

基于样本选择的启发式属性约简方法研究

杨习贝^{1,2} 颜 旭¹ 徐苏平¹ 于化龙¹

(江苏科技大学计算机科学与工程学院 镇江 212003)¹ (南京理工大学经济管理学院 南京 210094)²

摘 要 属性约简是粗糙集理论的核心研究内容之一。借鉴于贪心策略的启发式算法是求解约简的一种有效技术手段。传统的启发式算法使用了决策系统中的所有样本,但实际上每个样本对约简的贡献程度是不同的,这在一定程度上增加了启发式算法的时间消耗。为解决这一问题,提出了一种基于样本选择的启发式算法,该算法主要分为 3 步:首先从样本集中挑选出重要的样本;然后利用选取出的样本构建新的决策系统;最后利用启发式算法求解约简。实验结果表明,新算法能够有效地减少约简的求解时间。

关键词 信息系统,样本选择,粗糙集,属性约简

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.1.009

New Heuristic Attribute Reduction Algorithm Based on Sample Selection

YANG Xi-bei^{1,2} YAN Xu¹ XU Su-ping¹ YU Hua-long¹

(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)¹

(School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China)²

Abstract Attribute reduction is one of the important topics in rough set theory. Heuristic attribute reduction algorithm with greedy strategy is one of the widely used approaches to compute reduction. Traditional heuristic algorithm uses all of the samples in the information system. However, it should be noticed that for a data set, different samples contribute different importance to find reduction, and it follows that the time efficiency of the heuristic attribute reduction algorithm suffers from the redundant samples. To solve such problem, a heuristic attribute reduction algorithm based on sample selection was proposed. The algorithm is composed of three stages. Firstly, the most informative samples are selected. Secondly, a new information system is formed by using these selected samples. Finally, one of the reductions can be computed by using heuristic algorithm. The experimental results show that the proposed algorithm can efficiently reduce the computational time.

Keywords Information system, Sample selection, Rough set, Attribute reduction

1 引言

粗糙集理论^[1]是波兰学者 Pawlak 提出的一种用于处理不精确、不确定知识的数学工具。目前,粗糙集理论已被成功应用于人工智能、数据挖掘、模式识别等众多研究领域^[2-7]。从理论与应用的视角来考虑,属性约简是粗糙集方法研究的核心问题之一。众所周知,在数据集中,每个属性(特征)的重要性程度是不尽相同的,因而所谓属性约简,就是考虑在一些度量标准的前提下(如近似质量、熵、近似分布)选择出数据集中重要的属性并剔除冗余属性。经历了 30 余年的发展,目前粗糙集理论中常用的属性约简算法大致可归为 2 类:穷举算法与启发式算法。尽管穷举法可以求得信息系统的所有约简,如分辨矩阵方法^[8]及回溯方法^[9,10]皆属于穷举法范畴,但是其由于计算过程过于复杂,并不适合大尺度复杂数据的

约简;利用贪心策略设计的启发式算法虽然仅能求得数据集的一个约简,但其由于求解的时间效率较高而备受青睐。

值得注意的是,已有的很多穷举算法与启发式算法将数据集中所有的样本都视作是同等重要的,这成为了进一步提高约简求解效率的瓶颈。为解决这一问题,王熙照等人^[11]将样本选择的方法引入了基于分辨矩阵的属性约简过程中,该方法的特点是选择数据集中对分类影响比较重要的样本构成新的数据集来求解约简,该思想可以极大地提高约简求解的时间与空间效率。然而遗憾的是,在构造新的数据集后,文献^[11]依然利用分辨矩阵这一 NP-Hard 方法来进行约简的求解。从这一角度出发,笔者提出了基于样本选择的启发式属性约简算法,选择出对分类影响较大的样本进行启发式约简求解,相比于分辨矩阵方法来说,这可以进一步提高算法的时间效率。样本选择的概念最早由 Hart 提出,典型的算法有

到稿日期:2015-03-28 返修日期:2015-06-06 本文受国家自然科学基金(61100116, 61305058),江苏省青蓝工程中青年学术带头人人才项目,中国博士后科学基金项目(2014M550293)资助。

杨习贝(1980—),男,博士后,副教授,主要研究方向为粗糙集理论、粒计算与机器学习, E-mail: zhenjiangyangxibei@163.com; 颜 旭(1994—),男,主要研究方向为粗糙集理论, E-mail: huaianyanxu@163.com; 徐苏平(1991—),男,硕士生,主要研究方向为粗糙集理论、机器学习, E-mail: supingxu@yahoo.com; 于化龙(1982—),男,博士,副教授,主要研究方向为机器学习、生物信息学, E-mail: yuhualong@just.edu.cn.

