

相容关系三支聚类的治略效果评估研究

张天麒, 李英梅

(哈尔滨师范大学计算机科学与信息工程学院, 黑龙江 哈尔滨 150025)

摘 要: 三支决策是一种符合人类认知的“三分而治”模型, 衡量“治”的效果需基于特定的“分”法。现有的对“治”的研究往往基于等价类进行三分。本文基于相容关系进行三支聚类, 根据聚类假设对数据对象进行“治”略, 提出一种测量治略效果的算法——3WC-BC。通过“治”前后的比较表明“治”略的意义和有效性。实验结果表明, 三支决策中“治”略的研究具有较大意义, 能够为选择合理的或最大效益的“治”略选择提供方法, 与传统的离散表格索引的选择方法比较减少了因索引错误而造成的误差。

关键词: 三支决策; 相容关系; 三支聚类; 行动; 行动规则

中图分类号: TP391 **文献标识码:** A

Measure Effectiveness of Action and Strategy in Three-way Clustering Based on Compatible Relationship

ZHANG Tianqi, LI Yingmei

(School of Computer Science and Information Engineering, Harbin Normal University, Harbin 150025, China)

Abstract: The three-way decisions are a trisecting-and-acting model that conforms to human cognition. Measuring the effectiveness of action and strategy needs to consider the specific way of secting. The existing research on action is often based on equivalence classes for three points. An algorithm of 3WC-BC is proposed for measuring effectiveness of acting in three-way clustering based on compatible relationships. The comparison between before and after action and strategy shows the significance and effectiveness of action and strategy. Experimental results show that the study of action and strategy in three-way decisions has a great significance and further provide a method for choosing reasonable or most effective actions, which reduces deviation caused by indexing error compared with traditional discrete table index method.

Keywords: three-way decisions; compatible relation; three-way cluster; action; rule of action

1 引言(Introduction)

三支决策是从粗糙集理论逐渐演变成的一种符合广义思考的“分而治之”模型^[1,2], 在已有的一系列相关理论中, 三支决策可以形式化的表示为“trisecting-and-acting”这样一个两步走的活动^[3-6]。目前在三支决策的研究领域中多数的研究集中于对三分法的研究^[7-11], 对在不同区域采取适合的行动策略的研究相对较少。在本文中, 将三支决策的思想与行动规则的策略结合起来^[12], 建立一个三支决策行动策略的模型。

在本文中, 主要考虑对三支决策中治略模型的研究, 并使用在有限成本下获得最大效益的决策模型来检验其有效性。本文是按以下逻辑展开的: 第2章对三支治略前后三支聚

类的变化情况进行了描述, 同时对三支治略的有效性进行了检验。第3章对所做的实验进行了简要介绍并对提出的算法性能进行了评估。

2 治略模型(Action model)

本文提出的治略模型分为三部分, 一是在对数据集进行基于相容关系的三支聚类, 二是对聚类结果的三支治略, 三是将治略后的聚类结果与原结果进行对比。

在已有理论的支撑下, 本章提出一种基于相容关系的三支聚类结果的治略模型对效益和成本进行分析, 通过治略前后的对比, 评估治略的有效性。

定义1: 一个决策表可以用公式(1)表达:

$$S=(OB, AT, \{V_a\}, \{I_a\}) \quad (1)$$

其中, $AT=A_s \cup A_f \cup \{d\}$, $\{V_a | a \in AT\}$, $\{I_a | a \in AT\}$ 。

令 OB 是一个非空有限的对象集合, AT 是由三个属性子集构成的非空有限集, A_s 代表稳定属性, A_f 代表可变属性, d 代表决策属性。 V_a 是每个属性 a 的非空值集, $a \in AT$, I_a 是一个 OB 到 V_a 的映射, 对于每一个对象 $x \in OB$, 总有一个属性 $a \in AT$, 一个值 $v \in V_a$, $I_a(x)=v$ 代表了对象 x 的属性 a 的值为 v 。

定义2: 设 D 是对象集 OB 上的关系度量, $\delta(\geq 0)$ 是给定的阈值, 满足 D 是自反的(当且仅当 $\forall x \in OB$, 有 $D(x,x) \leq \delta$); D 是对称的(当且仅当 $\forall x,y \in OB$, 如果 $D(x,y) \leq \delta$, 则有 $D(y,x) \leq \delta$)。则称 D 为 OB 上的一个相容关系, OB 是关系 D 下的一个相容集^[13]。相容度可以理解为一对概念差异信息的程度, 即两对象之间的相容程度。

