基于预测决策同态理论的故障数据优化挖掘算法

陆青梅[1] 褚玉晓[2]
(中北大学电子与计算机科学技术学院 太原 030051)[1]
(郑州大学西亚斯国际学院电子信息工程学院 郑州 451150)[2]

摘 要 在一些大型的智能机械设备环境中,由于故障数据种类不断增加,形成了一个强冗余数据干扰的环境,这样的环境下,由于故障冗余关联规则的存在造成挖掘耗时。在充分研究关联挖掘算法的基础上,提出一种基于预测决策同态理论的强冗余数据挖掘算法,该算法通过对冗余关联中的数据以非同态信息增益惩罚因子构建同态区间,对区间内庞大的冗余关联数据进行关联约束,保证关联数据在距离较近的同态区间内,在邻近区间中采用预测决策方法进行故障的最终确认。实验证明,这种方法能够提高冗余环境下故障数据挖掘的准确率,其计算成本不高,鲁棒性较强。

关键词 预测决策,同态区间,数据挖掘
中图法分类号 TP391.4 **文献标识码** A

Fault Data Optimization Mining Algorithm Based on Theory of Prediction Decision Homomorphism

LU Qing-mei[1] CHU Yu-xiao[2]
(School of Electronic and Computer Science and Technology, North University of China, Taiyuan 030051, China)[1]
(School of Electronics and Information Engineering, Sias International University, Zhengzhou University, Zhengzhou 451150, China)[2]

Abstract In some large intelligent mechanical equipment environment, the fault data increases variety, forms a strong redundant interference environment, and in the environment, mining is time-consuming, because of the existence of unassociation rules. On the basis of full research of association mining algorithm, strong redundant data mining algorithm based on a prediction of the theory of decision homomorphism was proposed which constructs homomorphisms interval by the redundancy of the data with punish factor, constraints the interval of the huge redundant associated data correlation, ensures related data in the distance nears the homomorphism interval, and in the nearby interval, uses prediction methods of operation decision-making to make fault final confirmation. Experiments show that the method can improve the redundant environment, the accuracy of fault data mining, the calculation cost is not high, and it has a good robustness.

Keywords Forecast decision, Homomorphisms interval, Data mining

1 引言

随着数据挖掘技术应用得越来越广泛,数据挖掘算法的开发技术越来越成熟,尤其在一些大型的机械设备中,故障的处理早已实现了智能化;随着处理信息量的增长,故障信息存储数据库的存储能力也越来越强,一些不同属性、不同类型的机械故障数据存储在同一数据库中,这种现象称为混合存储,因此,快速准确的数据挖掘方法为大型机械故障数据的信息查找和开发提供了研究方法[1,2]。

现出现了不少相关的研究,Cooley R, Mobasher B等人[3]首次对数据挖掘做出了明确的定义,但是其没有把该方法应用到机械设备故障数据的挖掘中。Chen等人将数据挖掘技术应用于故障数据挖掘中,通过对故障日志数据进行分析和预处理,经过用户识别、会话识别后将用户会话分割成一系列的事务,通过关联规则分析,发现故障的最终原因[4,5]。我国国内的西安交通大学陆丽娜教授等人,通过基于事务法[6],研究了故障数据挖掘预处理过程和用户访问序列模式,在有向树模型的基础上,提出了一种用户浏览模式识别的故障数据挖掘方法。但是,这两种方法都存在一定的弊端,即过于依赖数据之间较为清晰的故障关联,一旦大量的干扰数据造成了大量的冗余数据关联,这两种方法都会陷入不收敛的情况,造成挖掘过程耗时[7,8]。

2 优化挖掘算法流程

基于上文中的概要分析,本文设计相应的基于预测决策同态理论的故障数据优化挖掘算法流程,如图1所示。

根据该流程图可知,基于预测决策同态理论的相关控制总体可分为3大步骤,描述如下:

Step1 对原始数据进行预处理,包括对数据的初始化,并且根据初始化的结果,对待挖掘数据进行有效的区间划分;

到稿日期:2012-09-10 返修日期:2012-12-10 本文受山西省2011年科学技术发展计划(20110321031)资助。
陆青梅(1979—),女,博士,讲师,主要研究方向为网络安全、数据挖掘,E-mail:wk_xiaolu@126.com;褚玉晓(1979—),女,博士,讲师,主要研究方向为算法分析与设计。

