

用自动模糊划分与改进 Apriori 算法生成 QAR 关联规则

乔永卫^a, 肖春景^b, 杨 慧^b

(中国民航大学 a. 工程技术训练中心; b. 计算机科学与技术学院 天津 300300)

摘 要: 针对属性粒度模糊划分需事先给定与 Apriori 算法效率低的问题, 提出基于自动模糊划分和改进 Apriori 算法的 QAR 关联规则生成方法。首先对 QAR 数据进行空缺值填补等预处理; 然后给出最佳聚类准则并根据给出的最佳聚类准则得到最佳聚类, 从而对 QAR 属性完成自动模糊划分及隶属函数的确定; 之后通过记录数据项位置及简化连接与剪枝过程来提高 Apriori 算法的效率; 并将其应用到 QAR 关联规则的生成过程; 最后通过品质和性能度量两方面的实验, 表明此方法在各方面的性能均优于经典方法。

关键词: 快速存取装置; 机载记录系统; 最佳聚类; 模糊划分; 关联规则

中图分类号: V19; TP131 文献标志码: A 文章编号: 1671-637X(2012)05-0036-06

Generation of QAR Association Using Rules Based on Automatic Fuzzy Partition and Improved Apriori Algorithm

QIAO Yongwei^a, XIAO Chunjing^b, YANG Hui^b

(Civil Aviation University of China, a. Engineering & Technical Training Center;

b. College of Computer Science & Technology, Tianjin 300300, China)

Abstract: Considering the problems that attribute granularity fuzzy partition should be given in advance and the Apriori algorithm has low efficiency, we proposed a method for generating association rules in Quick Access Recorder (QAR) based on automatic fuzzy partition and improved Apriori algorithm. First, preprocessing was made for QAR data by filling the vacancies value. Then, the optimum clustering criteria was given and an optimum clustering was obtained according to the criteria. Thus the fuzzy partition of QAR attributes was implemented and membership function was determined. It improved the efficiency of the algorithm by recording the location of the data items and simplifying the process of the pruning and connection. It was applied in the generation of the QAR association rules. Experiment was made on quality and performance of the method, and the result showed that the method is superior to the classical method in all aspects of performance.

Key words: Quick Access Recorder (QAR); airborne recording system; optimum clustering; fuzzy partition; association rules

0 引言

QAR(Quick Access Recorder)是飞机机载记录系统中的快速存储装置,由于其数据量大,涉及的参数多,参数的微小变化难以监控,使QAR无法得到及时有效的

分析,导致故障逐步恶化,最终造成严重的后果。因此,对QAR数据进行分析、挖掘对于飞行技术检查、安全评估、安全事件调查和飞机维护都有着重要的意义。目前西方发达国家对QAR数据的研究主要集中在故障预测与诊断方面,如著名的飞行操作质量保证(FOQA)项目。国内对QAR数据分析还处于起步阶段,如采用多元线性分析方法,利用QAR数据建立巡航阶段燃油油量模型^[1];利用小波和KPCA对QAR数据进行约简和降维^[2-3];根据不同的主题对数据进行存储^[4];并把QAR数据用于故障预测与诊断中^[5]。QAR数据属性间潜在的关联不仅能反映飞机的飞行状态,而且反映飞机的健

收稿日期: 2011-11-18

修回日期: 2011-12-25

基金项目: 国家自然科学基金项目(61103005); 国家自然科学基金与中国民航联合资助项目(61179063); 中央高校基本科研业务费(ZXH 2011B003)

