时序化生产预警有效影响因子的获取方法研究

李春生,邸京华,李少龙,张可佳,王 梅
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:在生产预警有效影响因子的筛选过程中,为了达到降低维度,增强影响因子集的有效性,从而提高生产异常预警准确率的目的,选取和分析所有原始项目,应用模糊综合评价法量化模糊限制语,采用 TRIMMEAN 内均法排除极端评估值。运用特征选择技术发现敏感特征因子,借鉴混合智能方法定义影响因子集的逻辑表达结构,基于粒度分析处理时序化数据,同时利用激剧判定函数摒弃无效元素完成对数据的降维以及筛选,得到高精细化的有效影响因子集。以此达到辅助深度挖掘数据内部潜在规律,解决信息杂乱等现象,运用于生产异常分析,提高预警准确率的目的。最后针对大庆油田某采油厂生产历史数据,完成时序化生产预警有效影响因子的获取。

关键词:生产异常预警;模糊综合评价法;特征选择;时序化数据;激剧函数
中图分类号:TP312 **文献标识码**:A **文章编号**:1673-629X(2016)07-0122-05
doi:10.3969/j.issn.1673-629X.2016.07.026

Research on Acquisition Method of Effective Impact Factors in Production Early Warning by Time Series

LI Chun-sheng, DI Jing-hua, LI Shao-long, ZHANG Ke-jia, WANG Mei
(College of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: In the screening process of effective impact factors for early warning, in order to reduce dimension, enhance the effectiveness of the impact factor set and improve accuracy in early warning of abnormal production, all original items are selected and analyzed, and fuzzy constraints is quantified based on fuzzy comprehensive evaluation method, using TRIMMEAN to eliminate extreme values. Then, sensitive feature factors are determined by using feature selection techniques. At the same time, the logical expression structure of the influence factor set is defined via the hybrid intelligent method, and the time sequence data is manipulated based on granularity analysis. Next, it finishes dimensionality reduction and selection of data through the dramatic function to achieve the effective impact factors of high precision. To reach the purpose that excavates potential law in the data deeply, and solves the phenomenon of information clutter, using the method in the production of abnormal analysis to improve accuracy in early warning. Finally, in combination with the history data of an oil production plant in Daqing Oilfield, the effective impact factors acquisition of the production early warning by time series is achieved.

Key words: early warning of abnormal production; fuzzy comprehensive evaluation method; feature selection; time series data; dramatic function

1 概 述

经过数十年的研究发展,针对生产异常的预警手段已经在大规模生产领域中得到广泛应用,并发挥着极大的作用。在数字化生产普及初期,各类数据尚不完善,预警过程完全依据专业技术人员感官评估和预测,对预警信息仅作验证测试,实为滞后预警。随着传感器技术及各种监测、测试手段的逐步普及,以数据处理、数据建模方法为依托,结合人工经验实现预警模式的探测式与监测式异常预警系统开始得到推广应用。例如 M600 系统、TSE/TEM 系统(用于发电预警监测)[1]。这类系统提供可以信任的推理依据,适用于生产模式变化不强,规律性明显,监控范围广泛的工业生产领域,但实际推理过程需要借助人工辅助分析,智能化推理支持度不高。随着数字化生产的进步以及与日俱增的数据量[2],挖掘数据内部存在规律成为智能化预警的关键。时序化数据处理方法的提出降低了大

收稿日期:2015-09-08 **修回日期**:2015-12-11 **网络出版时间**:2016-05-25
基金项目:黑龙江省科学基金项目(F2015020);东北石油大学校培育基金项目(XN2014102)
作者简介:李春生(1960-),男,博士,教授,博士生导师,研究方向为人工智能及其应用、数据挖掘与智能系统;邸京华(1990-),女,硕士研究生,研究方向为数据库与数据挖掘技术。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160525.1706.028.html

数据量分析过程中的耦合度,提高了数据处理的精细化,对于数据挖掘的有效性意义重大[3]。

探测监测手段以及数据处理方法对于生产异常预警领域至关重要。虽然通过工作人员和专家的丰富经验积累可以很好地推理业务领域内生产异常的影响因素,但仍然存在以下缺陷:

(1)多数异常间相互级联度较高,易并发,针对生产异常预警形成的实践经验和理论体系的通用性不强,经验和知识的松散度和针对性较高,不同专家对于异常的描述以及阈值的定义不同,存在无法融合的情况,针对生产异常情况缺少统一标准的影响因子集。