Step2　对小区域内的数据进行惩罚信息的约束，此阶段冗余关联效应可以得到有效避免；

Step3　通过预测决策方法对约束的故障数据进行挖掘，保证算法故障定位的准确性。

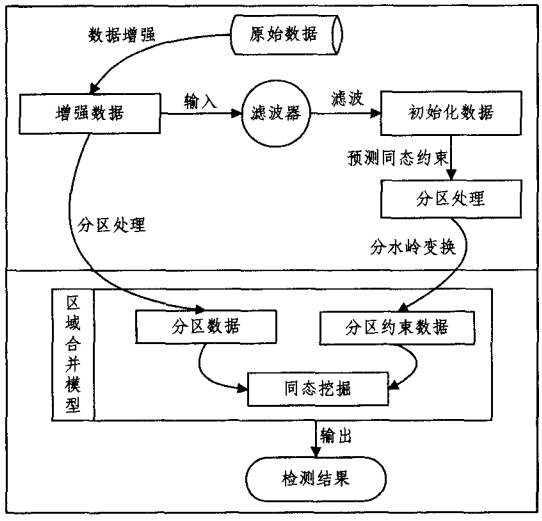

图 1　本文提出的基于预测决策同态理论的故障数据优化挖掘算法流程图

3　数据预处理与同态区间的划分

在介绍算法之前，先介绍本文提出的同态理论。同态的概率起源于生物学，在此表示数据具备某一种共同的属性，根据这种属性，能够对数据进行有效的聚类。与 Web 数据不同，机械故障数据有着自身固有的同态属性，其中，故障信息增益比可以很好地对故障数据进行有效的划分。但是由于大量冗余机械故障信息的存在，这种划分也存在着某种不确定性，需要加入特殊的约束条件才能保证区间划分的准确。

3.1　预处理与构建数据同态区间

本文运用直接惩罚信息增益的数量值进行区间构造，提出使用故障信息增益比作为冗余属性的分裂准则，对冗余故障数据属性进行惩罚，方法如下式所示：

$$Split_A(D)=-\sum_{j=1}^{v}\frac{|e_r|}{|D|}\times\log_2\left(\frac{e_r}{D}\right)$$

进行惩罚比率计算的公式如下：

$$GainRatio(A)=\frac{Gain(A)}{SplitI(A)}$$

求解故障信息增益比后，其作为故障区间划分的基础，可以有效避免因为故障冗余造成选择值不均匀、区间过宽、训练性能差的缺点。然后进一步使用 z 分数对故障数据的属性进行标准化，把故障数据标准化为一个均值为 0、标准差为 1 的一个 z 分数。

$$X'=(X-X')/S,S=\sqrt{\frac{\sum_{i=1}^{n}(X_i-X')^2}{n-1}}$$

可以证明，经过上述标准化的故障数据绝大部分都会包含在$[\overline{X}-3S,\overline{X}+3S]$的区间内，这样可以将区间以外的故障数据作为孤立点去除。对区间的数据进行进一步属性分裂时，处理的数据为离散型的数据，区间离散化数据应该满足离散划分点少、信息熵值小的特点。将机械设备的某项连续数据等宽划分为若干个区间，然后设定一个描述前后区间样本惩罚信息熵的绝对差值 m，如果小于 m 这个阈值，则样本区间继续合并，直到合并的值已经大于该给定的值 m。假设开始划分的区间为 k，防范如下：

$$m=\frac{I(s_1,s_2,\cdots,s_m)-\sum_{j=1}^{k}\frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s}I(s_1,s_2,\cdots,s_m)}{k-1}$$

如果前后区间的惩罚信息熵值计算差符合上述值，则将样本进行区间合并。经过以上分析，像水温故障这样的连续变量可以被划分为(小于 75)，(75,90)，(大于 90)3 个数值区间。有些因素的故障属性值无法使用数值来进行描述，但可以根据聚类的思想将其划分为不同的集合与簇，例如机械设备的天气情况可以划分为良好、一般、差等 3 个等级。设备运行性能分析的变量数众多，这些变量之间有些存在较大的相关性，比如设备运行温度与运行时间这样的一组变量，这些属性可以使用某一个或者几个属性的综合来表示，这样就可以将数据量庞大的属性进行进一步的规约，利用惩罚相关系数的分析来进行变量的规约，方法如下：

$$r=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})\sum_{i=1}^{n}(Y_i-\overline{Y})}}$$

求解每两个变量之间的惩罚相关系数，使用$|r|>0.8$的相关变量中的一个作为代表变量，如果 $0.5<|r|\leqslant 0.8$，则要根据分析需要进行取舍。进行数据划分的结果如图 2 所示。