本文使用余弦相似度^[14,15]对两个数据对象那个见的距离进行衡量, 假设对象 x 和 y 的属性向量分别为 $(x_1, x_2, x_3, \dots, x_n)$, $(y_1, y_2, y_3, \dots, y_n)$, 则它们的余弦相似度可以由公式(2)计算得到。

$$\cos \theta = \frac{(X_1, X_2, \dots, X_n) \cdot (Y_1, Y_2, \dots, Y_n)}{\sqrt{\sum_{i=1}^n (X_i)^2} \times \sqrt{\sum_{i=1}^n (Y_i)^2}} \quad (2)$$

第一步, 对数据进行预处理, 使用逻辑回归得到对决策属性影响最大的两个可变属性——BP和CP。选取每个属性中的最大值和最小值, 然后用最大值减最小值得到一个数值 m , 最后用该属性的所有值分别除以 m , 预处理过程可以用公式(3)表示:

$$V_{a(final)} = \frac{V_a}{m} \quad (3)$$

其中, $m = V_{\max} - V_{\min}$ 。

第二步, 计算两个对象间的余弦值作为相容度进行考察, 从而进行三支聚类的后续步骤。在以往的三支聚类中, 阈值往往是由相关领域专家给出的, 这在实际应用中很多时候不易得到。因此本文中首先对数据集在选取时要对决策属性进行一个初步的处理^[16], 在数据集中, 每个对象的决策属性可以被划分为正域结果, 边界域结果, 负域结果这三种情况。在正域结果中选取一个对象作为中心对象, 计算该数据对象与三个域中的其他所有数据对象的余弦相似度, 选取每个域比较后的最大值(即相似度最高的值)作为阈值, 据此方法产生三个阈值—— α 、 β 、 γ 。

第三步, 根据选取的三个阈值 α 、 β 、 γ 且满足 $\alpha < \beta < \gamma$, 可以构造三个域:

$$\begin{aligned} \text{POS} &= \{x, y \in OB | D(x,y) < \alpha\} \\ \text{BND} &= \{x, y \in OB | \alpha \leq D(x,y) \leq \beta\} \\ \text{NEG} &= \{x, y \in OB | \beta < D(x,y) < \gamma\} \end{aligned}$$

三个域分别对应上文所定义的正域结果对象所在的域, 边界域结果对象所在的域, 负域结果对象所在的域。如果 x 与 y 两个对象之间的相容度小于 γ , 则称两者之间存在相容关系。

根据三个域的特点可以构造三个关于行动策略的子集——DES、UND、IND。分别代表渴望的行动策略, 不渴望的行动策略, 不产生效益的行动策略,

在三支治略前首先要进行的是基于相容关系的三支聚类, 选取合适的阈值 α 、 β 、 γ 且满足 $\alpha < \beta < \gamma$, 算法的基本思想如下:

①构造相容簇的POS域: 从数据集中选取任一决策属性值为0的数据对象当作中心对象, 然后找出所有满足与中心对象相容度小于 α 的数据对象, 此时检查选取的数据对象, 若所有数据对象的决策属性都为0, 对所有选取的数据对象进行标记, 进行下一步, 否则重新选取中心对象。

②构造相容簇的BND域: 从数据集中找出所有满足与中心对象相容度大于 α 、小于 β 且决策属性为1的数据对象, 标记选取的所有对象并进行下一步。

③构造相容簇的NEG域: 从数据集中找出所有满足与中心对象相容度大于 β 、小于 γ 且决策属性为2的数据对象, 标记选取的所有对象并进行下一步。

④检查数据集中是否还有决策属性为0且未被标记的数据对象, 若存在, 则重复第一步, 若不存在, 则检查数据集中是否还有未被标记的对象, 若还有, 则标记为离散对象。

接下来要进行基于三支聚类的治略, 根据定义描述, 给定DES、UND、IND。可以确定的是, 在一个相容簇中, 所有不属于正域的数据对象都拥有其DES, 本文的治略算法给定的行动也都是满足DES的行动。使用 A_{xy} 表示从 x 域到 y 域的行动, 则用 A_{ND} 表示从NEG域到POS域的DES, 用 A_{BP} 表示从BND域到POS域的DES, 给出3WC-AS算法, 算法的基本思想如下:

①查找DES: 从三支聚类得到的相容簇集SC中取出一个相容簇, 找出所有DES。

②对边界域中的数据对象进行治略行动: 根据已知的DES, 将该相容簇边界域中的所有数据对象使用行动转移至正域。

③对负域中的数据对象进行治略行动: 根据已知的DES, 将该相容簇负域中的所有数据对象使用行动转移至正域。

④检索相容簇集SC中是否还有相容簇, 若存在, 返回第一步; 若不存在, 结束算法, 得到一个经过治略的新的相容簇集。算法描述如下:

上文中经过三支治略, 得到了一个新的相容簇集SC, 为与原相容簇集区分, 使用SC'来表示。下面根据基于相容关系的聚类和三支治略的方法。

对于拥有行动规则的数据对象, 定义: $\text{Optional} = \{x | \exists y, r_x \rightarrow r_y \in \text{DES}\}$ 。每一个 $x \in \text{Optional}$ 都是如上式所示拥有渴望行动方向的行动的。使用 C_{ij} 和 B_{ij} 来表示 x_i 的第 j 个可选策略的成本和效益, 使用 $a_{ij} \in [0, 1]$ 来表示 x_i 的第 j 个可选策略是否可行。

该模型采取基于动态规划的策略。令 $f(i, k)$ 表示对第 i 个行动对象在有限成本 k 下的最大效益, 通过推导得出 $f(i, k)$ 的一个完整公式, 即公式(4)。

$$f(i, k) = \begin{cases} 0 & i = 0 \\ \max_{j=0, \dots, n_i} \left\{ \begin{aligned} &f(i-1, k - C_{ij}) \\ &+ B_{ij} \mid C_{ij} \leq k \end{aligned} \right\} & i \neq 0 \end{cases} \quad (4)$$

根据基于相容关系的三支聚类与三支治略的过程, 可以设计出3WC-BC算法, 如下:

输入: D —关系度量; S —对象集; α 、 β 、 γ —阈值; C_a —给定成本。

输出: B —最大收益; a_{ij} —行动步骤。

步骤1: 根据相容关系确定POS域、BND域、NEG域中的元素。

步骤2: 找到所有的DES。

步骤3: 遍历每个DES, 求其成本与效益。

步骤4: 根据步骤3得到的成本与效益构造效益表与行动表。

步骤5: 根据效益表与行动表求得不同成本下的最大效益及其行动步骤。

步骤6: 算法结束, 返回结果。

3 实验结果(Experimental result)

文中的所有实验都是运行在windows 10 64bit上, 采用i5 2.50GHz CPU和8GB RAM。实验数据全部由UCI数据库提供。在此使用其中一个克利夫兰数据集^[17]。需要设置三个实验, 实验一对数据集进行聚类分析, 保证“三分”的有效性; 实验二对治略前后相容簇中的数据对象转移情况进行分析; 实验三对算法得到的策略行动步骤与随机选取的策略行动步骤做比较, 确保算法的性能。

3.1 聚类分析实验

实验一对数据集进行基于相容关系的三支聚类, 对于心脏病诊断数据集来说, 使用逻辑回归^[18,19]计算出对策属性影响最大的两个可变属性——BP和CP指标。作为横纵坐标在坐标轴上进行聚类分析。在聚类分析时需要涉及对相容关系阈值的选取, 此时使用余弦相似度, 按照前文阐述的规则选取阈值。聚类结果如图1所示。

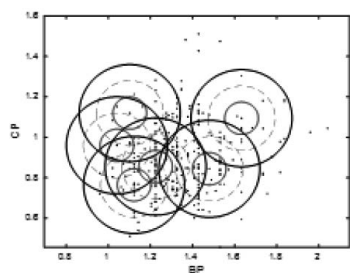


图1 心脏病数据聚类结果图

Fig.1 Data cluster result of heart disease

由图4可以看出, 基于相容关系的聚类允许一个数据对

象存在于多个相容集中, 这点和传统聚类有很大不同, 同时每个相容簇又有自己独有的数据对象, 与模糊聚类有着明显的不同。并且本文在一个相容簇中设置了POS域、BND域、NEG域, 采取“分而治之”的思想对在不同域中的对象采取不同的治略策略。在之后提出DES, 符合DES的数据对象作为Optional将拥有很多的可选策略, 而对应于不同行动策略的行动成本显然是不同的, 因此使用之后提出的算法进行计算, 求其最优的行动方案。

