

引用格式:岳根霞. 基于关联规则的医疗大数据挖掘算法[J]. 微电子学与计算机,2019,36(4):105-108. [YUE G X. Big Data Mining Algorithm Based on Association Rules[J]. Microelectronics & Computer,2019,36(4): 105-108.]

基于关联规则的医疗大数据挖掘算法

岳根霞

(山西医科大学汾阳学院,山西 汾阳 032200)

摘 要: 针对当前医疗大数据挖掘过程中医疗大数据挖掘效果差的问题,提出一种基于关联规则的医疗大数据挖掘算法,通过对医疗大数据的频繁项集进行查找,产生强关联规则,并对医疗大数据集的平均值和标准差的计算,实现对关联规则的调整,确定医疗大数据属性,构建合适的适应度函数,根据适应度函数值进行选择、交叉、变异操作得到约简的分类属性组合,对约简的分类属性信息熵进行计算,并通过计算期望信息得到分类属性的信息增益,建立决策树挖掘模型,实现医疗大数据挖掘. 实验结果表明,所提算法能够有效提高医疗大数据预处理的水平,数据挖掘的效果较好.

关键词: 关联规则;医疗;大数据;挖掘;算法

中图分类号: TP311

文献标识码: A

文章编号: 1000—7180(2019)04—0105—04

Big Data Mining Algorithm Based on Association Rules

YUE Gen-xia

(Fenyang College of Shanxi Medical University,Fenyang 032200,China)

Abstract: In view of the current medical data mining process, the medical data mining problems, proposes a mining algorithm based on association rules of medical data, searching through frequent itemsets of medical data, generate strong association rules, the calculated average value and standard deviation of large medical data sets, implementation the association rules adjustment, determine the medical data attributes, construct the appropriate fitness function, according to the classification attribute combination value of fitness function selection, crossover and mutation reduction, calculation of information entropy on classification attribute reduction, classification and attribute information gain by calculating the expected information, the establishment of decision tree mining model, implementation of medical data mining. The experimental results show that the proposed algorithm can effectively improve the level of medical data preprocessing, and the effect of data mining is better.

Key words: association rules;medical;big data;mining;algorithms

1 引言

医疗大数据挖掘已成为实现医疗信息化建设过程中的重要手段^[1]. 随着对医疗大数据挖掘研究的深入,关联规则算法逐渐受到相关专家学者的重视,已产生一些较为成熟的理论和应用,但仍存在一些问题,需要后续专家学者研究发展^[2].

文献[3]提出一种基于关联规则映射的医疗大

数据挖掘算法,通过引入挖掘因子和相对误差建立医疗数据之间的关联映射关系,根据医疗数据之间的关联程度,确定医疗大数据属性子空间,并对属性子空间中数据集进行区分,实现医疗大数据挖掘,但这种算法挖掘的计算过程较为复杂,导致挖掘过程中,数据容易出现数据失真的问题. 文献[4]提出一种基于 ICA 的医疗大数据挖掘算法,但这种算法数据挖掘的效果有待提高. 文献[5]提出一种基于 Ma-

pReduce 与相关子空间的医疗数据挖掘算法,但这种算法没有对数据的类别进行分析,影响挖掘效果.

针对上述问题,提出一种基于关联规则的医疗大数据挖掘算法,实验结果表明,所提算法能够有效提高医疗大数据挖掘质量,为该课题的深入研究发展提供理论依据.

2 基于关联规则的医疗大数据挖掘算法

2.1 关联规则的建立与调整

Apriori 算法作为关联规则中应用广泛的算法,利用其实现医疗大数据分析,需要对医疗大数据的频繁项集进行查找,生成关联规则,确定医疗大数据中项集间的关联关系,具体过程如下所述.

设定置信度表示医疗大数据中项集 A 和 B 同时发生的概率,其描述项集中同时包含 A 和 B 的项数与所有项数的比值^[6],则有:

$$\text{Support}(A \Rightarrow B) = P(A \cup B) = \frac{\text{Support_count}(A \cup B)}{\text{Total_count}} \quad (1)$$

上式中, $\text{Support_count}(A \cup B)$ 表示项集中同时包含 A 和 B 的项数, Total_count 表示项集中所有项数, $P(A \cup B)$ 表示项集中 A 和 B 同时包含的概率,

利用置信度建立关联规则,可以通过下式进行表示:

$$\text{Confidence}(A \Rightarrow B) = P(B | A) = \frac{\text{Support_count}(A \cup B)}{\text{Support_count}(A)} \quad (2)$$

上式中, $\text{Support_count}(A)$ 表示项集中包含 A 的项数.