作者简介: 乔永卫(1976—)男,山西祁县人,硕士,讲师,研究方向为飞行器发动机与机务维修。

康状况 关联规则能很好地挖掘出 QAR 数据属性的潜在联系,为故障诊断和预防性维护提供良好的理论支撑。经典 Apriori 算法是关联规则挖掘中最重要的方法,但存在产生大量候选项集和多次扫描数据库两个致命缺点,使其运行效率低、I/O 性能差,严重影响算法的普及和推广。为了提高算法的效率,文献[6-8]中提出一系列改进算法,对算法的连接和剪枝过程进行各种优化,主要从减少扫描数据的次数、减少候选集数量、减少计算每个候选集频率所需的时间等方面进行改进,使其性能得到一定程度的提高。但对数据属性粒度的划分,无论采用模糊集、网格、聚类技术还是分层、加权等处理^[9-15],都必须事先给定属性的模糊集划分和隶属函数或设定聚类类别,而这导致划分的模糊集不能与实际的数据相适应,不具有实际意义。针对以上问题,提出基于自动模糊划分与改进 Apriori 算法的关联规则生成方法,并将其应用到 QAR 数据中。它无需事先给定模糊划分或聚类个数,能更好地适应数据特性,在规则生成过程中提高算法的运行效率,改善其性能。

1 数据预处理

QAR 数据量大,存在采样频率不同,格式不统一等问题,针对这一现象,首先采用空缺值填补的数据预处理,使 QAR 数据格式统一。假设 $a_1, a_2, \dots, a_m, \dots, a_n$ 为某一属性值, a_r, a_q 分别为第一个和最后一个非空缺的属性值,则空缺值的填补公式为

$$\begin{cases} a_i = a_{i+1}, & 1 \leq i < r \\ a_{i+(i+j)/2} = \frac{a_i + a_j}{2}, & r \leq i < j \leq q \\ a_i = a_{i-1}, & q < i \leq n \end{cases} \quad (1)$$

式中: i, j 为最近两个非空缺值。从而得到格式统一,没有空缺值的 QAR 属性值。

2 模糊集自动划分

关联规则生成过程一般首先要对数据属性粒度进行划分,由于大部分聚类算法都要求事先给定聚类数、最小密度、初始值等参数,导致随着输入参数的不同数据集的划分形式也不同,而且在很多情况下得到的属性划分与实际数据特性不符,不具有实际意义。

2.1 最佳聚类准则

要想得到最佳聚类,首先给出聚类类别的允许取值范围 c ,并使之达到最优。由于 c 一般是由行业专家给出,这就能保证聚类结果符合属性特性,并且此方法适用于任何聚类算法,而不仅仅局限于本文使用的模糊 c -均值(FCM)聚类。

对于属性 X 和聚类类别取值 c, x_1, x_2, \dots, x_n 为各

属性值,则偏差 $\delta^2(X)$ 定义为

$$\delta^2(X) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 \quad (2)$$

式中: $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$ 为属性均值,它表示属性 X 的差异,它的值越大,说明 X 的差异越大,值越小,说明 X 中的值越相似。同理,聚类 i 的偏差为

$$\delta^2(X_i, \bar{X}_i) = \frac{1}{n_i} \sum_{k=1}^{n_i} (x_{ik} - \bar{X}_i)^2 \quad (3)$$

式中: $X_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$ 为第 i 类中的数据元素; \bar{X}_i 为 X_i 的均值。 c 个聚类的平均分配偏差为

$$\delta(X, \bar{X}) = \frac{\frac{1}{c} \sum_{i=1}^c \delta^2(X_i, \bar{X}_i)}{\delta^2(X)} \quad (4)$$

式中: \bar{X} 为 c 聚类平均值; $\delta(X, \bar{X})$ 表示聚类的平均紧密度,值越小,表明 c 个聚类中的元素值越接近,此值随着聚类之间距离的增大而增大。但最佳聚类的类间距离应尽可能大, c 个聚类间平均距离定义为

$$D(X, R) = \frac{1}{2c} \sum_{i=1}^c \sum_{j=1}^c |r_i - r_j| \quad (5)$$

式中: $D(X, R)$ 为类间距离的度量; R 为聚类中心集合; r_i, r_j 分别为聚类 i, j 的中心。 $D(X, R)$ 值越大,类间元素距离越大,聚类结果越好,因此期待此值最大化。

