文章编号: 1006-4710(2012)01-0106-05

关联规则分类的数据流挖掘方法在水电机组故障诊断的应用

苏立[1], 南海鹏[1], 余向阳[1], 吴罗长[1], 王瑾[2]

(1. 西安理工大学 水利水电学院, 陕西 西安 710048;
2. 中国水电顾问集团 贵阳勘测设计研究院,贵州 贵阳 550081)

摘要: 引进一种数据流关联规则分类法(AC-DS),并应用该方法对 UCI 机器学习库中标准数据集进行分类验证,验证结果表明该方法准确且有效。然后将该方法应用到水电机组故障的诊断分类中,证明该方法的分类精度随着测试样本的增加而增加。该方法对现场不同类型机组故障分类有一定意义。

关键词: 水电机组; 故障分类; 数据流; 关联分类

中图分类号: TM312 **文献标志码**: A

The Application of Data Stream Mining Method of Associative Rule Classification in Fault Diagnosis of Hydro-Turbine Generating Unit

SU Li[1], NAN Haipeng[1], YU Xiangyang[1], WU Luochang[1], WANG Jin[2]

(1. Faculty of Water Resources and Hydroelectric Engineering, Xi'an University of Technology, Xi'an 710048, China;
2. Hydrochina Guiyang Engineering Corporation, Guiyang 550081, China)

Abstract: This paper introduces a kind of data stream associative rule classification method, and this method is used to carry out the classification test of the standard datasets in the UCI machine learning repository. The test results indicate that this method is accurate and effective. And then, this method is used in the fault diagnosis classification of hydro-turbine generating unit, whereby proving that the classification accuracy of this method is improved with an increase in the testing samples. Accordingly, this method is of the certain significance for the in-situ fault classification of generating units of different types.

Key words: hydroelectric units; fault classification; data streams; associative classification

数据挖掘技术[1-2]是一门主要由统计学和人工智能学组成的交叉学科,它的主要任务是从海量的数据集中提取隐藏的、有用的信息。目前,数据挖掘技术已被广泛应用于电信、金融、网络安全、天气预报等领域[3]。随着软硬件技术的高速发展,经常有大量需要处理的数据以很快的速度产生,利用传统的数据挖掘方法已经不能满足目前数据处理的要求。一种新型的解决方案——基于关联规则分类的数据流挖掘方法应运而生,其最大特点是待处理的数据以一种动态、流式的形式出现,即数据流(data streams)[4],对数据流中的数据只能按顺序进行一次或有限次的访问[5]。

在水电机组运行过程中,振动信号包含了很多机组运行信息。据统计,水电机组约有 80% 的故障或事故都在振动信号中有所反映[6],而在机组在线监测过程中,机组振动信息数据量非常大,并且随着监测时间的延长而增长,信息数据成为大规模海量数据集,这就给信号的处理分析带来一定的难度。一般的方法是通过典型标准故障样本对分类器进行训练,但是个体机组之间则存在着差异,这会影响到分类的精度,而数据流分类正好符合这种连续数据流分类,在对数据进行测试的同时也进行对自身分类器的更新,因此对水电机组监测数据流进行挖掘分析是一种有效的方法。

收稿日期: 2011-11-15

基金项目: 国家青年科学基金资助项目(50809054)。

作者简介: 苏立(1982-),男,山西太原人,博士生,研究方向为水电机组监测与控制。E-mail: sunan971@163.com。
南海鹏(1963-),男,陕西乾县人,博导,研究方向为水电机组监测与控制。E-mail: hxnhp@163.com。

