遗传算法的数据挖掘技术在医疗大数据中的应用研究

陈 萌

（山东中医药大学，济南 250355）

[摘 要]由于医疗行业具有较强的复杂性与特殊性，且医疗大数据中的数据信息十分复杂，为了能够对患者病情进行更好地分析与决策，相关部门要发挥医疗大数据的作用，通过应用数据挖掘技术，实现对患者的有效治疗。基于此，本文分析了遗传算法中的数据挖掘技术，研究了其在医疗大数据中的实际应用，旨在为相关研究提供借鉴。

[关键词]遗传算法；医疗；大数据；数据挖掘技术

doi：10.3969/j.issn.1673－0194.2019.08.077

[中图分类号]TP311.13；TP18 [文献标识码]A [文章编号]1673-0194（2019）08-0173-02

0 引 言

随着科学技术水平的提升，计算机网络技术被广泛用于各个行业之中，尤其是在医疗事业中，不仅改善了医疗服务质量，还提升了服务水平。治疗过程中会产生大量的数据，像医疗器械信息、患者个人信息数据等，为了发挥出医疗数据的最大价值，医疗机构要加强数据挖掘技术在医疗大数据中的应用力度。

1 数据挖掘技术的概述

数据挖掘指从海量的数据信息中挖掘出有效的知识或模式，其在应用上主要由数据、算法以及知识3种要素构成，其中，数据是数据挖掘的基础，算法是重要手段，获取知识是最终目的。以下是对数据挖掘3种要素的介绍。

1.1 数据

数据的形式多种多样，如文本数据、影像数据以及音频数据等。数据的描述主要从两个方面进行，一是记录数，二是属性数。在大数据时代，数据的记录数量多，属性涵盖范围广。此外，属性又被称为特征、变量或维度，是刻画对象特征或性质的一种方式，会随着对象及时间的改变而改变。数据的分析技术方式是由属性决定的，对于属性类型的判定可以通过明确数值性质的方式找到对应性质。常用的数值性质有“=、≠”的相异性，以及“>、<、+、－、×、÷”等符号。

当数值性质固定后，即可定义其属性类型。首先是标称，如颜色、医嘱类型等，如果只是数值的名称具有差异，一般通过“=、≠”加以区分；其次是序数，如收入水平等，数值能够明确对象的序，一般通过“>、<”加以区分；再次是区间，如摄氏度、华氏温度等，数值间的差具有一定的意义，一般通过“+、－”加以区分；最后是比率，如药量、体重等，数值的差与比率存在实际意义，一般通过“×、÷”加以区分。此外，标称与序数在一般状况下被合称为定性属性或是分类属性，而区间与比率被合称为定量属性或是数值属性。

[收稿日期]2019－03－11

在明确数据属性类型后，为了构建出各个数据间的逻辑关系及模型，可以对相关数据信息进行描述性地统计与分析，主要从数据的均值、众数以及中位数分析数据中心趋势，从方差、极差以及标准差等方面分析数据离散趋势，从而制作出直方图、折线图等描述性的图表。

1.2 算法

算法是探寻数据间的规律，将其转变为人类可理解的形式，主要分成6类。第一类是分布探索。了解多个数据间的客观分布状况，一般采取聚类分析技术进行数据挖掘。第二类是关系探索。了解不同事物之间及变量之间的关系，一般采取关联规则技术进行数据挖掘。第三类是特征选择。了解高维变量事物的重要特征，一般采取特征抽取技术进行数据挖掘。第四类是异常探索。了解高维变量事物的个性离群案例，一般采取异常侦测技术进行数据挖掘。第五类是推测探索。按照已知变量数量判断目标标量的值。第六类是趋势探索。一般会按照时间次序对事物的变化趋势进行考察与推测，一般采取时间序列进行数据挖掘。

1.3 知识

通常情况下，数据挖掘技术会被应用在固定数据挖掘任务模式中，此模式主要分为描述性与预测性两种类别。其中描述性模式能够刻画出数据的基本性质，包含离群点、频繁模式等；而预测性模式能够归纳已上传数据，从而实现有效预测，可以对预测分析进行分类及回归等。此外，描述性模式中的频繁模式具有多种形式，包括频繁子结构、序列模式以及项集等。其中，频繁项集是在事务数据中经常出现的物品集合，如医生开具的医嘱中经常性出现的药品；而频繁序列模式是在医生开具检验项目后，根据检测结果所做出的治疗方案及药物使用的过程。

2 数据挖掘在医学大数据中的应用

随着医疗事业信息化建设速度的不断提升，其产生的信息数据数量也在不断增多，包含实验室数据、患者治疗信息以及

临床研究数据等，这些数据蕴含着许多高价值信息，需要使用数据挖掘技术进行价值挖掘。因此，在医疗大数据中使用数据挖掘技术，不仅能够提升整个医疗事业的服务质量与水平，还能提高治疗效率与质量。在实际使用过程中，若想对数据对象进行科学研究，医疗机构要适当改进挖掘算法，从而增强数据挖掘的效果，优化医疗服务水准。

