

基于医疗数据的属性约简聚类分析算法^{*}

李晓雪 郑静晨 李 明 郝昱文

(中国人民武装警察部队总医院 北京 100039)

[摘要] 通过医疗数据预测疾病的发展状况具有十分重要的意义。介绍聚类分析和属性约简的概念, 阐述典型的聚类分析流程和新的基于属性的约简聚类分析算法, 通过对医疗数据进行属性约简, 可极大提高聚类分析过程的效率, 同时通过与现有经典聚类算法的比较, 证明该算法的有效性。

[关键词] 医疗数据; 数据属性约简; 聚类分析; 数据挖掘

[中图分类号] R-056 [文献标识码] A [DOI] 10.3969/j.issn.1673-6036.2016.04.013

Clustering Analysis Algorithm Based on Attribute Reduction of Medical Data LI Xiao-xue, ZHENG Jing-chen, LI Ming, HAO Yu-wen, General Hospital of Chinese People's Armed Police Forces, Beijing 100039, China

[Abstract] It is very significant to predict the development status of a disease based on medical data. The paper introduces the concepts of clustering analysis and attribute reduction, and explains the typical clustering analysis process and a new clustering analysis algorithm based on attribute reduction. To reduce attributes of medical data can greatly enhance the efficiency of the clustering analysis process. As compared with existing classical clustering algorithms, this algorithm is proved to be effective.

[Keywords] Medical data; Data attribute reduction; Cluster analysis; Data mining

1 引言

近年来, 伴随各种先进医疗设备的应用, 在医疗诊断中产生了海量的数据。人们在进行体检或诊断疾病中几乎所有的信息都是以数据的形式进行展现。随着时间的推移, 这些数据将不断地增加。这些海量的体检数据除了能够帮助及时判断当前病人的体征外, 也可以帮助预测疾病的发展趋势或者发

现新的疾病, 以便提前做好防护措施, 保证人们的健康生活。数据容量及规模大大超过人类的直接处理能力, 在进行疾病预测的过程中, 为了准确地理解和表示这些数据, 需要利用计算机进行自动而有效的处理^[1]。因此, 本文首先对现有的数据进行必要的聚类操作, 将具有相同特征的数据分类, 简化数据分析的过程, 提高数据处理的效率。

2 聚类算法概述

2.1 聚类分析

聚类分析就是以某种标准将数据分为不同的数据组, 保证每组内部数据之间相似性最大而不同组之间的数据差异最大^[2]。可以人为的方式将少量的

[修回日期] 2016-02-25

[作者简介] 李晓雪, 博士, 副主任医师; 通讯作者: 郝昱文, 硕士, 高级工程师。

[基金项目] 发改办高技(2013)2140号; 卫计委公益性行业科研专项(项目编号: 201302003)。

数据进行划分,但是对于已经超出人力承受能力的海量数据,必须借助计算机进行处理。聚类分析算法就是通过计算机按照某种标准自动地将数据分为不同的组。聚类分析可以看成是一种无监督学习,在进行聚类的过程中只依赖数据的属性来区分相似程度。聚类分析算法作为一种有效的数据分析算法,广泛用在数据挖掘、机器学习以及图像识别等领域。目前聚类算法主要分为层次化聚类方法、划分式聚类方法、基于密度的聚类方法、基于网格的聚类方法、基于核的聚类算法、基于谱的聚类方法、基于模型的聚类方法、基于遗传算法的聚类方法、基于 SVM 的聚类方法、基于神经网络的聚类方法等^[3-4]。