依据最佳聚类要求类内距离最小,类间距离最大的原则,得到最佳聚类的评价准则为

$$S(X, R) = \delta(X, \bar{X}) + \frac{1}{D(X, R)} \quad (6)$$

式中: $S(X, R)$ 值越小,表明聚类结果越好。求最佳聚类的目的就是要最小化 $S(X, R)$ 。对于任意给定的 c 运行 FCM,并利用最佳聚类准则来评估,当 $S(X, R)$ 取得最小值时得到最佳聚类数和聚类结果。

2.2 模糊集划分及隶属函数的确定

对于某一数据属性,利用最佳聚类准则找到最佳聚类数和聚类中心后,即可把属性值划分成 c 个模糊集,进而得到其模糊划分。

对于属性变量,可采用梯型、正态型或柯西型函数作为样本相似性量度的主要依据。本文采用三角型隶属度函数,并给出其相关定义,如图1所示。

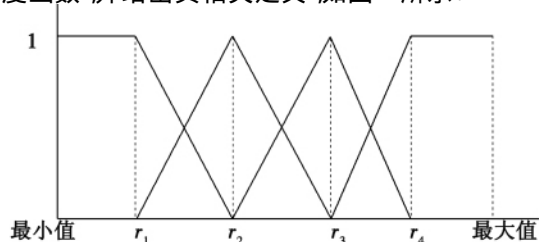


图1 三角形隶属函数

Fig. 1 Triangular membership function

$\{r_1, r_2, \dots, r_i, \dots, r_c\}$ 为聚类中心集合, 对于每一个模糊集的隶属函数定义为

$$\mu_i(x) = \begin{cases} 1, & \text{最小值} \leq x < r_1 \\ \frac{1}{r_i - r_{i-1}}(x - r_{i-1}), & r_{i-1} < x < r_i \\ \frac{1}{r_i - r_{i+1}}(x - r_i), & r_i < x \leq r_{i+1} \\ 1, & r_c \leq x \leq \text{最大值} \end{cases} \quad (7)$$

3 关联规则生成

3.1 数据离散化

对于某一点 x 其在各模糊集的隶属函数值分别为 $\mu_1(x), \mu_2(x), \dots, \mu_n(x)$,

$$\mu_m(x) = \begin{cases} 1, & \mu_m(x) = \max(\mu_1(x), \mu_2(x), \dots, \mu_n(x)) \\ 0, & \text{其他} \end{cases} \quad (8)$$

这样, 就把复杂的隶属度集合转化成简单的 0-1 矩阵, 能更方便地应用到关联规则生成过程, 且易于通过重新定义频繁项和项集分别表示的定义的方式来应用到更复杂的数据中。

3.2 改进 Apriori 算法

Apriori 算法在迭代过程中多次扫描数据库和产生大量的候选项集形成了算法的性能瓶颈。为了提高算法的效率, 在生成 1-项集时通过记录数据项位置使规则生成过程只需扫描一次数据库, 在从 $k-1$ -项集生成 k -项集过程中通过设定的阈值过滤候选项, 减少频繁项的生成, 提高运算效率, 改善其性能。

3.2.1 T_{ID} 集合的定义

设为数据库 D 中每个交易 T 设置一个唯一编号 T_{ID} , 则其项 k -项集定义为

$$R_k = \langle X_k, T_{ID}(X_k) \rangle \quad (9)$$

式中: $X_k = (I_{i1}, I_{i2}, \dots, I_{iq})$, $q < m$ 且 $I_{ij} \in I$; $T_{ID}(X_k)$ 为数据库中包含 X_k 的交易 T 的编号 T_{ID} 的集合, 则

$$T_{ID,S}(X_k) = \{T_{ID}; X_k \in T, \langle T_{ID}, T \rangle \in D\} \quad (10)$$