1 数据流挖掘概念

1.1 数据流

所谓数据流就是大量连续到达的、随时间不断变化的有序数据序列。设 $\boldsymbol{I}=\{i_1,i_2,\cdots,i_m\}$ 是一组集合，则对于任意 $\boldsymbol{X}\subseteq\boldsymbol{I}$，称 $\boldsymbol{X}$ 为一个项集（如果 $|\boldsymbol{X}|=k$，则称 $\boldsymbol{X}$ 为 k 项集）[7]。事务数据库 $\boldsymbol{D}_S=\{\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_N\}$ 是一个序列阵，用矩阵标识符 $\boldsymbol{B}_i$ 来标识此处的每一个矩阵，$\boldsymbol{B}_N$ 是最后一个矩阵。任何一个矩阵 $\boldsymbol{B}_i$ 是由一组事务组成的，即为 $\boldsymbol{B}_i=[T_1,T_2,\cdots,T_k]$，$k>0$。因此，数据流的流长定义为：

$$CL=|\boldsymbol{B}_1|+|\boldsymbol{B}_2|+\cdots+|\boldsymbol{B}_N|$$

一个项目集 $\boldsymbol{X}$ 的频率项定义为 freq($\boldsymbol{X}$)，表示事务 $\boldsymbol{B}$ 中项目集 $\boldsymbol{X}$ 的个数。对 $\boldsymbol{X}$ 的支持度定义为 freq($\boldsymbol{X}$)/N，N 是接收到的事务总数。如果 $sup(\boldsymbol{X})\geqslant\min\boldsymbol{Support}$，$\boldsymbol{X}$ 即频繁项目集（Frequent Itemset）。此处最小支持度（$0\leqslant\min\boldsymbol{Support}\leqslant1$）是用户定义的最小支持阈值。

1.2 关联分类法

基于关联规则的分类是针对数据集挖掘频繁模式而建立的分类器，以实现对未知样本分类的方法[8]。设 $\boldsymbol{D}$ 是数据库，$\boldsymbol{I}$ 是 $\boldsymbol{D}$ 中一组项目集，$\boldsymbol{C}$ 是一组类标记。记 $\boldsymbol{D}_i\in\boldsymbol{D}$，$\boldsymbol{X}\subseteq\boldsymbol{I}$ 的一组关联规则（class association rule）是表示从 $\boldsymbol{X}$ 到 $\boldsymbol{c}$ 的关联规则，此处 $\boldsymbol{X}\subseteq\boldsymbol{I}$，$\boldsymbol{c}\in\boldsymbol{C}$。

Bing Liu[9] 首次提出了关联分类（AC）。使用关联分类法进行分类与传统的频繁项目集挖掘的区别是，可以同时执行多种频繁项目集挖掘程序。关联分类（AC）数据挖掘构架常由两步组成：

① 生成关联分类规则，其中 $\boldsymbol{i}_{\text{set}}\geqslant\boldsymbol{c}$，此处 $\boldsymbol{i}_{\text{set}}$ 是一个项目集，$\boldsymbol{c}$ 是一个类。

② 根据产生的关联分类规则创建分类器。一般来说选择关联规则的一个子集作为一个分类器。关联分类法是根据置信度来选择规则的。

大多数的分类算法都是在寻求频繁项目集，其中有一些是利用决策树来分类数据流。本文运用一种挖掘类关联规则的方法，通过挖掘数据的关联规则，然后制作一个数据流的分类器。

2 数据流的关联分类

2.1 问题的定义

数据流是快速连续产生的大量无约束数据元素。由于数据流的实时性特点，单程的算法不得不牺牲在其允许误差范围内分析结果的正确性。但是项目集 $\boldsymbol{X}$ 真实的可信度由 $T_{\sup}(\boldsymbol{X})$ 来表示，它是关于对项目集子集的数量预期，项目集 $\boldsymbol{X}$ 的预计支持度由 $E_{\sup}(\boldsymbol{X})$ 表示，是由一次扫描方法得到的存储了总体的数据结构 $\boldsymbol{X}$ 预计的可信度，此处的 $E_{\sup}(\boldsymbol{X})\leqslant T_{\sup}(\boldsymbol{X})$。如果 $T_{\sup}(\boldsymbol{X})\geqslant$ 最小支持度（min*Support*），项目集 $\boldsymbol{X}$ 就称作频繁项目集。目标是在边界窗口模式下尽可能少地使用主存储器来发展单程算法分类数据流。