K-Means 均值聚类算法^[12]、压缩近邻算法(CNN)^[13]等。文中使用 K-Means 均值聚类算法进行样本选择。

本文第 2 节简要介绍粗糙集的相关知识;第 3 节提出了基于样本选择的启发式属性约简算法,并把该算法与传统的启发式属性约简算法(HAAR)、基于样本选择的分辨矩阵属性约简算法(ISDMAR)^[11]进行对比分析;最后总结全文。

2 基本知识

形式化地,一个信息系统 IS 可以表示为一个四元组 $IS = \langle U, AT, V, f \rangle$, 其中 U 表示所有样本的集合,称为论域; AT 表示描述样本的属性集合; $V = \bigcup_{a \in AT} V_a$ 为所有属性的值域,其中 V_a 为属性 a 的值域; $f: AT \times U \rightarrow V$ 为信息函数,表示 $\forall x \in U$, 有 $f(a, x) \in V_a$ 。特别地,若 $AT = C \cup D, C \cap D = \emptyset$ (其中 C 为条件属性集合, D 为决策属性集合), 则 IS 信息系统也被称为决策系统, 记为 DT 。

定义 1 给定一决策系统 $DT, \forall A \subseteq C$, 定义一个二元关系:

$$IND(A) = \{(x, y) \in U^2 : \forall a \in A, f(a, x) = f(a, y)\} \quad (1)$$

称 $IND(A)$ 为由属性集 A 生成的不可分辨关系。

定义 2 给定一决策系统 $DT, \forall A \subseteq C, \forall X \subseteq U, X$ 基于不可分辨关系 $IND(A)$ 的下近似集合 $\underline{A}(X)$ 定义为:

$$\underline{A}(X) = \{x \in U : [x]_A \subseteq X\} \quad (2)$$

其中 $[x]_A = \{y \in U : (x, y) \in IND(A)\}$ 表示 x 的等价类。

定义 3 给定一决策系统 DT , 设 $A \subseteq C$, 决策属性 D 关于 A 的正域定义为:

$$POS_A(D) = \bigcup_{i=1}^k \underline{A}(X_i) \quad (3)$$

其中, $U/D = \{X_1, X_2, \dots, X_k\}$ 是由决策属性 D 所导出的论域 U 上的划分。

3 属性约简

属性约简是粗糙集理论研究中的核心问题之一。所谓属性约简, 就是考虑在一些度量标准的前提下, 删除其中不相关或冗余的属性, 以达到简化知识表示的目的。

定义 4 给定一决策系统 DT , 设 $A \subseteq C$, 若有 $POS_A(D) = POS_C(D)$, 则称 A 为 DT 的一个正域保持约简。

3.1 基于启发式的属性约简算法

在决策系统 DT 中, $U/D = \{X_1, X_2, \dots, X_k\}, \forall A \subseteq C, \forall a_i \in A, a_i$ 的重要程度为:

$$Sig_{in}(a_i, A, D) = \frac{|\sum_{j=1}^k \underline{A}(X_j) \oplus \underline{A-\{a_i\}}(X_j)|}{|U|} \quad (4)$$

其中, $X \oplus Y$ 表示集合 X 与集合 Y 的对称差。由式(4)可以看出, $sig_{in}(a_i, A, D)$ 反映了从当前属性集 A 中删除 a_i 后正域的变化程度。同理, 可以定义:

$$Sig_{out}(a_i, A, D) = \frac{|\sum_{j=1}^k \underline{A}(X_j) \oplus \underline{A \cup \{a_i\}}(X_j)|}{|U|} \quad (5)$$