这样, 只在生成 1-项集时扫描一次数据库, 在迭代过程中利用编号集合 T_{ID} 进行频繁项集和规则的生成, 不再需要扫描数据库。

3.2.2 连接

在对两个 $k-1$ -项集连接生成 k -项集时, 只有当 $k-1$ -项集的前 $k-2$ 相等时才进行连接, 即对于 $L_{k-1}(i) = \langle X_{k-1}, T_{ID,S}(X_{k-1}) \rangle$ 和 $L_{k-1}(j) = \langle Y_{k-1}, T_{ID,S}(Y_{k-1}) \rangle$, 只有当 $X_{k-1}[k-2] = Y_{k-1}[k-2]$ 时,

$$L_{k-1}(i) \propto L_{k-1}(j) = \langle X_{k-1} \cup Y_{k-1}, T_{ID,S}(X_{k-1}) \cap T_{ID,S}(Y_{k-1}) \rangle = \langle X_k, T_{ID,S}(X_k) \rangle = R_k \in L_k \quad (11)$$

否则, 生成的 k -项集要么是非频繁项集, 要么是重复结构。这样, 简化了连接过程, 通过集合的并、交运算完成频繁项的生成, 减少了很多计算量, 大大提高了算

法的运算效率。

3.2.3 剪枝

在生成 $T_{ID,S}(X_k)$ 情况下, 因为候选项的最小支持度和最小支持数小于给定阈值时必定是非频繁项集, 所以在频繁项集生成的剪枝过程中, 对大量候选项可利用最小支持度和最小支持数进行过滤筛选来减小计算量。

k -项集的支持度和支持数分别定义为

$$D_{\text{sup port}}(R_k) = \frac{|T_{ID,S}(X_k)|}{|D|} = \frac{|\{T_{ID}; X_k \in T, \langle T_{ID}, T \rangle \in D\}|}{|D|} \quad (12)$$

$$N_{\text{sup num}}(R_k) = D_{\text{sup port}}(R_k) * |D| = |T_{ID,S}(X_k)| \quad (13)$$

式中: $|D|$ 为数据库中交易的总数。剪枝时仅当 $N_{\text{sup num}}(R_k)$ 小于其最小值时, 从 L_k 中删除 R_k 。

这样, 只在生成 1-项集时扫描一次数据库, 在迭代时不再需要扫描数据库, 降低了 I/O 负载, 使得频繁项集的发现速度大大提高; 利用最小支持数的阈值对候选频繁项过滤, 减小了频繁项集, 并对项目集连接过程进行简化, 仅仅使用集合的并、交运算即可完成, 不需要复杂的计算, 使得该算法不但易于实现且计算量小。

4 实验过程及结果分析

为了研究和比较算法的性能和优越性, 利用它对波音 737-800 飞机的实际 QAR 数据进行了实验。QAR 数据包含从起飞、巡航到降落的 11908 条, 每条包含 256 数据项的数据。因为 QAR 数据的采集频率不同, 如 ALTITUDE(ALT)、COMPUTED AIRSPEED(CAS) 等采集频率为 4 次/s, 而 SELECTED N1 INDICATED(N1)、SELECTED N2 ACTUAL(N2) 等为 2 次/s, SELECTED OIL PRESS(OIP) 为 1 次/s, SELECTED OIL TEMP(OIT) 为 0.5 次/s, 故首先对 QAR 数据进行空缺值填补的数据预处理, 并对 QAR 中描述发动机性能的 ALT, CAS, SELECTED EGT(T495)(EGT), SELECTED FUEL FLOW(FF), N1, N2 进行实验与分析, 利用 Matlab 仿真实现。