3.2 治略有效性度量实验

实验二依照文中治略的方法对实验一得到的聚类结果进行了治略, 原数据对象经过治略中对数据对象进行移动的一系列行动, 心脏病数据集中数据对象在不同域的转移情况如图2所示, 图中对POS域、BND域、NEG域, 以及其他区域中的数据对象在给定成本为0、274、589时各域中的对象数量进行了表示, 从图中可以看出给定成本为0时聚类初始状态下各个域中都有数据对象; 给定成本为274时, BND域中的对象转移的最多, 体现了治略的有效性。

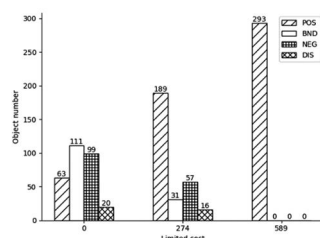


图2 医疗数据对象转移比较图

Fig.2 Medical objects transfer status

4 结论(Conclusion)

本文对基于相容关系的三支聚类治略的有效性进行了评估, 通过策略的成本和效益分析, 设计了三个算法模型对三支聚类的治略效果进行了评估。算法中使用分段函数进行策略行动成本与效益的检索, 避免因小数而出现检索错误的情况。

对心脏病数据集的三个实验结果表明, 基于相容关系的三支聚类不同于传统的聚类方法, 增加了灵活性; 通过三支治略有效地提高了相容簇的质量, 并将数据对象从负域或边界域转移到了正域; 在有限成本下求最大收益的算法具有理想的效果, 相比于随机选取行动, 算法能够更早的达到期望的目的。接下来的研究方向将研究属性和策略行动成本之间的关系, 可以考虑使用属性的一部分来获取相应的部分效益, 同时需要的成本也相对较低。在实际问题中, 某些属性对结果的影响可能是决定性的。因此, 需要对属性进行更有效的加权处理, 并使之通用性更强。同时, 不同的策略间可能存在相关的隐藏约束, 在研究中可以继续考虑放宽本文关于策略行动模式的约束, 加强行动策略的泛化能力。

参考文献(References)

- [1] Jiang C, Wu J, Li Z. Adaptive thresholds determination for saving cloud energy using three-way decisions[J]. Cluster Computing, 2018(1):1-8.

- [2] Yang X, Li T, Fujita H, et al. A unified model of sequential three-way decisions and multilevel incremental processing[J]. Knowledge-Based Systems, 2017.
- [3] Yao Y. Three-Way Decisions and Cognitive Computing[J]. Cognitive Computation, 2016, 8(4): 543–554.
- [4] Hu B Q. Three-way decisions space and three-way decisions[M]. Elsevier Science Inc, 2014.
- [5] 贾修一, 商琳, 周献中, 等. 三支决策理论与应用[M]. 南京大学出版社, 2012.
- [6] 李金海, 邓硕. 概念格与三支决策及其研究展望[J]. 西北大学学报: 自然科学版, 2017, 47(3): 321–329.
- [7] Deng X, Yao Y. Decision-theoretic three-way approximations of fuzzy sets[J]. Information Sciences, 2014(279): 702–715.
- [8] 汪璐, 贾修一, 顾雁因. 三支决策贝叶斯网络分类器[J]. 南京大学学报(自然科学), 2016, 52(5): 833–843.
- [9] 刘盾, 李天瑞, 李华雄. 粗糙集理论: 基于三支决策视角[J]. 南京大学学报(自然科学), 2013, 49(5): 574–581.
- [10] Gao C, Yao Y. Determining Thresholds in Three-Way Decisions with Chi-Square Statistic[C]. International Joint Conference on Rough Sets. Springer, Cham, 2016.
- [11] Jing-Tao Y, Herbert J P. A game-theoretic perspective on rough set analysis[C]. 2008国际知识技术论坛, 2008.
- [12] Raś, Zbigniew W, Tsay L S, et al. Tree-based algorithms for action rules discovery[J]. Studies in Computational Intelligence, 2009: 153–163.
- [13] 万仁霞, 王立新, 刘振文, 等. 一种基于相容关系的聚类算法[J]. 计算机应用研究, 2009, 26(4): 1302–1304.
- [14] 陈大力, 沈岩涛, 谢槟竹, 等. 基于余弦相似度模型的最佳教练遴选算法[J]. 东北大学学报(自然科学版), 2014, 35(12): 1697–1700.
- [15] Wang D, Lu H, Bo C. Visual Tracking via Weighted Local Cosine Similarity[J]. IEEE Transactions on Cybernetics, 2017, 45(9): 1838–1850.
- [16] Zhang X, Miao D. Reduction target structure-based hierarchical attribute reduction for two-category decision-theoretic rough sets[J]. Information Sciences, 2014, 277(2): 755–776.
- [17] Gao C, Yao Y. Actionable Strategies in Three-way Decisions[J]. Knowledge-Based Systems, 2017: 183–199.
- [18] Kianifard F. Logistic Regression: A Self-Learning Text[J]. Technometrics, 1995, 37(1): 116.
- [19] 毛毅, 陈稳霖, 郭宝龙, 等. 基于密度估计的逻辑回归模型[J]. 自动化学报, 2014, 40(1): 62–72.