2.2 属性约简

在进行数据聚类的过程中,不同的属性可能具有不同的重要性,有些属性对于聚类起着至关重要的作用,而部分属性可能对聚类结果没有任何影响。为了找出这些属性的重要性,一般的方法是将某个属性去掉之后再考察没有此属性聚类结果如何变化。如果去掉这个属性对聚类结果无影响,则此属性重要性比较小,反之重要性较大。属性约简是粗糙集理论研究的重要内容之一,是在保证数据有效性的基础上对数据的属性进行约简,消去不重要的属性。目前,属性重要度计算的方式主要有 3 种:基于属性依赖度、基于信息量和基于区分矩阵中属性出现频率^[5]。属性约简对于提高数据处理的效率具有重要的意义,可极大地减少后续聚类算法的运行时间。对数据进行聚类可以有效地处理数据,但是现有的聚类算法中,大部分聚类算法都是采用基于数据全部属性进行分类。在数据进行聚类的过程中,部分属性对数据聚类的结果没有显著的影响甚至无直接影响。因此在进行聚类分析之前,可以对数据的属性进行简化,去掉不必要的属性,从而提高聚类算法运行效率。

3 基于属性约简的聚类分析算法

3.1 典型的聚类分析流程

给定一个对象集合 $X = \{x_1, x_2, \dots, x_n\}$, 假

设每个对象 $x_i (i=1, \dots, n)$ 含有 m 个特征,在此用向量的方式表示对象的特征, $x_i = (l_1, l_2, \dots, l_m)$, 聚类分析的过程就是根据对象的特征来分析对象之间的相似程度,根据某种聚类决策准则获得聚类结果。聚类的结果用 $c = \{c_1, c_2, \dots, c_k\}$, 聚类结果满足以下条件^[6]: $c_i \neq \Phi, i=1, 2, \dots, k, \bigcup_{i=1}^k c_i = X; c_i \cap c_j = \Phi, i \neq j, i, j=1, 2, \dots, k$ 。

典型的聚类过程可以描述为:(1) 准备数据。主要是对数据进行简单处理,包括对非数值数据进行数值化或归一化等操作。(2) 提取数据特征。针对数据中属性的有效性,选择对聚类分析影响较大的属性或根据数据属性重要性设置属性权重,保证后续相似度计算的有效性。(3) 计算数据属性的相似度、选择合适的距离函数或者是自己构建计算数据距离的数据函数进行数据之间相似度的计算。(4) 聚类。步骤 3 中数据之间相似度进行聚类,使组内数据之间的相似度最大,组间的相似度最小。(5) 聚类结果评估。是指对聚类结果进行评估;通过观察数据的聚类结果,对数据聚类分析算法的有效性做出评估^[7]。

3.2 新的基于属性约简的聚类分析算法

为提高聚类算法的效率,简化计算的过程,本文提出一种基于医疗数据的属性约简聚类分析算法。在算法中首先利用部分样本空间数据计算数据的属性权重,删除权重比较小的属性,根据样本空间数据属性的简化结果简化所有数据的属性,而后利用现有的方法再次进行聚类分析。基于医疗数据的属性约简聚类分析算法的流程,见图 1。基于医疗数据的属性约简聚类分析算法可以描述为:(1) 选取样本空间。随机选取一定量数据 k_0 作为样本空间。(2) 获取属性数据方差。扫描样本中的每个数据的属性,将数据的属性进行归一化操作并计算每个属性数据的方差。(3) 定义方差临界值 m_0 。定义方差的临界值 m_0 ,若数据属性方差值大于 m_0 ,则此属性为重要属性,若属性方差值小于 m_0 ,则此属性为次要属性,将此属性忽略。(4) 聚类。根据步骤 3 确认的属性对所有的数据进行聚类。