其中, $Sig_{out}(a_i, A, D)$ 反映了从当前属性集 A 中加入属性 a_i 后正域的变化程度。

根据上述属性重要度的定义, 在决策系统中, 传统的启发式属性约简算法如算法 1 所示。

算法 1 启发式属性约简算法(HAAR)

输入: $DT = \langle U, C \cup D, V, f \rangle$

输出: 约简 Red

步骤 1: 计算 $POS_C(D)$ 。

步骤 2: $Red \leftarrow \emptyset$ 。

步骤 3: $\forall a_i \in C$, 计算属性 a_i 的重要程度 $Sig_{in}(a_i, C, D)$ 。

步骤 4: 选择 a_j, a_j 满足 $Sig_{in}(a_j, C, D) = \max\{Sig_{in}(a_j, C, D) : \forall a_j \in C\}$, $Red \leftarrow a_j$, 计算 $POS_{Red}(D)$ 。

步骤 5: 若 Red 不满足 $POS_C(D) = POS_{Red}(D)$, 则重复以下循环, 否则转至步骤 6。

(1) $\forall a_i \in (C - Red)$, 计算重要程度 $Sig_{out}(a_i, C, D)$;

(2) 选择 a_j, a_j 满足 $Sig_{out}(a_j, Red, D) = \max\{Sig_{out}(a_j, Red, D) : \forall a_j \in (C - Red)\}$, $Red = Red \cup \{a_j\}$;

(3) 计算 $POS_{Red}(D)$ 。

步骤 6: $\forall a_i \in Red$, 若 $Red - \{a_i\}$ 满足 $POS_{Red}(D) = POS_{Red - \{a_i\}}(D)$, 则 $Red = Red - \{a_i\}$ 。

步骤 7: 输出 Red 。

文献[11]的研究成果表明, 决策系统中的样本并不是同等重要的。其实, 在计算属性约简的过程中, 有些样本对于属性约简的贡献度很小, 甚至毫无贡献。显然, 在这种含有冗余样本的决策系统中进行启发式属性约简会在一定程度上增加算法的时间消耗。如果将那些对分类重要或对约简贡献度大的样本挑选出来组合成为一个新的决策系统, 然后在这个新的决策系统上求解约简, 就可以提高传统的启发式约简算法的时间效率。基于这种思想, 本文提出了基于样本选择的启发式属性约简算法。

3.2 基于样本选择的启发式属性约简算法

使用决策系统中对分类影响较大的样本来进行启发式属性约简, 可以大大缩小样本的规模, 并在一定程度上提高约简的时间效率。一般来说, 这些对于分类比较重要的样本都分布在边界区域上^[11]。在此, 采用 K-Means 均值聚类算法对这些样本进行选择。新的属性约简方法可以分为 3 步: 首先, 利用 K-Means 均值聚类算法对原始决策系统进行样本聚类, K 为决策类的类别数; 其次, 选择距离聚类中心较远的样本, 也就是那些靠近边界的样本, 并将选择出的样本组合成为一个新的决策系统; 最后, 在新的决策系统中执行启发式属性约简算法。具体的算法描述如算法 2 所示。

算法 2 基于样本选择的启发式属性约简算法 (ISHAAR)

输入: $DT = \langle U, C \cup D, V, f \rangle$

输出: 约简 Red

步骤 1: 从 U 中选择 K 个样本作为 K 个簇的中心, K 由数据集的类别数决定(即决策类的个数)。

步骤 2: 分别计算剩下的样本到 K 个中心点的距离, 将样本分别划归到距离最近的簇。

步骤 3: 根据聚类结果, 重新计算 K 个簇各自的中心。

步骤 4: 将 U 中全部样本按照新的中心重新聚类。

步骤 5: 重复步骤 4, 直到聚类结果不再发生变化。

步骤 6: 将各个簇中样本到中心点的距离大于簇中所有样本到中心点的平均距离的样本加入新的样例集 U' 中, 得到新的决策系统 $DT' = \langle U', C \cup D, V, f \rangle$ 。