4.1 最佳聚类的确定

设定 FCM 允许的聚类取值范围为 $[2, 10]$, 得到部分不同属性的最佳聚类准则值, 如图 2 所示。

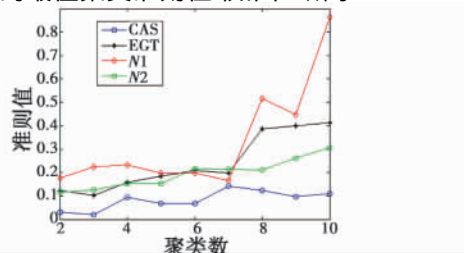


图 2 不同属性的最佳聚类准则值

Fig. 2 The criteria value of the best clustering for the different attributes

图2表示QAR属性值的聚类准则值随着聚类数的变化情况,可得到其最佳聚类数和聚类中心,如表1所示。由表1可知,每个属性具有不同最佳聚类数目,这样能使聚类效果更好,而如果指定聚类个数,必然存在着某些属性的聚类数不是最佳的,从而使得聚类结果不能很好地反映数据特性。

表1 各属性聚类数和聚类中心表

Table 1 The number of clusters and cluster centers for the different attribute

QAR 属性	聚类数	聚类中心
ALT/inch	2	2187.4; 28407
CAS/kn	3	45.358; 166.26; 271.79
EGT/°C	3	497.96; 616.56; 781.94
FF/(lb·h ⁻¹)	2	1003.7; 2904.9
N1/%	7	20.245; 30.544; 42.457; 60.754; 72.004; 83.864; 96.896
N2/%	2	62.3; 92.285

4.2 关联规则生成

在QAR关联规则的生成与分析过程中主要进行了质量和性能度量两方面的实验,并与指定聚类个数和经典Apriori算法相结合的经典算法进行了比较。在质量度量方面比较了两种方法在最小置信度一定,最小支持度变化和最小支持度一定,最小置信度变化两种情况下的频繁项数目、规则数目、平均支持度和平均置信度。在第2个实验中通过比较两种方法在上述两种情况下的执行时间,说明了两种算法的适应性。

4.2.1 品质度量

品质度量中首先给出最小置信度和最小支持度 $C_{\min, \text{conf}} = 0.6, S_{\min, \text{sup}} = [0.05 \ 0.3]$ 时实验结果,如图3所示。结果表明两种算法具有相似的特性,随着最小支持度的增加,频繁项集数、关联规则数减少,而平均支持度、平均置信度增加,且本文方法中各方面的性能均优于经典方法。图3a与图3b表明,本文方法用更少的频繁项集生成有用的关联规则,在最小支持度小于0.15时生成的关联规则数略少于经典方法,大于0.15时几乎相同,但由于本文方法的频繁项少,故计算量应更小,效率更高。从图3c和图3d看出,本文方法的平均支持度明显高于经典方法,平均置信度略高于经典方法,并在最小支持度为0.25时有较大的飞跃。

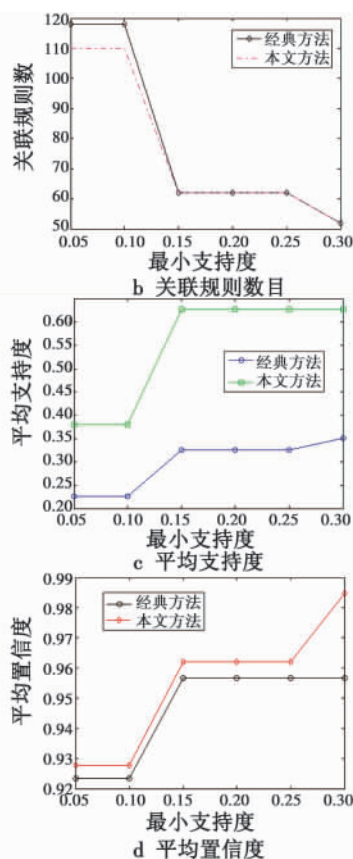
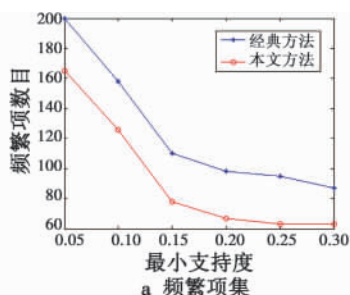
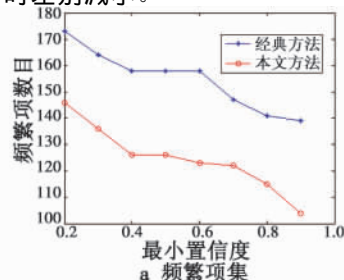