2.2 数据流的关联分类算法

数据流的分类任务可以描述为：从数据流中发现所有频繁且精确的规则集合，并从该规则集合中选取合适的规则子集构成分类器[10]，由图 1 可以知道整个算法，算法允许用户指定两个参数：最小支持度阈和窗口尺寸 $S_{\text{window}}=|\boldsymbol{B}_i|$（用 N 表示流的当前长度）。每当记录一条新的数据组时，算法根据选出关联规则能够预报它的类的标识。每一个规则都有一个置信度 $E_{\sup}(\boldsymbol{X})$，它的值大于最小置信度。首先通过给定数据块 $\boldsymbol{B}_i$ 进行训练，计算以确定事件的频繁 1-项集，只要这一组 1-项目集不是空集，则接着执行下一个，如此迭代来寻找频繁（$m+1$）-项目集。当得到所有的频繁项目集，将计算规则的置信度和排列顺序记录在内存中。当训练好的分类器对数据流进行分类时，如果有新的数据 $\boldsymbol{B}_i$，对数据进行分类的同时也会对分类算法中记录的规则进行修改和存储，同时也删除低置信度的规则，这样就形成了对数据流时变分类，即在对样本进行测试的过程中自动地更新分类器的分类规则，从而实现实时更新。

```
输入: D_S---一个数据流，每一个记录都有 N 个项。
      S_window---窗口大小, S_window=|B_i|。
      minSupport--- 最小支持度阈值。
输出: M---具有高于 50%准确率且可信度高于最小置信度的多种关联规则所组成的分类器。
方法:
  原始的规则存储  M=Φ
  Do
  读取数据块 B_i={T_1, T_2, ..., T_k}
    m=0; A_m=Φ            //清除候选中的项目集 A
    A_{m+1}=Gen(B_i, A_m)   // 生成 n 候选频繁项目集
                            // 项目集_1,项目集_2,...,项目集_n,
                            //A 中的每一个项目集 i 有 1 个项目.
      While A≠Φ
        For  i=1 to n
          S=Supp(Itemset_i)      //计算 Itemset_i 的置信度
            If  S>= minSupport  then
              M←M+ Itemset_i      //将项目集存储
            End if
        End for
        m=m+1
        A_{m+1}=Gen(B_i, A_m)     //生成 A(m+1)
      Endwhile
      M=Rank(M)          //按照置信度进行排序
      M=Decay(M)         //替换存储的分类规律
  While
```

图 1 关联分类的算法

Fig. 1 Associative classification algorithm

2.3 数值试验

通过文中的方法对 UCI (UCI Repository of Machine Learning Databases)机器学习库中标准数据集进行分类验证,在 2.93 GHz Pentium PC 处理器、2G 内存、Microsoft Windows XP 系统上做实验来检测方法的性能。为了验证本文算法可信与可靠,从 UCI 机器学习库中选择数据集 nursery,在对连续特性的离散化的过程中应用了熵法,是对于数据库中的每个连续属性,先将它的取值范围划分为若干区间,每个区间对应一个不重复值;然后选择两个毗邻区间进行合并,使合并前后的熵差最小[11]。

由图 2 可以看出,当初始化窗口尺寸设置为 500、1000、2000 和 4000,最小支持度阈值设置为 10% 时,可以得到类似预计分类结果,数据随着测试样本的增加精度也逐渐增加,最后测试精度基本相同;研究不同窗口尺寸对花费时间的影响时,发现在数据集为同一数据集时,窗口较小时花费时间较长,这是由于当窗口较小时,出现一些非频繁规则较多,因此耗费的时间较长。当窗口变大时,非频繁规则减少,花费时间也相应减少。但是窗口比例增大,运行时间并非成比例地减小,因此选择合适的窗口满足流挖掘过程时间即可。