步骤 7: 利用算法 1 计算 DT' 的属性约简 Red。

步骤 8:输出 Red。

在算法 2 中, K-Means 均值聚类的时间复杂度为 $O(KTU)$, 其中, K 为簇数, T 为迭代次数, U 为论域 U 中样本的个数。在论域 U 和论域 U' 中执行启发式属性约简算法的时间复杂度分别为 $O(U \times C^2)$ 和 $O(U' \times C^2)$ 。其中, U' 为选择出的样本个数, C 为条件属性的个数。由此可知, 算法 2 的时间复杂度为 $O(KTU + U' \times C^2)$ 。

将算法 1 和算法 2 的时间复杂度进行比较。假设 $U \times C^2 \geq KTU + U' \times C^2$, 则 $KTU \leq U \times C^2 - U' \times C^2$, 即 $KT \leq (1 - \frac{U'}{U}) \times C^2$ 。因此, 当条件 $KT \leq (1 - \frac{U'}{U}) \times C^2$ 成立时, $O(U \times C^2) \geq O(KTU + U' \times C^2)$ 。条件 $KT \leq (1 - \frac{U'}{U}) \times C^2$ 在数据库上一般都能满足。

因此, 算法 2 的时间复杂度要优于算法 1。

3.3 实验分析

为了验证算法的有效性, 选取了 8 组 UCI 数据集, 并进行了 2 组实验。这 8 组数据集的基本信息如表 1 所列。实验 1 是将本文提出的算法 (ISHAAR) 与传统的启发式属性约简算法进行对比分析。实验 2 是将本文提出的算法 (ISHAAR) 与王熙照等人^[11]提出的基于样本选择的分辨矩阵属性约简算法 (ISDMAR) 进行对比分析。两组实验的实验环境皆为 PC 机, 双核 2.1GHz CPU, 2GB 内存, Windows 7 操作系统,

Matlab 2012b 实验平台。

表 1 实验数据的基本信息

数据集	对象个数	属性个数	决策类个数
Breast	699	9	2
Iris	150	4	3
Monk'S	1711	6	2
Glass	214	10	7
Tic-Tac-Toe	958	8	2
Wine	178	13	3
Spect Heart	267	22	2
Wine-Quality	4898	11	11

实验结果如表 2 所列。从表 2 可以看出, 本文提出的算法在时间消耗上明显低于传统算法。例如在 Iris 数据集上, 传统的算法用时为 0.11s, 而本文提出的算法仅需用时 0.01s。本文提出的算法通过样本选择, 剔除了对分类贡献度较小的样本, 极大地缩小了样本的规模; 且在所有的数据集上, 本文提出的算法的分类精度都要大于或等于传统算法的分类精度。所以, 与传统的启发式属性约简算法相比, 本文提出的算法更加有效。此外, 本文提出的算法求取约简的时间效率要远高于 ISDMAR, 例如, 在 Wine 数据集上, 使用 ISDMAR 求取约简需要耗时 252.18s, 而使用本文提出的算法仅需耗时 0.04s; 且在大多数数据集上, 本文提出的算法的最佳分类精度要高于 ISDMAR, 只是在 Breast 和 Tic-Tac-Toe 数据集上略低于 ISDMAR。

表 2 3 种算法的对比分析

数据集	对象个数			最小约简的属性个数			分类精度/%			时间/s		
	HAAR	ISHAAR	ISDMAR	HAAR	ISHAAR	ISDMAR	HAAR	ISHAAR	ISDMAR	HAAR	ISHAAR	ISDMAR
Breast	699	203	203	3	3	3	96.13	98.03	98.52	3.8	0.60	76.87
Iris	150	14	14	4	1	1	96	100.00	98.89	0.11	0.01	0.26
Monk'S	1711	823	823	6	6	6	64.47	71.81	71.78	15.62	3.34	5.18
Glass	214	26	26	7	4	5	61.68	69.23	67.64	1.06	0.23	1.60
Tic-Tac-Toe	958	450	450	8	8	8	77.45	78.22	78.31	11.54	3.34	1.708
Wine	178	86	86	1	1	1	100	100	100	0.08	0.04	252.18
Spect Heart	267	128	128	11	10	16	79.03	92.97	84.49	9.2	3.56	72.68
Wine-Quality	4898	1721	1721	5	5	5	55.55	57.99	57.99	102.98	91.78	432.87