通过上述论述,利用 Apriori 算法通过计算支持度,确定医疗大数据的频繁项集,利用频繁项集产生强关联规则,实现医疗大数据分析^[7],这种算法数据之间的关联性较弱,FP-Growth 算法作为关联规则中的另一个重要算法,通过将 FP-Growth 算法引入到 Apriori 算法中对关联规则进行调整,具体实现如下所述.

设定 T 表示医疗大数据的数据集, a_i 表示医疗大数据的属性, a_k 表示第 k 个医疗大数据的属性, x_j 表示类别, $T_{x_j} \in T$. μ_{ai} 和 σ_{ai} 分别表示医疗大数据集的平均值和标准差, $\Delta\mu_{\min}$ 表示允许类 x_j 中的平均值 a_i 与剩余项集中 a_i 的平均值的最小误差, $\sigma_{\max}$ 表示类别中 a_i 的最大标准差, $\gamma_{\min}$ 表示最小置信度,其计算公式分别可以表示为:

$$\mu_{ai}(T) = \frac{\sum_{k \in T} (a_k)}{|T|} \quad (3)$$

$$\sigma_{ai}(T) = \sqrt{\frac{\sum_{k \in T} (a_k - \mu_{ai}(T))^2}{|T|}} \quad (4)$$

$$\gamma_{\min} = \mu_{ai}(T_{x_j}) - \mu_{ai}(T - T_{x_j}) \quad (5)$$

每个属性 a_i 的平均值和标准差分别由类 x_j 产生,当观察值时最小样本时,医疗大数据挖掘满足以下条件的属性:

(1) a_i 在类 x_j 中存在不同于其他类的行为;

(2) a_i 在类 x_j 中具有一个统一行为.

为满足上述条件,对关联规则进行调整,得到调整后的关联规则可以表示为.

$$\text{Confidence}(A \Rightarrow B) = \frac{\mu_{a_i}(T_{x_j}) - \mu(T - T_{x_j})}{\frac{\sigma_{a_i}(T_{x_j})}{\sqrt{(|T_x|)}}} \quad (6)$$

根据上述论述,利用关联规则 Apriori 算法对医疗大数据的强关联规则进行计算,并通过 FP-Growth 算法对关联规则进行调整,利用关联规则对医疗大数据信息进行提取,确定医疗大数据属性.

2.2 基于遗传算法的数据属性约简

在关联规则进行数据挖掘的过程中,引入遗传算法,将问题的可行解表示成染色体,建立适应度函数,确定问题的最优解,利用遗传算法进行数据属性约简,具体过程如下所述.

设定通过 Apriori 算法和 FP-Growth 算法实现医疗大数据提取,得到的解为 $x = (x_1, x_2, \dots, x_n)$,通过二进制编码,生成医疗大数据属性的初始种群,通过逐代进化,对属性进行约简,初始种群通过随机的方式生成,其作为进化的第一代,设定 $Z = \{X_i = (x_{i1}, x_{i2}, \dots, x_{in}), i = 1, 2, \dots, n\}$ 表示一个医疗大数据属性的普通群体, X_i 表示群体中的染色体个体, x_i 表示该染色体中的基因,采用实数编码的方式,则 $x_i \in (a_i, b_i)$.

在遗传算法中,适应度函数的设计质量直接影响算法是索引效果,对数据挖掘的结果产生影响.采用遗传算法进行数据属性约简,设定目标函数为 $f(x)$,存在类 $C_{\max}$ 和 $C_{\min}$,使得 $C_{\min} < \{f(x)\}$, $C_{\max} > \{f(x)\}$, $C_{\max} > 0$. 寻找 $f(x)$ 的最大值,若染色体个体 X 对应的解为 x ,则个体 X 的适应度可以表示为:

$$\text{Fit}(X) = \begin{cases} -C_{\min} + f(x), & f(x) > C_{\min} \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

或:

$$\text{Fit}(X) = \frac{1}{C_{\max} - f(x)} \quad (8)$$

对目标函数 $f(x)$ 的最小值进行计算,根据上述,同理得到个体 X 的适应度可以表示为:

$$Fit(X) = \begin{cases} C_{\max} - f(x), & f(x) < C_{\max} \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