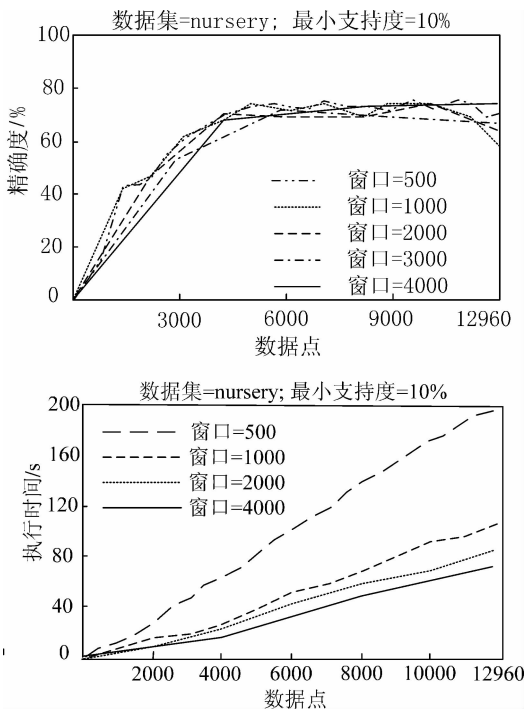

图 2 不同窗口分类算法的准确性和执行时间

Fig. 2 The accuracy and time of different block size

由图 3 可看出,随着支持度阈值的减小,能够得到准确率类似的分类结果。但是同一时刻支持度越小的分类器得到的规则越多,所消耗的时间也就越多。从以上实验可看出,关联分类算法在分类精度、运行时间、对非频繁规则更新速度等方面都有不错的表现,在应用过程中,选择合适的窗口大小和支持度,满足算法消耗时间小于数据流更新数据时间即可。对于精确度,当分类器自学习到一定程度时间,在数据流没有发生概念漂移的情况下精确度基本相同。

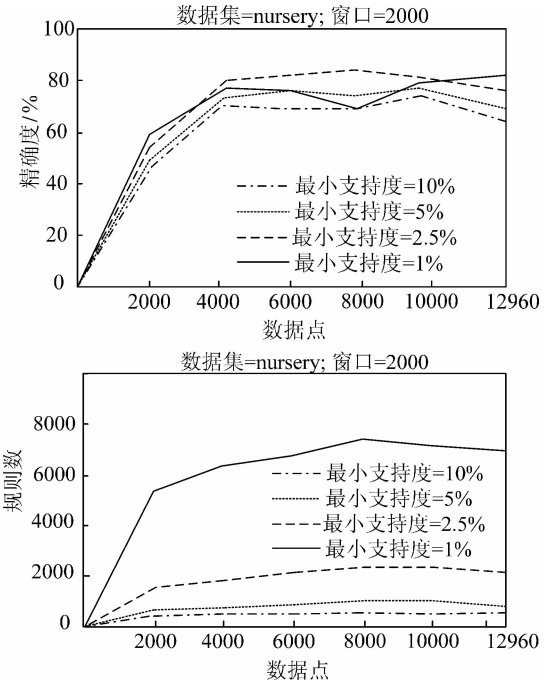

图 3 不同最小支持度分类算法的准确性和规则数

Fig. 3 The accuracy and rules of different support thresholds

3 水电机组振动挖掘实例验证

引起水电机组振动的原因很多,大致可归纳为机械、水力和电气 3 个方面。在对水电机组测试的过程中,由于测点、振动强度和幅值等各个因素不同,测试出来的数据组也有所不同。但是根据经验,水力机械振动故障的征兆大多由频域特征来表达,因此可利用频率分量作为数据流进行分类。通过对监测的振动信号进行 FFT 变换后得到其频谱特性,通常选择(0.15 ~ 0.5 (平均量)) x、$1x$、$2x$、$3x$、50 Hz、100 Hz、导叶数 × 转频频率分量进行比较(其中 x 为转频),对 FFT 变换后的各个频率幅值进行归一化处理[12]。

目前,已有不少文献对水电机组的振动故障机理进行分析及试验研究,本文根据文献[13] ~ [15]收集到的有关机组振动的典型故障征兆,做出总结如表 1 所示,通过表 1 对分类器进行训练。利用典型故障数据训练好的分类器对监测机组振动的数据流进行分类,首先将机组正常运行的数据输入分类