结束语 为了进一步提高启发式算法求解约简的时间效率, 本文提出了基于样本选择的启发式属性约简算法, 并从实验分析的角度将其与传统的启发式算法进行了对比分析。由实验结果可知, 基于样本选择的启发式属性约简算法可以在保证分类精度基本不变或有所提高的前提下, 极大地减少算法运行的时间消耗。因此本文提出的算法具有一定的现实意义。

在本文工作的基础上, 笔者下一步将把样本选择的思想引入模糊环境中, 对模糊粗糙集的快速约简方法进行深入探讨。

参 考 文 献

[1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356

[2] Luo G Z, Yang X B. Limited dominance-based rough set model and knowledge reductions in incomplete decision system[J]. Journal of Information Science and Engineering, 2010, 26(6): 2199-2211

[3] Yang H L, Li S G, Wang S Y, et al. Bipolar fuzzy rough set model on two different universes and its application[J]. Knowledge-Based Systems, 2012, 35: 94-101

[4] Wang G Y, Zhang Q H, Ma X A, et al. Granular computing models for knowledge uncertainty[J]. Journal of Software, 2011, 22(4): 676-694

[5] Chen J K, Li J J. An application of rough sets to graph theory[J]. Information Sciences, 2012, 201(19): 114-127

[6] Celotto E, Ellero A, Ferretti P. Short-medium term tourist services demand forecasting with Rough Set Theory[J]. Procedia Economics and Finance, 2012, 3(12): 62-67

[7] Yeh C C, Lin F, Hsu C Y. A hybrid KMV model, random forests and rough set theory approach for credit rating[J]. Knowledge-Based Systems, 2012, 33(3): 166-172

[8] Skowron A, Rauszer C. The discernibility matrices and functions in information systems[M]. Kluwer Academic Publishers, 1992: 331-362

[9] Min F, He H P, Qian Y H, et al. Test-cost-sensitive attribute reduction[J]. Information Sciences, 2011, 181(22): 4928-4942

[10] Yang X B, Qi Y S, Song X N, et al. Test cost sensitive multi-granulation rough set: Model and minimal cost selection[J]. Information Sciences, 2013, 250(11): 184-199

[11] Wang X Z, Wang T T, Zhai J H. An attribute reduction algo-

rithm based on instance selection[J]. Journal of Computer Research and Development, 2012, 49(11): 2305-2310(in Chinese)

王熙熙, 王婷婷, 翟俊海. 基于样例选取的属性约简算法[J]. 计算机研究与发展, 2012, 49(11): 2305-2310

- [12] Wu X D, Kumar X, Quinlan J R, et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems, 2008, 14(1): 1-37

(上接第 21 页)

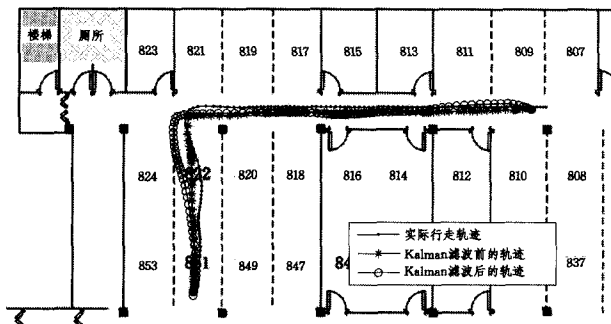


图 8 本文轨迹生成方法在 Kalman 滤波前后的轨迹对比

结束语 本文提出一种融合多模传感器的室内实时高精度轨迹生成方法, 将 Wi-Fi 定位与传感器定位相结合, 最后对得到的位置信息进行 Kalman 滤波, 得到移动终端在室内的实时运动轨迹, 解决了 Wi-Fi 定位不连续、周期长以及传感器定位的误差累积问题。在实验中, 首先得到系统中的参数, 然后将本文的方法分别与纯 Wi-Fi 定位以及惯性导航技术、行人航迹推算技术等不同的传感器定位技术进行对比, 结果表明本文方法得到的用户室内移动轨迹的精确度比同类方法的更高。

