出现的概率。

$InfGain$  为数据分类的隶属度，当  $InfGain$  超过任何的指定的阈值时，就认为该数据为核心的数据。

模糊性指的是将核心的数据作为医院本质特性的数据，在分类处理的过程中，对于待分类的数据在某种程度上都属于一个同等的类别，并非是绝对的二值的逻辑关系。根据传统数据模型进行分类处理的时候，往往得到的是待分类的数据属于各个分类级别的数值信息，一般情况下根据最大的结果来判定数据分类的结果，但这种方法往往会丢失大量的数据信息，会产生较大的误差。而模糊的分类处理技术则是根据数据本身的模糊性来使用模糊的规则进行数据的处理，从而更加真实反应出医院的原始数据。因此，为了提高数据分类的智能性与准确性进行了模糊的分类处理。首先应该对数据分类的模型进行模糊化的处理，然后将具体的数据信息转换成为模糊的变量，然后根据实际情况进行相应的模糊分类处理，具体的过程如图 2 所示。

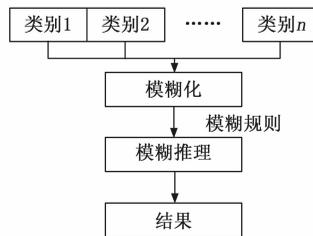


图 2 模糊分类的处理流程

由图 2 可知，不同数据的类别进行模糊化的处理，根据相应的模糊规则进行模糊的推理，从而得出结果。

使用模糊分类处理技术，可以较好地处理数据分类过程中出现的兼类、拒类的情况，如果只有一个类别的处理准确性较

高的时候，那么该数据的类别就是对应的分类；如果同时出现两个或者更多的数据分类处理准确性都较高的时候，那么该数据的类别就可以被分成多种的类别，被称作兼类；如果出现多个数据分类处理准确性都较低的时候，那么该数据的类别就被称作拒类。在分类的处理过程中能够根据数据的特征进行准确的分类，提高了实际应用的高效性。由以上步骤，实现了对医院临床数据的智能分类。

### 3 实验结果与分析

为了验证模糊分类处理技术下的数据智能分类的有效性进行了如下的实验，实验的条件如下：

实验中所用的数据选取自某医院 2015 年 1 月 1 日至 2016 年 1 月 1 日的临床数据，数据的总量为 300 M，经解析储存在数据库中。数据库的服务器采用的是 2.0 GHz，智能的联想微机搭建以及百兆的局限网络使用环境；数据库的服务安装 0.1.2.2 的企业版管理的系统。

#### 3.1 实验步骤

选取医院临床数据 200 个，其中，传统的数据处理技术下的数据为 100 个，基于模糊分类处理技术下的数据为 100 个，模拟实际医院临床的环境将 2 中技术下数据分类处理的准确率进行对比。根据医院临床实际的数据情况，对准确率的要求较高，相对数据的召回率可以适当的将要求放宽，保证准确率较高的前提下确保召回率一定即可，具体的对比情况如表 2 所示。

表 2 不同阈值下的数据分类的准确率

阈值	准确率
0.5	0.892
0.7	0.927
0.9	0.993

由表 2 可知：阈值越高，那么数据分类处理的准确率就越高，因为阈值降低会造成一些数据的误判，大大降低数据分类处理的准确率。

通过对医院临床数据进行智能的分析，可以对临床门诊的预约情况、入住的病人数、临床门诊护士的排班、临床医护人员平均工作量、参加门诊的次数、护士对病人的注射次数的数据进行检索，并使用模糊分类处理技术进行分类，具体的处理情况如表 3 所示。

表 3 不同临床指标的分类处理相似度

临床指标	相似度高数据所占百分比	相似度低数据所占百分比
临床门诊的预约情况	89.2	9.1
入住的病人数	76.1	9.0
临床门诊护士的排班	90.2	7.3
临床医护人员的平均工作量	54.6	31.2
参加门诊的次数	82.2	9.4
护士对病人的注射次数	70.1	12.5

由表 3 可知：不同临床指标相似度强的数据占有分类处理数据的百分比与相似度弱的数据占有分类处理数据的百分比要大。因此，模糊分类处理技术的数据处理的相似度对于不同临床指标来说具有较大的影响。

#### 3.2 实验结果分析

根据上述实验的过程可知不同阈值影响数据分类的准确

率，因此可将阈值恒定，对传统分类技术与基于模糊分类处理的技术的分类处理结果进行对比，结果如表 4 所示。

表 4 不同技术的数据分类处理准确率

	传统分类技术	模糊分类处理的技术
处理的时间/s	4	4
处理的准确率/%	0.902	0.997
数据的召回率/%	0.923	0.929

由表 4 可知：在相同的处理时间下，传统分类技术与模糊分类处理的技术的召回率相似，但是准确率相差较大。具体数据分类精度对比如图 3 所示。

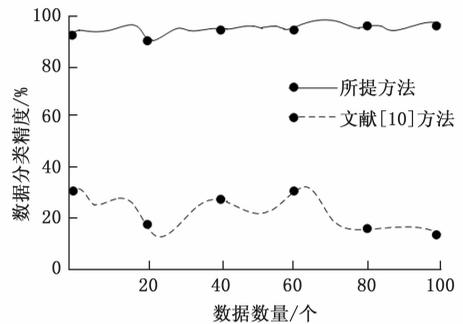


图 3 不同方法数据分类精度对比

由图 3 可知，对相同的 100 个数据进行分类处理，通过传统方法与所提方法进行对比出现了明显的分类精度差别，传统方法随着分类数据的数量增多，精度一直处于 40% 一下，且实验过程精度变化幅度较大，而所提方法数据分类精度一直保持在 90% 左右，且稳定性高，上下波动非常小，这进一步说明所提方法对医院临床数据的分类精度高，且该方法稳定性好，实用性更强。

而分类的速度的对比情况如图 4 所示。

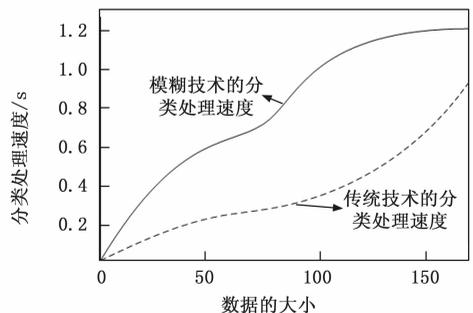


图 4 不同技术下数据分类处理速度

由图 4 可知：在数据的大小一致的情况下，传统分类处理技术的速度要小于模糊分类处理的技术。

由此可得出结论：模糊分类处理的技术对医院临床数据分析的时候，利用固定的内存进行实时的数据处理，具有较高的时空效率，其分类的性能远远高于传统的分类性能，将分析优势与智能的分类处理技术紧密的结合起来对数据进行分类处理，满足了人们对数据处理的需求，随着技术的不断进步，该分类的处理技能能够实现大容量的数据处理功能，发挥更大的作用。