作者简介:

张天麒(1995–), 男, 硕士生. 研究领域: 机器学习, 粗糙集理论.

李英梅(1970–), 女, 硕士, 教授. 研究领域: 计算机软件与理论. 本文通讯作者.

(上接第12页)

6 结论(Conclusion)

数据的准确性是Web日志挖掘中最重要的基础和前提, 使用web服务器日志, 挖掘分析用户行为特征, 为网站业务拓展提供数据支撑, 已经得到了广泛的应用^[10]. 比如常用的第三方网站统计分析工具: Webtrends、Google分析、百度统计等. 虽然第三方工具方便、高效, 但是并不利于特色业务的针对分析. 所以使用Hadoop平台开发专用web日志分析系统, 具有积极的现实意义.

本文提出了基于Hadoop的Web日志分析平台, 充分利用Hadoop处理海量数据的优势, Hive在开发上的便利, 有效提高了挖掘效率, 可以根据不同的用户业务, 定制日志分析模块的功能代码, 方便进行业务扩展. 例如: 将用户历史访问行为进行分析后, 将分析结果构建成行为特征库, 再实时分析用户当前的访问记录, 就可以开发异常行为入侵检测系统.

参考文献(References)

- [1] 陈明. MapReduce分布编程模型[J]. 计算机教育, 2014(11): 6–11.
- [2] 刘常伟. 基于Hadoop的Web日志数据分析系统的设计与实现[D]. 哈尔滨工业大学, 2017: 13–14.

- [3] 任建新. 基于Hadoop平台的大数据应用系统架构的研究与实现[D]. 北京邮电大学, 2014: 16–17.
- [4] 王慧娟. 基于云计算的会计大数据分析平台构建研究[J]. 山西财经大学硕士论文, 2015(9): 22–27.
- [5] 陈龙. 基于Hive的水利普查数据仓库[J]. 计算机与现代化, 2018(08): 12–20.
- [6] 程苗, 陈华平. 基于Hadoop的Web日志挖掘[J]. 计算机工程, 2011(06): 11–37.
- [7] 朱鹤祥. Web日志挖掘中数据预处理算法的研究[D]. 大连: 大连交通大学, 2010: 11–13.
- [8] 夏靖波, 韦泽鲲, 付凯, 等. 云计算中Hadoop技术研究与综述[J]. 计算机科学, 2016, 43(11): 6–11.
- [9] 于兆良, 张文涛, 葛慧, 等. 基于Hadoop平台的日志分析模型[D]. 计算机工程与设计, 2016, 37(2): 338–344.
- [10] 杨怡玲, 管旭东, 陆丽娜. 一个简单的日志挖掘系统[J]. 上海交通大学学报, 2000(7): 35–37.

作者简介:

何璇(1993–), 女, 硕士生. 研究领域: 数据分析与处理.

马佳琳(1974–), 女, 硕士, 副教授. 研究领域: 电子商务应用.