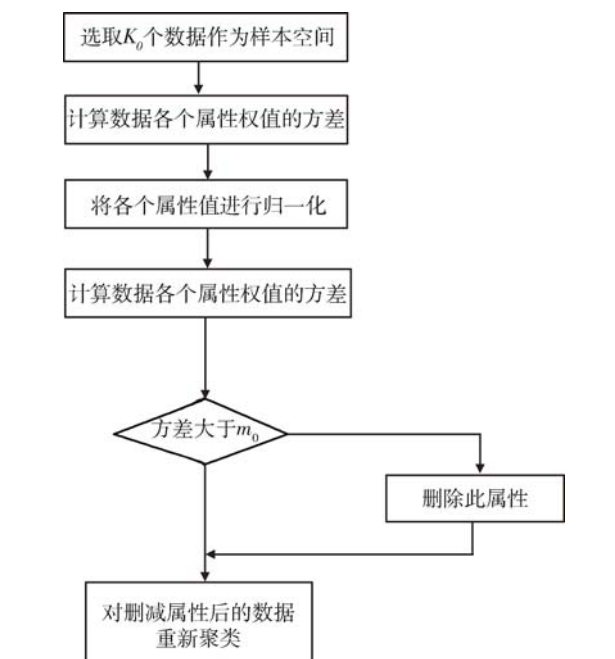


图 1 属性约简聚类分析算法流程

4 算法实验及分析

4.1 概述

为验证算法的有效性, 针对相同的数据, 在本文中分别用 k-means 方法和 PCM 方法进行比较, 首先利用 k-means 方法和 PCM 方法对医院医疗数据进行聚类比较, 记录现有算法对医疗数据的聚类结果和运行时间, 同时采用属性约简算法对医疗数据的属性进行约简, 然后再进行聚类, 比较聚类结果的有效性和运行时间。

4.2 实验数据

本试验所使用的数据集从医院的体检数据中获得, 该数据集的详细描述, 见表 1。可看出该数据集是多类高维数据。

表 1 数据集描述

数据集	样本总数	属性数
体检数据	1 000	12

4.3 实验结果与分析

在聚类的过程中, 不断调整算法中的方差基值 m 和样本个数 k , 通过不断调整, 找到合适属性约

简的基数, 同时调整 k 找到合适的样本个数。固定样本数 $k = \text{数据总个数} \times 80\%$, 不断调整参数 m , 争取找到合适的阈值 m_0 。由于在数据约简的过程中先对数据的属性值进行了约简, 先假定 m 的初始值为 0.001, 将数据属性值方差值小于 m 的值约简, 再进行聚类, 查看聚类算法聚错数据点个数。再不断对 m 进行 +0.001。利用聚错的样本数作为评价标准, 聚错的样本数指各类中聚错的样本数的和。实验的结果, 见表 2。

表 2 m 与聚错样本个数关系

m	聚错样本个数	m	聚错样本个数
0.001	80	0.006	97
0.002	80	0.007	100
0.003	95	0.008	123
0.004	95	0.009	135
0.005	95	0.010	160

由于设定 m 在每次递增的过程中只增加 0.001, 可能在递增一次没有对属性进行简化, 所以出现了不同 m 值相同的聚错样本个数现象。从表 2 可以看出, 在 $m_0 = 0.007$ 的时候, 聚错样本个数小于等于总样本数的 10%, 在本文中认为 10% 的样本空间即为样本最大的容忍度, 因此在后续计算 k 的数值的时候均设置 m_0 为 0.007。通过以上步骤确定 m_0 , 设 $m_0 = 0.007$, 不断改变 k 的值, 从而判断 k 值的变化对聚类结果的影响。具体样本比例与聚错率的关系, 见图 2。在 $k = 65\%$ 数据总数时, 聚类的错误率开始小于 10%, 本文设定 $k = 65\%$, 在后续算法比较的过程中, 假定 $m_0 = 0.007$, $k_0 = 65\%$ 。为证明算法有效性, 设定好 m_0 和 k_0 后, 分别针对 k-means 和 PCM 方法进行属性约简。以聚错样本数、平均准确度以及运行时间作为比较的标准。其中平均准确度的计算方法如下: 设原数据集有 k 个类, 用 C_i 表示第 i 类, n_i 为 C_i 这类中样本的个数, m_i 为 C_i 中聚类正确的个数, 则 m_i/n_i 为类 C_i 中的精度 ($i = 1, \dots, k$)。则平均准确度为 P 的计算如式 (1):