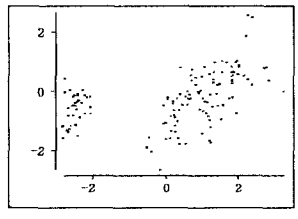

图 2　区间划分结果

3.2　故障数据的预测决策方法

在 3.1 节建立的同态区间的基础上，采用故障数据的预测决策方法进行故障数据的最终定位。该方法首先建立决策树结构。决策树是类似于数据流程图的一种树结构，每一个非树叶阶段的属性值都代表了该属性的一个测试即一个分裂标准属性测试，使用决策树进行属性的判别清晰明了，能够处理海量的历史数据。故障挖掘的决策树，以汽车系统的故障数据为例，建立的决策树如图 3 所示。

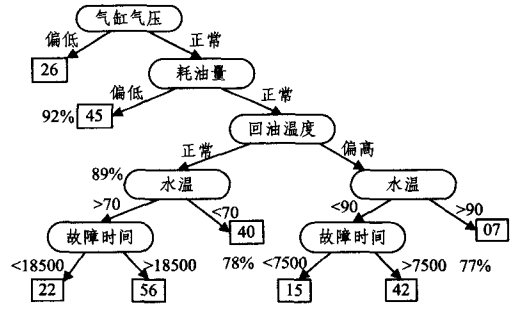

图 3　故障决策树

选择哪种属性的分裂准则是决策树预测效果的关键点，可以根据前面建立的惩罚信息熵的方法对属性进行计算，将其作为判别关联能力的度量。惩罚信息熵可以对整个信息源的不确定性进行度量，在机械设备的运行数据中，某种属性要作为预测属性，比如结构断裂或设备无法进气等，其自身可能的状态具有不同的发生概率，如果有 n 种状态分别表示为

$x_1, x_2, \cdots, x_n$，每一种结果出现的概率表示为 $P(x_1), P(x_2), \cdots, P(x_n)$，则有 $\sum_{i=1}^{n} p(x_i)=1$，则惩罚信息的熵可以表示为：

$$H(X)=p_1 I(x_1)+p_2 I(x_2)+\cdots+p_n I(x_n)$$

$$=-\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

将惩罚信息增益作为属性判别的标准，在每一个非叶节点进行故障数据测试，可以求出该样本的最大信息量。使用这种方法进行进一步的分类后，生成的决策树的深度最小，到达叶子节点的路径最短，提高了机械故障预测的准确度与速度。信息增益如下定义：

机械设备的某项数据集中一共有 m 类的样本，每一类的样本数量为 $p_i, i=1,2,\cdots,m$，某一个属性作为故障预测决策树的根，该属性假设为 A，具有 v 个值 $v_1, v_2, \cdots, v_v$，这些值将数据训练样本划分为 v 个子集，如果某一个子集 e_r 属于类 C_i 的概率为 p_i，可以使用 $\frac{|C_{i,D}|}{|D|}$ 进行估计，则该子集的信息熵可以表示为：

$$I(e_r)=-\sum_{i=1}^{m} p_i \log_2(p_i)$$

信息的期望为：

$$E(A)=\sum_{j=1}^{v}\frac{|e_r|}{|D|}\times I(e_r)$$

如果故障测试集的数量分别为 p, n，则这种决策树分类的信息量如下所示：

$$I(p,n)=-\frac{p}{p+n}\log_2\frac{p}{p+n}-\frac{n}{p+n}\log_2\frac{n}{p+n}$$

$$E(A)=\sum_{i=1}^{v}\frac{p_i+n_i}{p+n}I(p_i,n_i)$$

上述属性为根的决策树的信息增益表示为：

$$gain(A)=I(p,n)-E(A)$$

对每一个测试属性进行上述信息增益的计算，并且从大到小进行排列，形成属性的分裂树，对故障信息进行预测确认。

4 实验结果分析

4.1 实验数据

本文利用某公司 35 辆某类型机械车辆的 10000 条车辆运行发动机记录数据对该算法的可行性进行分析，整个实验仿真的平台为 Windows xp 下的 Spss Clementine 实验数据决策树构造平台。将整个样本根据车辆服役年龄属性进行分层，在每一层中随机抽取样本的单位。部分车辆发动机记录表如表 1 所列。

表 1 部分车辆发动机记录表

样本	时间	水温	油温	油温度	故障编号
10	235	80	66	正常	56
13	563	85	75	正常	45
16	354	92	85	正常	37
20	542	96	75	偏低	26
22	236	85	60	偏低	15
29	475	85	96	偏低	07
30	156	75	60	正常	42
35	489	80	85	偏低	36
36	600	90	93	正常	35
37	177	80	60	偏低	28
42	230	76	75	正常	40
45	560	85	85	偏低	17
50	623	88	95	偏低	16