图3 不同最小支持度下的品质度量

Fig. 3 The quality measurement under different minimum support

图4给出了 $S_{\min, \text{sup}} = 0.2, C_{\min, \text{conf}} = [0.2 \ 0.9]$ 下的实验结果。由图4可以知道,结果与图3相似,随着最小置信度的增加,频繁项集、关联规则数减少,平均支持度、平均置信度增加,且各方面的性能仍然优于经典方法。

从图4a和图4c可以看出,本文方法生成的频繁项明显少于经典方法,而规则的平均支持度却明显高于它,表明本文方法用更小的频繁项集生成了平均支持度更高的规则,不但计算量更小且规则质量高。图4b和图4d表明,在最小置信度小于0.6的情况下,本文方法生成的关联规则数明显少于经典方法,平均置信度却明显高于经典方法,说明本文方法生成的规则置信度高,避免了过多冗余规则的生成。当最小置信度大于0.6时差别减小。



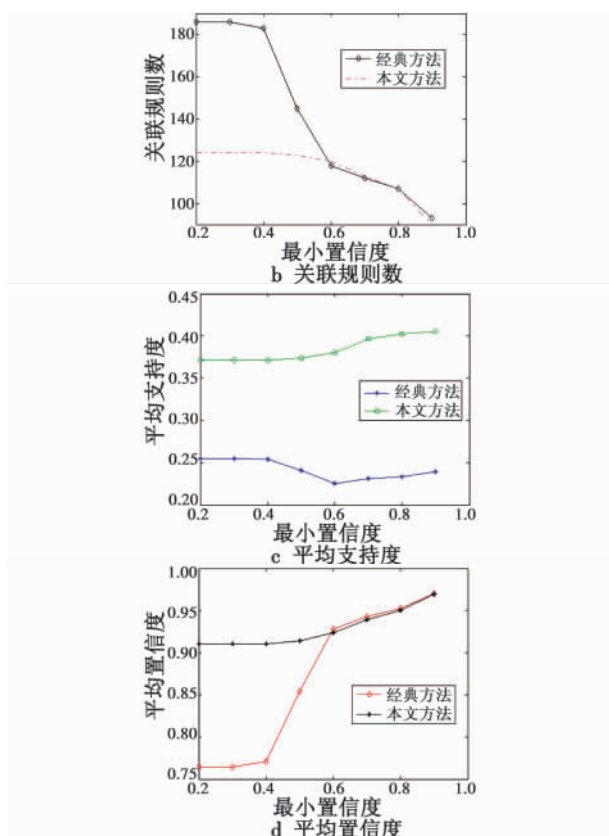


图 4 不同最小置信度下的品质度量

Fig.4 The quality measurement under different minimum confidence

4.2.2 性能度量

在性能方面主要比较了 $C_{\min \text{ conf}} = 0.6, S_{\min \text{ sup}} = [0.05 \ 0.3]$ 和 $S_{\min \text{ sup}} = 0.2, C_{\min \text{ conf}} = [0.2 \ 0.9]$ 两种情况下算法的执行时间,如图 5 所示。