器，人为地将正常数据替换为故障数据输入到分类器中，分类器根据训练好的分类规则进行分类，得出故障类型，如表 2 所示。

表 1 学习样本表

Tab. 1 Learning samples for fault diagnosis

	故障样本	样本输入故障类型							故障类型
		(0.15~0.5)x	1x	2x	3x	50 Hz	100 Hz	导叶数×转频	
训练样本	涡带偏心	0.88	0.22	0.02	0.04	0.06	0.05	0.05	Class1
	涡带偏心	0.90	0.20	0.05	0.02	0.02	0.01	0.03	Class1
	不对中	0.02	0.52	0.40	0.32	0.10	0.08	0.05	Class2
	不对中	0.01	0.41	0.43	0.34	0.15	0.09	0.01	Class2
	不平衡	0.04	0.98	0.10	0.10	0.02	0.01	0.08	Class3
	不平衡	0.02	0.90	0.08	0.05	0.03	0.04	0.04	Class3
	定子铁芯松动	0.02	0.13	0.15	0.08	0.80	0.75	0.05	Class4
	定子铁芯松动	0.05	0.18	0.07	0.11	0.76	0.77	0.04	Class4
	三相负荷不平衡	0.92	0.05	0.89	0.12	0.05	0.08	0.01	Class5
	三相负荷不平衡	0.88	0.09	0.77	0.15	0.04	0.11	0.20	Class5
	转轮导叶开口不均	0.02	0.10	0.13	0.14	0.05	0.04	0.93	Class6
	转轮导叶开口不均	0.03	0.14	0.22	0.18	0.07	0.06	0.98	Class6

表 2 测试结果表

Tab. 2 Testing samples for fault diagnosis

故障类型		输入故障类型							结果输出类型
		(0.15~0.5)x	1x	2x	3x	50 Hz	100 Hz	导叶数×转频	
测试	不平衡	0.02	0.91	0.08	0.01	0.02	0.05	0.08	Class3
	不对中	0.01	0.48	0.48	0.36	0.10	0.08	0.10	Class2

虽然水轮机的故障样本与典型故障特征基本一致，但针对于每个机组，也会存在一些差异。普通静态分类器在经过训练后其分类器规则是固定不变的，本文的流挖掘的方法在对测试样本进行测试的过程中也可以通过输入故障结果自身进行学习更新，这样分类器精度也会根据测试样本的增加而增加，如表 3 所示。这样对于现场实际测试，准确率随着测试样本的增加，对故障的分类精度也会增加，对现场不同类型机组故障分类有一定意义。

表 3 不同方法结果比较

Tab. 3 Results comparison of different methods

分类方法	第一次测试准确率	第二次测试准确率	第三次测试准确率
贝叶斯	78.3%	78.1%	78.4%
关联规则流挖掘	77.9%	80.2%	81.5%

4 结 语

数据流上的分类挖掘是不同于基于数据库（即静态数据集）分类挖掘的，其待处理的数据以动态、流式的形式不断出现。本文通过关联规则的流挖掘方法对 UCI 标准数据集进行分类，分析和实验证明了该方法是准确、有效，并将该分类方法运用到水电机组的振动信号的故障分类中，充分发挥流挖掘在连续数据中快速分类同时进行自更新的特点，具有一定的实用价值。

参考文献：

[1] Han J, Kamber M. Data mining: concept and techniques [M]. 2nd Edition. San Fransisco: CA. Higher Education Press, 2001:1-7.

[2] Tan Pangning, Sreinbach M, Kumar V. 数据挖掘导论 [M]. 北京:北京大学出版社, 2006.

[3] 王涛，李周军，颜跃进，等. 数据流挖掘分类技术综述 [J]. 计算机研究与发展, 2007,44(11):1809-1815.
Wang Tao, Li Zhoujun, Yan Yuejin, et al. A survey of classification of data streams[J]. Journal of Computer Research and Development, 2007,44(11):1809-1815.