本文将 7000 条车辆运行的发动机记录历史数据进行决策树构造后，使用 3000 条数据进行决策树的准确率判别，这些测试数据涵盖了 35 辆车辆的 1 个月的运行数据。本文使用以下的性能进行定量的分析，假设 I_M 代表某辆车的记录数，I_c 为预测的结果符合实际情况的记录数，则 $p=\frac{I_C}{I_M}$ 代表预测的准确率，故障数据随机分布图如图 4 所示。

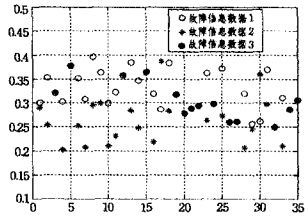

图 4 故障数据分布图

这里以上述的信号采集数据作为实验对象，设计了系统故障检测的仿真实验平台，来验证本文所提算法的有效性。

4.2 实验结果分析

首先通过实验得到车辆在正常工作时发动机的电流信号 TN，然后在平台运动过程中给平台施加一个额外的阻力（模拟故障发生的情况），得到一个故障模拟电流数据信号，设为 TF。通过加入大量的非阻力信号，模拟冗余数据关联环境。通过对强冗余关联环境下的故障电流信号的提取来检测故障信号。挖掘到的 TN、TF 信号，以及 TE 随时间的变化曲线分别如图 5 和图 6 所示。

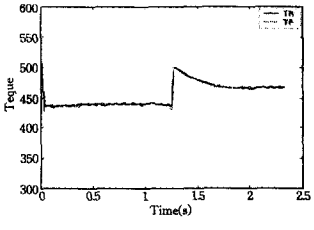

图 5 TN、TF 变化曲线

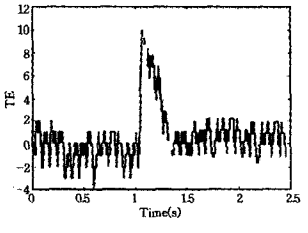

图 6 TE 随时间变化曲线

从图 6 可以看出，TE 信号中存在较多的干扰信息。图 7 和图 8 分别是采用优化前与优化后的两种方法对 TE 进行处理得到的数据信号曲线。

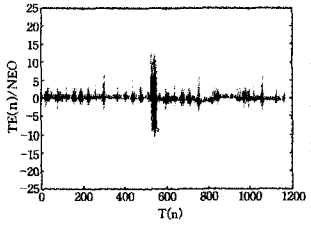

图 7 干扰状态下的数据描述结果曲线

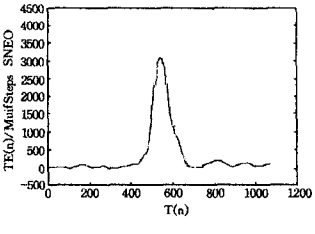

图 8 预测决策同态理论的故障数据优化挖掘算法曲线

比较图 7 和图 8 可以看出，本文方法下采集的数据能够更好地消除冗余关联的干扰，得到的故障数据趋势更加清晰明显，提取的效果更好。相比传统的方法，其采集信息的准确性明显更高。

结束语 本文提出一种基于预测决策同态理论的强冗余数据挖掘算法，该算法通过对冗余关联中的数据以非同态信息增益惩罚因子构建同态区间，对区间内庞大的冗余关联数据进行关联约束，保证关联数据在距离较近的同态区间内，在

邻近区间中采用预测决策方法进行故障的最终确认。实验证明，这种方法能够大大提高机械设备故障预测的准确率，对实际的机械维护有很强的指导意义。

参 考 文 献

[1] Su T, Dy J. A Deterministie Method for Initializing K-Means Clustering[C]// Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence. Boca Raton, Florida, 2009: 784-786

[2] 齐继阳，竺长安. 设备故障智能诊断方法的研究[J]. 仪器仪表学报，2009，27(10)：34-36

[3] Cooley R, Mobasher B, Srivastava J. Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns[R]. TR 97-021. University of Minnesota, Dept. of Computer Science, Minneapolis, 2011: 218-229

[4] 郭岩，白硕，于满泉. Web 使用信息挖掘综述[J]. 计算机科学，2005，32(1)：21-27

[5] Cooley R, Srivastava J. Grouping Web page references into transactions for mining world wide Web browsing patterns[C]// Proceedings of KDEX'97. NewportBeach, CA, USA, 1997: 2-7