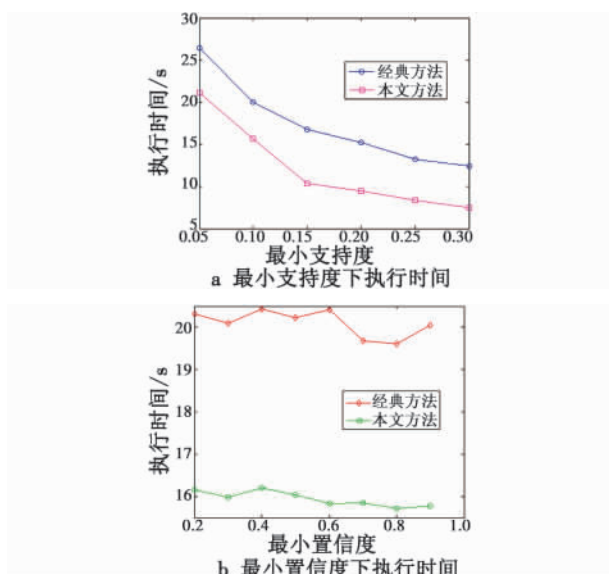


图 5 执行时间比较图

Fig.5 The execution time comparison

从图 5a 和图 5b 可知,在两种情况下,本文方法的

执行时间均小于经典方法,效率更高。

由图 3a 和图 5a 可知,由于频繁项集越大,候选项越多,生成的规则越多,执行时间也越长,所以执行时间随着频繁项集的增大而增加。

4.2.3 QAR 关联规则比较

表 2 和表 3 分别给出两种方法生成的部分规则,图 6 表示某些 QAR 数据属性值,图 6b 中用每分钟转速的百分率表示。从图 6 可以看出,这些属性值之间基本上遵循正比的关系,即 EGT 越大,FF 越大,N1 越大则 N2 也越大,从而可看出,本文生成的关联规则更符合数据属性间的潜在关系。

表 2 经典方法生成的关联规则

Table 2 The association rules generated by the classic method

if ALT 低,EGT 低 then CAS 低
if ALT 中等,EGT 中等 then CAS 高
if ALT 中等,EGT 中等,FF 中等 then CAS 高
if ALT 中等,EGT 中等,FF 中等,CAS 高 then N1 高

表 3 本文方法生成的关联规则

Table 3 The association rules generated by this method

if ALT 低,EGT 低 then CAS 低
if ALT 中等,EGT 中等 then CAS 高
if ALT 高,EGT 中等,FF 高 then CAS 高
if ALT 高,EGT 中等,FF 高,CAS 高 then N1 高

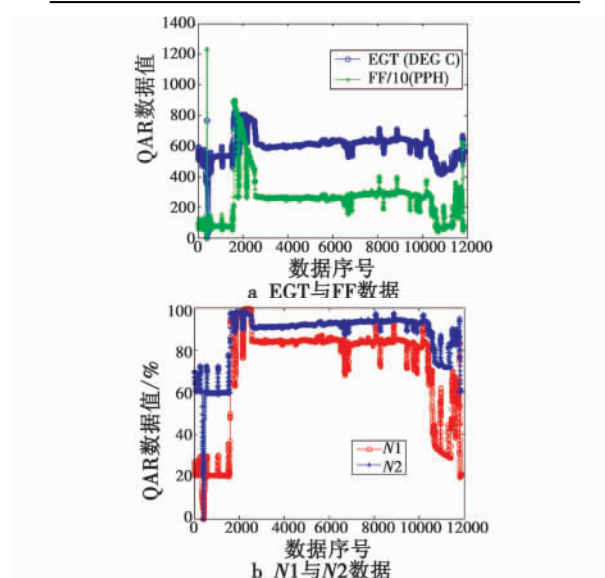


图 6 部分 QAR 属性值

Fig.6 Partial property value of QAR

因此,本文方法不但在质量品质和性能两方面优于经典方法,而且生成的关联规则也更能体现 QAR 数据项之间的潜在联系,反映飞机健康状况,更好地为故

障诊断与预防性维护提供理论支持。

5 结论

针对模糊划分需事先给定与 Apriori 效率低的问题,提出基于自动模糊划分和改进 Apriori 算法的关联规则生成方法,并将其应用到 QAR 数据中,进行了品质度和性能两方面的实验,实验结果表明,该方法不仅各方面性能优于经典方法,且生成的关联规则能更好地反映 QAR 数据的特性、属性间的潜在联系及飞机的飞行状态、健康状况,为故障诊断和预防性维护提供了理论支撑。