[4] Babcock B, Babu S, Datar M, et al. Models and issues in data streams[C]//proc of ACM Sump on Principles of Database Systems. New York: ACM Press,2002:1-16.

[5] 金澈清，钱卫宁，周傲英. 流数据分析与管理综述[J]. 软件学报,2004,15(8):1172-1179.
Jin Cheqing, Qian Weining, Zhou Aoying. Analysis and management of streaming data: a survey [J]. Journal of

Software, 2004,15(8):1172-1179.

[6] 赵道利,马薇,梁武科,等. 水电机组振动故障的信息融合诊断与仿真研究[J]. 中国电机工程学报,2005,25(20):137-142.
Zhao Daoli, Ma Wei, Liang Wuke, et al. On data fusion fault diagnosis and simulation of hydroelectric unit vibration [J]. Proceedings of the CSEE, 2005, 25(20):137-142.

[7] 孟彩霞. 面向数据流的频繁模式挖掘研究[J]. 计算机应用研究,2009,26(11):4054-4056.
Meng Caixia. Research on mining frequent patterns in data streams[J]. Application Research of Computers, 2009,26(11):4054-4056.

[8] 李宏,李博,吴敏,等. 一种基于关联规则的多类标分类算法[J]. 控制与决策,2009,24(4):574-578.
Li Hong, Li Bo, Wu Min, et al. Multi-label classification algorithm based on association rules[J]. Control and Decision, 2009,24(4):574-578.

[9] Liu Bing, Wynne Hsu, Ma Yiming. Integrating classification and association rule mining: proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York,1998.

[10] 赵传申,何顺刚,杨吉宏. 基于多分类-关联规则的数据流分类算法[J]. 计算机工程,2010,36(9):38-40.
Zhao Chuanshen, He Shungang, Yang Jihong. Data stream classification algorithm based on multiple class-association rules[J]. Computer Engineering,2010, 36(9): 38-40.

[11] 贺跃,郑建军,朱蕾. 一种基于熵的连续属性离散化算法[J]. 计算机应用,2005,25(3):637-638.
He Yue, Zheng Jianjun, Zhu Lei. An entropy-based algorithm for discretization of continuous variables[J]. Computer Applications, 2005,25(3):637-638.

[12] 白亮,王瀚,李辉,等. 基于时间序列相似性挖掘的水电机组振动故障诊断研究[J]. 水力发电学报, 2010, 29(6):229-236.
Bai Liang, Wang Han, Li Hui, et al. Vibration fault diagnosis based on time-series similarity mining for hydropower units[J]. Journal of Hydroelectric Engineering, 2010,29(6):229-236.

[13] 张利平,孙美凤,王铁生. 新型的RBF神经网络在水轮发电机组故障诊断中的应用[J]. 水力发电学报, 2009,28(6):219-223.
Zhang Liping, Sun Meifeng, Wang Tiesheng. Application of a novel RBF algorithm to fault diagnosis of hydro-turbine generating unit[J]. Journal of Hydroelectric Engineering, 2009,28(6):219-223.

[14] 彭文季,罗兴锜,赵道利. 基于频谱法与径向基函数网络的水电机组振动故障诊断[J]. 中国电机工程学报, 2006,26(9):155-158.
Peng Wenji, Luo Xingqi, Zhao Daoli. Vibrant fault diagnosis of hydro-turbine generating unit base on spectrum analysis and RBF network method[J]. Proceedings of the CSEE,2006,26(9):155-158.

[15] 贾嵘,洪刚,武桦,等. 基于IPSO优化LSSVM的水轮发电机组振动故障诊断[J]. 水利学报,2011,42(3):373-377.
Jia Rong, Hong Gang, Wu Hua, et al. Vibration fault diagnosis of hydroelectric generating unit by Least Squares Support Vector Machine based on Improved Particle Swarm Optimization[J]. Journal of Hydraulic Engineering,2011,42(3):373-377.

(责任编辑 王卫勋)