[6] 田卫. RS-485 总线分支线短路故障检测技术[J]. 微电子学与计算机，2011，28(4)：116-117

[7] Mannila H, Toivonen H, Verkamo A I. Efficient algorithms for discovering association rules[C]//KDD-94: AAAI Workshop on Knowledge Discovery in Database. Seattle, Washington, 2009: 181-192

[8] Buchner A G, Mulvenna M D. Discovering Internet Marketing Intelligence Through Online Analytical Web Usage Mining[J]. SIGMOD Record, 2008, 27(4): 54-61

(上接第 221 页)

[19] Langleyand P. Selection of relevant features in machine learning [C]// Proceedings of A AAI Fall Symposium on Relevance. 1994

[20] Wang Y, Tetko I V, HallMark A, et al. Gene selection from microarray data for cancer classification-a machine learning approach[J]. Computation Biology and Chemistry, 2005, 29(1): 37-46

[21] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1-3): 389-422

[22] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Science, 1982, 11: 341-356

[23] 李衍达，孙之荣. 生物信息学基因和蛋白质分析的实用指南[M]. 北京：清华大学出版社，2000

[24] Li Ding-fang, Zhang Wen. Gene selection using rough set theory [C]// Rough Sets and Knowledge Technology 2006 (RSKT 2006). Lecture Notes in Artificial Intelligence, Chongqing, 2006, 4062: 778-785

[25] Skowron A, Komorowski J, Pawlak Z, et al. Rough sets perspective on data and knowledge[M]. Handbook of data mining and knowledge discovery. NewYork: Oxford University Press, 2002

[26] Banerjee M, Mitra S, Banka H. Evolutinary-Rough Feature Selection in Gene Expression Data[J]. IEEE Transaction on Systems, Man, and Cybernetic, Part C: Application and Reviews, 2007, 37: 622-632

[27] Momin B F, Mitra S, Datta G R. Reduct Generation and Classification of Gene Expression Data[C]// Proceeding of First International Conference on Hybrid Information Technology (ICHICT06). 2006: 699-708

[28] Valdes J J, Barton A J. Gene discovery in leukemia revisited: a computational intelligence perspective[C]// Proceedings of the 17th International Conference on Industrial & Engineering Applications of Artificial International Conference & Expert Systems. Springer Verlag, 2004: 118-127

[29] 苗夺谦. 粗糙集理论中连续属性的离散化方法[J]. 自动化学报，2001，27(3)：296-302

[30] 权光日，等. 连续属性空间上的规则学习算法[J]. 软件学报，1999，10(11)：1225-1232

[31] 叶东毅，黄翠微，赵斌. 基于逼近精度的一个粗糙集属性约简算法[J]. 福州大学学报：自然科学版，2000，28(1)：7-10

[32] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets[J]. International Journal of General Systems, 1990, 17: 191-209

[33] Zadeh L A. 模糊集合，语言变量及模糊逻辑[M]. 北京：北京科学出版社，1982

[34] Xu F F, Miao D Q, Wei L. Fuzzy-rough attribute reduction via mutual information with an application to cancer classification [J]. Computers & Mathematics with Applications, 2009, 57(6): 1010-1017

[35] Bhatt R B, Gopal M. On fuzzy-rough sets approach to feature selection[J]. Pattern Recognition Letters, 2005, 26(7): 965-975

[36] Hu Qing-hua, An Shuang, Yu Da-ren. Soft fuzzy rough sets for robust feature evaluation and selection [J]. Information Sciences, 2010, 180(22): 4384-4400

[37] Jensen R, Shen Qiang. Fuzzy-rough data reduction with ant colony optimization[J]. Fuzzy Sets and Systems, 2005, 149(1): 5-20

[38] Chen De-gang, Zhao Su-yun. Local reduction of decision system with fuzzy rough sets. Fuzzy Sets and Systems, 2010, 161(13): 1871-1883

[39] Priness I, Maimon O, Ben-Gal I. Evaluation of gene-expression clustering via mutual information distance measure, BMC Bioinformatics, 2007, 8: 111

[40] Chow T W S, Huang D. Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information [J]. IEEE Trans. Neural Networks, 2005, 16(1): 213-224

[41] 苗夺谦，王珏. 粗集理论中概念与运算的信息表示[J]. 软件学报，1999，2：113-116

[42] 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

[43] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of Max-Dependency, Max-Relevance, and Min-Redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238

[44] Maji P, Paul S. Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data [J]. Int. J. Approx. Reason, 2011, 52(3): 408-426

[45] West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles [C]// Proceedings of the National Academy of Science. USA 98, 2001(20): 11462-11467