参考文献

- [1] 耿宏,揭俊.基于 QAR 数据的飞机巡航段燃油流量回归模型[J].航空发动机,2008,34(4):46-50.
- [2] 冯兴杰,李胜,邹秀霞.基于小波尺度系数的民航 QAR 数据约简及其性能分析[J].计算机工程与设计,2009,30(5):1255-1258.
- [3] 冯兴杰,冯小荣,王艳华.基于 KPCA 的 QAR 数据分析[J].计算机工程与应用,2009,45(14):207-209.
- [4] 邹秀霞,王红.基于数据仓库的 QAR 数据分析[J].计算机工程与设计,2008,29(10):2685-2688.
- [5] 卿立勇,黄圣国,林钰森.基于 QAR 数据的飞机系统故障预测与故障诊断支持系统研究[J].江苏航空,2006(2):11-12.
- [6] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules [C]//Proc. of the 20th VLDB Conference, 1994: 478-499.
- [7] FU A W C, WONG M H, SZE S C, et al. Finding fuzzy sets for the mining of fuzzy association rules for numerical attributes [C]//Proc. Int. Symp. on Intelligent Data Engineering and Learning (Ideal98), Hong Kong, 1998: 263-268.
- [8] 殷剑锋,徐建城,李伟强.改进 Apriori 算法的网格实现[J].计算机仿真,2010,27(2):145-148.
- [9] WU Jian, LI Xingming. An effective mining algorithm for weighted association rules in communication networks [J]. Journal of Computers, 2008, 3(10): 20-26.
- [10] 高原,倪世宏,王彦鸿,等.一种基于改进遗传量子算法的飞行状态规则提取方法[J].电光与控制,2011,18(1):28-31.
- [11] 胡傲,冯新喜,王冬旭,等.遗传模糊聚类算法在数据关联中的应用[J].电光与控制,2010,17(3):30-34.
- [12] LOTFI S, SADREDDINI M H. Mining fuzzy association rules using mutual information [C]//Proceeding of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, Hong Kong, 2009: 684-689.
- [13] 李成军,杨天奇.一种改进的加权关联规则挖掘算法[J].计算机工程,2010,36(7):55-57.
- [14] 李闯,杨胜,李仁发.一种最优分类关联规则算法[J].计算机工程与科学,2009,31(4):63-65.
- [15] KHAN M S, MUYEBA M, COENEN F. Weighted association rule mining from binary and fuzzy Data [C]//ICDM'08 Proceedings of the 8th Industrial Conference on Advances in Data Mining, Springer-Verlag Berlin Heidelberg, 2008: 200-212.

(上接第26页)

内,完全能满足捷联寻北对精度的要求;在达到3"的求角精度时只需22 kB的制表空间,如果采用直接查表法,要达到3"的计算精度将需要3.3 MB的存储空间,因此相对于直接查表法,分段线性近似求角法节省了153倍的存储空间;由于在整个过程中只需要存储两个小数组,再通过简单的比较和计算就能方便地得出载体的方位角,而查表法需要采用轮询遍历的方式才能求得载体的方位角,采用分段线性近似求角法就能极大地提高代码的执行效率。因此,这种在8区间细分的基础上再对正切函数进行分段线性处理的方法是

一种简单有效的近似求解反正切角度的方法。

参考文献

- [1] 张恒,冯旭升,薛东方,等.基于正切算法的轴角数字转换器设计[J].机械工程学院学报,2010,22(3):40-43.
- [2] 简小军,史耀耀,汪文虎,等.光栅信号软件细分技术及其误差分析[J].工具技术,2006,40:72-74.
- [3] 于恩祥.同步机角度——数字转换的一种新方法及其实现[J].吉林大学自然科学学报,2001(1):53-56.
- [4] 唐小琦,刘世峰,王平江,等.正切法莫尔条纹信号幅值分割细分的误差分析[J].测量与设备,2007(2):8-11.

欢迎订阅 欢迎刊登广告