$$P = \frac{1}{k} \sum_{i=1}^k \frac{m_i}{n_i} \quad (1)$$

算法聚类的效果比较, 见表 3。由于基于属性约简的聚类分析算法在进行聚类分析算法前对数据

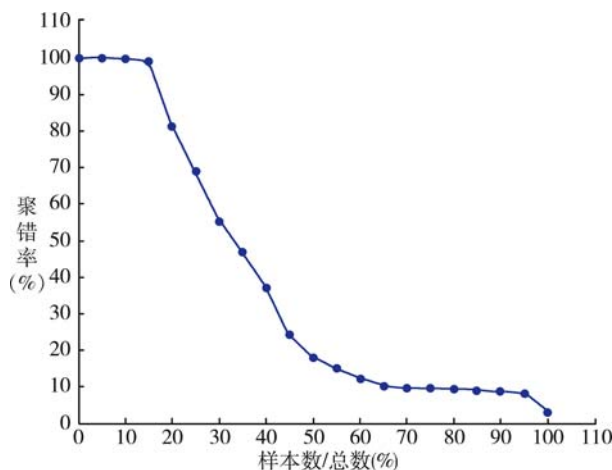


图2 样本比例与聚错率关系

的属性进行了化简,因此聚类过程中算法的运行时间明显降低。但是由于在聚类过程中忽略了部分属性,虽然算法的平均准确度有所下降,但是在可接受的范围内。

表3 聚类效果比较

所用算法	聚错的 样本数	平均 准确度	运行 时间(秒)
k-means	75	0.592 62	93
属性约简算法(k-means)	105	0.571 47	75
PCM	159	0.606 26	135
属性约简算法(PCM)	172	0.605 26	112

5 结语

通过计算机技术聚类分析算法进行医疗检测数据的分析过程中,减少医护人员不必要的工作。而通过对数据属性进行约简,在计算的过程中避免一些不必要的操作,提高算法的运行效率。在保证聚类结果有效性的基础上,显著提高了对数据的分析能力。

参考文献

- 1 贺海平,李捷,贺明. 基于 SQL Server 2005 数据库仓库构建与数据挖掘 [J]. 中国数字医学, 2008, (3): 58-59.
- 2 刘伟. 基于综合评判的蚁群聚类分析 [J]. 医学信息学杂志, 2014, 35 (7): 42-45.
- 3 Jing L, NgM K, Huang J Z. An Entropy Weighting K-means Algorithm for Subspace Clustering of High-dimensional Sparse Data [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19 (8): 1026-1041.
- 4 黄亚明,李新伟,崔雷,等. 科技期刊互引关系研究-Kohonen 神经网络和系统聚类法应用比较 [J]. 医学信息学杂志, 2007, 28 (5): 440-444.
- 5 何磊. 一种改进的基于区分矩阵的属性约简算法 [J]. 成都大学学报: 自然科学版, 2014, (33): 135-137.
- 6 李智玲,胡彧. 改进的属性约简算法在数据挖掘中的应用研究 [J]. 计算机技术与发展, 2012, 22 (10): 47-50.
- 7 任永功,于戈. 一种多维数据的聚类算法及其可视化研究 [J]. 计算机学报, 2006, 28 (11): 1861-1865.
- 8 杨明. 一种基于一致性准则的属性约简算法 [J]. 计算机学报, 2010, 33 (2): 231-239.
- 9 孟祥军,马志庆. 云计算在医学方面的发展与应用 [J]. 通讯世界, 2015, (11): 106.
- 10 詹庆. 云计算与智慧医疗系统建设 [J]. 信息系统工程, 2013, (30): 145-146.
- 11 张鑫,王连根,胡海荣. 云技术在医院信息化管理中的应用 [J]. 医学信息学杂志, 2014, 35 (5): 38-42.
- 12 焦铜,侯俐. 善用物联网技术打造城乡智慧医疗平台 [J]. 科技传播, 2013, (18): 12-13.
- 13 李萍. 云计算与大数据时代医院信息化的三个转变 [J]. 中国医院管理, 2013, 33 (12): 80-81.
- 14 刘天娇,刘成彦,林巧民. 移动医疗信息系统的设计与实现 [J]. 福建电脑, 2014, (10): 18-19, 42.
- 15 王鑫. 基于云计算的移动医疗服务平台的构建 [J]. 电子世界, 2013, (13): 17.
- 16 吕美丹,祝锡永,傅慧,等. 基于 GAE 云计算的区域移动医疗服务平台 [J]. 计算机系统运用, 2013, 22 (6): 29-33.

(上接 42 页)