参考文献

- [1] Park M H, Kim H C, Lee S J. Implementation results and service examples of GPS-Tag for indoor LBS and message service[C]// 2013 15th International Conference on Advanced Communication Technology (ICACT). IEEE, 2013: 367-370
- [2] Enge P, Misra P. Special Issue on Global Positioning System[J]. Proceedings of the IEEE, 1999, 87(1): 3-15
- [3] Sun Chao-qi, Song Bing-long, Jia Bin, et al. The Research and Application of the Vehicle Monitoring System Based on Beidou Second Generation[J]. Science & Technology Association Forum, 2012(10): 77-78(in Chinese)
- 孙超奇, 宋秉龙, 贾斌, 等. 基于北斗二代的车辆监控系统研究与应用[J]. 科协论坛, 2012(10): 77-78
- [4] Want R, Falcao V, Gibbons J. The Active Badge Location System[J]. ACM Transactions on Information Systems, 1992, 10(1): 91-102
- [5] Priyantha N B, Chakraborty A, Balakrishnan H. The Cricket Location-Support System [C]// Proceedings of the 6th ACM International Conference on Mobile Computing and Networking (MobiCom 2000). Boston, Massachusetts, August 2000: 32-43

- [13] Hart P E. The condensed nearest neighbor rule [J]. IEEE Transactions on Information Theory, 1968, 14(3): 515-516

- [14] Miao D Q, Hu G R. A heuristic algorithm for reduction of knowledge[J]. Journal of Computer Research and Development, 1999, 36(6): 681-684(in Chinese)

苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6): 681-684

- [6] Ni L M, Liu Yun-hao, Lau Y-C, et al. LANDMARC: Indoor Location Sensing Using Active RFID [J]. Wireless Networks, 2004, 10(6): 701-710

- [7] Gonzalez-Castano F J, Garcia-Reinoso J. Bluetooth location networks [C] // Global Telecommunications Conference, 2002 (GLOBECOM '02). IEEE, 2002: 233-237

- [8] Brunato M, Bartiti R. Statistical learning theory for location fingerprinting in wireless LANs [J]. Computer Networks, 2005, 47: 825-845

- [9] Chung J, Donahoe M, Schm C, et al. Indoor Location Sensing Using Geo-Magnetism [C]// ACM MobiSys. 2011

- [10] Sun Yu-shan, Dai Tian-jiao, Zhao Zhi-ping. Error Analysis of Dead-reckoning Navigation System for Autonomous Underwater Vehicle[J]. Ship Engineering, 2010, 32(5): 67-72(in Chinese)

孙玉山, 代天骄, 赵志平. 水下机器人航位推算导航系统及误差分析[J]. 船舶工程, 2010, 32(5): 67-72

- [11] Levi R W, Judd T. Dead reckoning navigational system using accelerometer to measure foot impacts; US, US5583776 A[P]. 1995

- [12] Lee S, Kim B, Kim H, et al. Inertial Sensor-Based Indoor Pedestrian Localization with Minimum 802. 15. 4a Configuration[J]. IEEE Transactions on Industrial Informatics, 2011, 7(3): 455-466

- [13] Zhang Ming-hua. Study on WLAN Based Indoor Location Estimation Technology[D]. Shanghai: Shanghai Jiao Tong University, 2009 (in Chinese)

张明华. 基于 WLAN 的室内定位技术研究[D]. 上海: 上海交通大学, 2009

- [14] Liu Jun-fa, Gu Yang, Chen Yi-qiang, et al. Incremental Localization in WLAN Environment with Timeliness Management[J]. Chinese Journal of Computers, 2013, 36(7): 1448-1455(in Chinese)

刘军发, 谷洋, 陈益强, 等. 具有时效机制的增量式无线定位方法[J]. 计算机学报, 2013, 36(7): 1448-1455

- [15] Song M R, Moon J Y, Bae S H. Efficient Indoor Positioning By Hybrid Algorithm [J]. CES-CUBE, 2013, 25: 118-122

- [16] Zhao Yong-xiang, Zhou Huai-bei, Chen Miao, et al. Application of Kalman Filter in Indoor Positioning System for Real-Time Tracking[J]. Journal of Wuhan University(Natural Science Edition), 2009, 55(6): 696-700(in Chinese)

赵永翔, 周怀北, 陈淼, 等. 卡尔曼滤波在室内定位系统实时跟踪中的应用[J]. 武汉大学学报(理学版), 2009, 55(6): 696-700