或:

$$Fit(X) = \frac{1}{-C_{\min} + f(x)} \quad (10)$$

遗传操作包括选择、交叉和变异,设定种群中存在 n 个个体,分别为 $x_1, x_2, \dots, x_n$, $Fit(x_i)$ 表示 x_i 适应能力大小,即适应度函数值,则该种群任意个体 x_i 被选择遗传到下一代的概率 $P(x_i)$ 可以表示为:

$$P(x_i) = \frac{Fit(x_i)}{\sum_{j=1}^n Fit(x_j)} \quad (11)$$

通过将概率 $P(x_i)$ 作为约简的阈值,当数据属性值大于概率则保留数据属性,否则提出数据属性,实现简约对医疗大数据的属性进行判断简约。

通过遗传算法的全局寻优能力对医疗大数据的属性集进行约简,得到约简的分类属性组合。

2.3 基于决策树算法的医疗大数据挖掘

决策树算法具有较好的分类效果,且分类规则易于被理解,成为数据挖掘分类中最为常用的技术,通过分析 ID3 和 C4.5 决策树算法在疾病分类诊断中的应用,建立医疗大数据决策树模型,实现对医疗大数据挖掘。

设定约简后医疗大数据的样本数据属性集合 S 中存在 s 个样本数据, $C_i (i = 1, 2, \dots, m)$ 表示样本数据分类的类别数量, s_i 是类 C_i 中的样本数量,数据集 S 对应的信息熵可以表示为:

$$I(s_1, s_2, \dots, s_m) = - \sum_{i=1}^m P_i \log_2(P_i) \quad (12)$$

式中, $P_i \approx s_i/s$ 表示样本集中任意样本属于类别 C_i 的概率. 设定医疗大数据属性 A 中具有 v 个不同的值,利用属性 A 将 S 分割成 v 个子集 $\{S_1, S_2, \dots, S_v\}$, 其中 S_j 表示 A 上具有 a_j 值的样本,设定 A 为当前的测试属性, s_{ij} 表示子集 S_j 中类别为 C_i 的样本个数,则属性 A 划分样本的信息熵为:

$$E(A) = \sum_{j=1}^v \frac{s_{1j} + \dots + s_{mj}}{s} I(s_{1j}, \dots, s_{mj}) \quad (13)$$

对于给定的子集 S_j , 期望信息 $I(s_{1j}, \dots, s_{mj})$ 可以通过下式进行计算:

$$I(s_{1j}, \dots, s_{mj}) = - \sum_{i=1}^m p_{ij} \log_2(p_{ij}) \quad (14)$$

式中, $p_{ij} = s_{ij}/|s_j|$ 表示 S_j 中的样本属于类 C_i 的概率; $|s_j|$ 表示在属性 A 上具有 a_j 值样本的个数. 根

据计算得到的信息熵和期望信息,得到属性 A 划分样本集 S 所获的信息增益为:

$$Gain(A) = I(s_{1j}, \dots, s_{mj}) - E(A) \quad (15)$$

通过上式可以看出,信息增益随属性取值的大小变化,属性取值越大. 但在实际应用中,属性取值大小不能决定分类属性结果,影响 ID3 算法对数据属性分类的效果,引入 C4.5 算法对分类属性进行优化. 得到属性 A 对应的初始信息量为:

$$SplitInf_A(S) = - \sum_{j=1}^v \frac{|S_j|}{S} \times \log_2\left(\frac{|S_j|}{S}\right) \quad (16)$$

则信息增益率的计算公式为:

$$GainRatio(A) = \frac{Gain(A)}{SplitInf_A(A)} \quad (17)$$

建立的决策树挖掘模型可以表示为:

$$\psi = \sum_{j=1}^v \frac{s_{1j} + \dots + s_{mj}}{s} \cdot Gainratio(A) \quad (18)$$

通过上述论述,对医疗大数据属性训练样本集进行预处理,计算各属性的信息增益率,选取信息增益率最大的属性作为决策树的根节点,并从候选属性集中去除已使用属性,直到候选属性子集为空集,通过分配类别属性,合并及约简相同的类别属性,生成医疗大数据属性决策树挖掘模型,实现医疗大数据挖掘。

3 实验结果与分析

为了验证本文所提基于关联规则的医疗大数据挖掘算法的有效性和实用性,进行了一次实验,实验过程中,采用 CPU 为 Intel i7 的计算机作为实验硬件环境,利用 matlab 对实验过程进行模拟,将某医院医疗数据库中数据作为实验数据。