2.1 明确数据挖掘对象

为了使数据挖掘技术在医疗大数据中发挥出最大效用，要先明确数据挖掘的对象。从实际医疗事业发展上看，数据挖掘对象主要包括存在于互联网中的一些患者信息、费用信息以及药物信息、医疗设施信息等，只有明确好数据挖掘对象，才能增强数据挖掘技术效果。此外，在进行正式数据挖掘前，要研究出挖掘的主要流程。数据挖掘主要分为 7 个步骤：第一，对挖掘问题进行定义，并选择相应的数据信息进行分析，此步骤关乎后续数据挖掘的有效性与合理性，具有重要作用；第二，预先处理好所选择的数据，并将正确、有效、合理的数据输入至数据库中；第三，进行数据集成，处理好有关数据的共享问题；第四，清理数据，将不合理、或存在漏洞的数据进行删除处理；第五，交换数据，此步骤能够确保挖掘形式与数据一致，增强数据挖掘效果；第六，数据规约，通过删除某列或某行等方式，保证挖掘运算量合理；第七，数据挖掘，采集目标信息，进行最终结果评价与展示。

2.2 以遗传算法为基准的 K-means 聚类算法

K-means 聚类算法是一种距离聚类迭代算法，将相似性较高的一些数据以点聚集的方式集中在某簇中，将相似性较小或具有差异的数据归置到其他簇中，按照有关约束规定实现数据的有效迭代。为了加强数据挖掘技术在医疗大数据中的使用效果，医疗机构要对 K-means 聚类算法进行适当地改进，从而促进我国医疗事业的进一步发展。

首先，要制订编码方案，做出种群的初始解。在对数据挖掘算法进行改进时，要将 K-means 聚类算法和遗传算法进行有效结合。先定义出实际中心坐标，将其设定为 d 维，再设定每簇染色体的长度为 $k \times d$；每条染色体为 $\{P_1, P_2, P_3, P_4, \cdots, P_k\}$，$P_i=\{P_{j1}, P_{j2}, P_{j3}, P_{j4}, \cdots, P_{jd}\}$。在制订好编码方案，做好种群初始化后，相关机构要以随机的方式从多个对象中明确 K 个初始中心坐标。其次，为了确保医疗大数据在使用数据挖掘技术上具有较强的合理性，要选择好适应函数。所谓的适应函数就是计算各个数据的适应度，该过程有利于相关机构获得最优解。一般情况下，适应函数的公式如下。

$$F=\frac{1}{E}=\frac{1}{\sum_{i=1}^{k}\sum_{p\in C_i}\left|p-\overline{p_i}\right|^2} \quad (1)$$

最后，进行操作选择。为了增强实际操作性能，在数据迭代时可以适当引入一些免疫机制，以此实现更好的操作。改进算法的流程如下：先输入原始数据，设 n 为迭代数，n=1；再对染色体编码，产生初始种群；接着使用 K-means 操作优化种群个体；然后计算出存在于种群中每个个体的适应度值；对失传因子进行免疫机制筛选，做好自适应动态调整，若最后不满足终止条件，则需要重新计算适应度值，待满足终止条件后，选择适应度最大的个体作为最优解输出，完成整个医疗大数据的数据挖掘工作。

2.3 案例分析

按照现阶段医疗事业的发展情况，本文以医疗费用数据为例，对其进行数据挖掘。由于医疗费用数据具有冗杂性、隐蔽性及多样性等特征，与医疗大数据特征相符，因此，通过应用数据挖掘技术，能够增强费用结算与查询效果的功能，提高医疗服务水平。在进行实际数据挖掘时，医疗机构通过信息采集系统获取患者的数据信息，再根据数据信息选择出分类算法。一般情况下会选择四分位数法，该方法相对于其他算法更易理解，通过将患者进行分组，获得其主要数据信息，包括年龄、疾病或医疗费用等。本案例将医疗费用的 25%、50%、75% 作为分界线，对数据进行区间化处理。

按照传统的四分位分类算法可以得出，在费用 25% 时，其医疗费用为 853.01 元，换病例数为 42；在费用 50% 时，其医疗费用为 1 446.28 元，在 25%~50% 间的患病例数为 44；在费用 75% 时，其医疗费用为 3 184.52 元，在 50%~75% 之间的患病例数为 44，高于 75% 的患病例数为 42。按照 K-means 聚类算法对以上医疗费用数据进行挖掘时，可以将其分为 4 个聚类中心，其总费用分别为 1 123.48 元、3 581.53 元、8 828.64 元、14 369.25 元，患病例数分别为 112 例、50 例、5 例、5 例。通过对该算法进行分析，可以看出 K-means 聚类算法的分类挖掘效果十分显著，可以表述出不同的聚类中心，得到更为详细的医疗费用使用情况与例数。

3 结 语

医疗行业的快速发展，使大数据技术的应用范围逐渐扩大，通过应用合理的数据挖掘技术，采集多种医疗数据信息，进行医疗大数据挖掘，不仅能够提升医疗服务质量，还能提高医疗信息利用率。此外，在医疗大数据中使用数据挖掘技术时，要按照选择的数据对象对数据挖掘技术进行适当地调整。

主要参考文献

[1] 张晴，李洁莉，朱家沐，等 . 基于物联网的健康医疗大数据深层挖掘的应用与研究 [J]. 中国医学装备 ,2019(1).

[2] 罗堃，代冕 . 数据挖掘技术在医疗大数据中的应用研究 [J]. 信息与电脑 ,2016(6).

[3] 樊小毛，何晨光，卢东昕，等 . 医疗大数据特征挖掘及重大突发疾病早期预警 [J]. 网络新媒体技术 ,2014(1).