对医疗大数据进行预处理,能够剔除利用价值较低的数据,提高医疗大数据的查询效率. 数据的完整性是评价医疗大数据处理效果的重要因素,分析本文算法进行预处理后医疗数据的丢包情况,能够对数据的质量进行分析. 引入文献[4]算法和文献[5]算法,对数据集中数据预处理,确定处理后出现丢包情况的数据数目,通过实验,得到的结果如表 1 所示。

表 1 不同算法处理后数据的丢包情况对比

数据/条	本文算法/条	文献[4]算法/条	文献[5]算法/条
200	3	4	5
400	4	8	10
600	8	11	12

通过表 1 可以看出,文献[4]算法和文献[5]算法集中出现丢包的数据较多,本文所提算法出现丢包的数据较少,说明本文算法进行医疗大数据预处理后,处理后数据的完整性较好,保证了医疗大数据的质量.

通过实验检测 Apriori 算法、FP-Growth 算法以及改进关联规则算法的实际应用性能,分别利用三种算法对医疗数据进行处理,分析三种算法执行时间,通过实验,得到的结果如表 2 所示.

表 2 三种算法的执行时间对比

数据 /MB	Apriori 算法 /s	FP-Growth 算法/s	改进关联规则算法/s
10	5.4	4.2	2.4
20	7.1	5.3	3.3
30	8.8	6.9	4.2

通过表 2 看出,改进的关联规则算法执行时间少于 FP-Growth 算法和 Apriori 算法,执行速度快.综合上述实验,改进关联规则算法处理后数据的完整性和准确性较好,说明改进关联规则算法具有良好的使用效果.

对数据采用本文方法进行挖掘,将挖掘结果与医生做出的结果进行比较,得到的结果如下图所示.

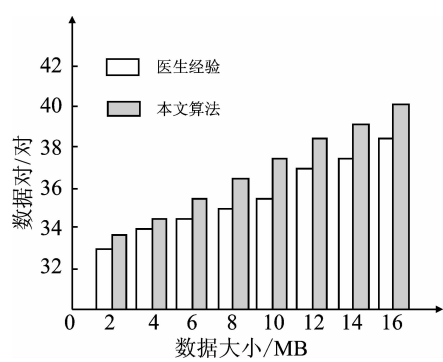


图 1 本文算法数据挖掘效果分析

通过图 1 可以看出,临床经验丰富的医生确定的数据对数量少于本文算法确定的数据对数量.说明本文算法对医疗数据挖掘的较为深入,能够发现

医疗大数据之间关联关系,有效提高了数据挖掘的效果.

4 结束语

数据挖掘是实现数据分析利用的重要手段,对提高医疗大数据的利用效率具有重要意义.针对当前医疗大数据挖掘算法存在的挖掘效果较差问题,提出一种基于关联规则的医疗大数据挖掘算法.实验结果表明,所提算法数据的处理效果较好,且数据挖掘内容较为深入,能够发现数据之间的关联关系.

参考文献:

[1] 张继荣,王向阳. 基于 XML 数据挖掘的 Apriori 算法的研究与改进[J]. 计算机测量与控制, 2016, 24(6): 178-180.

[2] 张新英,付川南. 一种高效的多类型数据挖掘算法[J]. 中国电子科学研究院学报, 2017, 12(4): 359-364.

[3] 唐晓东. 基于关联规则映射的生物信息网络多维数据挖掘算法[J]. 计算机应用研究, 2015, 32(6): 1614-1616.

[4] 王莉君,何政伟,冯平兴. 基于 ICA 的异常数据挖掘算法研究[J]. 电子科技大学学报, 2015, 44(2): 211-214.

[5] 张继福,李永红,秦啸,等. 基于 MapReduce 与相关子空间的局部离群数据挖掘算法[J]. 软件学报, 2015, 26(5):1079-1095.

[6] 王慧,张翠羽. 基于改进遗传算法的网络差异数据挖掘算法[J]. 计算机仿真, 2015, 32(5):311-314.

[7] 宋万洋,李国和,吴卫江,等. 基于平衡准确率和规模的决策树剪枝算法[J]. 科学技术与工程, 2016, 16(16):79-82.

作者简介:

岳根霞 女,(1977-),硕士,副教授.研究方向为大数据、数据挖掘. E-mail:yuegenxia@163.com