(2)异常预警所涉及的影响因子繁多,且样本具有不确定性,描述异常样本的特征维数高。

(3)针对生产早期微弱异常以及数据缺失状况,非敏感影响因子的隐蔽性较强[4],异常表征不明显,为推理可持续性采用预测数据填补,使得影响因子集存在不准确性。

针对上述问题,研究时序化生产预警有效影响因子的获取方法。下文将首先针对预警目标获取相应粗糙原始数据,应用模糊综合评价法量化专家对影响因子的模糊语义描述,采用 TRIMMEAN 内均法排除极端数据点。运用特征选择技术从特征相关性和冗余性出发,通过 CF-ISF 权重计算方法发现敏感特征因子,剔除冗余数据。其次,定义影响因子集的逻辑表达结构,通过提出 ND 模型建立影响因子与粗糙原始项目的映射关系。引入时间序列,结合业务数据的特点与同一模式多重粒度的思想,选取最佳数据粒度进行数据处理。最后,利用激剧判定函数对数据进行降维以及有效化筛选,得到高精细化的有效影响因子集,完成对时序化生产预警有效影响因子的获取。

流程如图 1 所示。

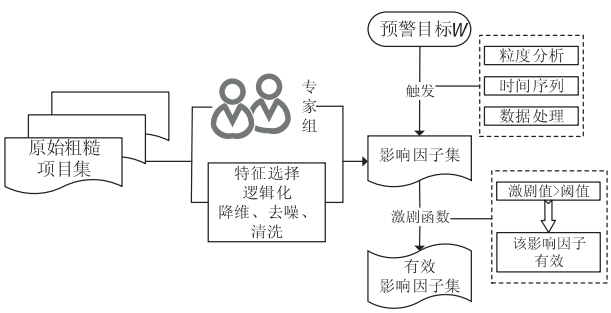

图 1　有效影响因子集的获取

2　原始影响因子集的获取

原始项目组是数据有序化构成的信息项集合,是影响因子集的构建基础。对原始数据的合理化处理是实现有效数据项获取的前提。因此,文中针对某异常预警目标进行粗糙原始项目选定,定义影响因子的表达结构,建立影响因子与粗糙原始项目的映射关系,完成原始影响因子集的组织和获取,为生产原始数据抽象化打下重要基础。

2.1　粗糙原始项目的选定

为降低原始项目分析的难度,在全域内剔除完全无关的原始项目。针对某异常预警目标,选定粗糙原始项目,应用模糊综合评价法,通过对领域内专家组语义倾向性评价确定隶属度以及权重,并对自然语言的语义限制词(模糊限制语)进行定量化描述[5],如表 1 所示。

表 1　模糊限制语的数学和图例表示

模糊限制语	数学表示	图例表示
A little	$[\mu_A(x)]^{1.3}$	
lightly	$[\mu_A(x)]^{1.7}$	
Very	$[\mu_A(x)]^{2}$	
Extremely	$[\mu_A(x)]^{3}$	
Very very	$[\mu_A(x)]^{4}$	
More or less	$\sqrt{\mu_A(x)}$	
Somewhat	$\sqrt{\mu_A(x)}$	
Indeed	$2[\mu_A(x)]^2$ $0 \leqslant \mu_A \leqslant 0.5$	

通过表 1 内图例中虚线与实线间的面积可以比较出不同模糊限制语的隶属度大小,面积越大则隶属度越高。

设定领域内某异常预警目标 W,于是 $\forall$ 与 W 相关的所有原始项目集合 U 为:

$$U = \{\mathrm{op}_1, \mathrm{op}_2, \cdots, \mathrm{op}_n\}$$

其中,op 包括对原始项目的基本属性描述和领域内专家组对于该原始项目有效性的语义倾向描述;n 表示原始项目数量。

利用模糊综合评价法基于模糊数学的隶属度理论通过对语义限制词的定量化映射可得专家组对于集合 U 的量化结果:

$$\boldsymbol{Q} = \begin{vmatrix} v_{11} & \xi_1 v_{12} & v_{13} & \cdots & v_{1(r-1)} & \xi_k v_{1r} & \cdots & v_{1n} \\ v_{21} & \xi_1 v_{22} & v_{23} & \cdots & v_{2(r-1)} & \xi_k v_{2r} & \cdots & v_{2n} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ v_{m1} & \xi_1 v_{m2} & v_m & \cdots & v_{m(r-1)} & \xi_k v_{mr} & \cdots & v_{mn} \end{vmatrix}$$

其中,ξ 表示项目的权重系数; m 表示专家组专家数量。

则某 op_n 的量化结果为:

$$Q_r = \{v_1, v_2, \cdots, v_m \mid m > 2\}$$

为避免极端评估降低评价准确性,采用 TRIMMEAN 内均法进行 op 评估,取剔除数据点的比例为 20%,则评估结果 opv - 可表示为:

$$\bar{opv} = \text{TRIMMEAN}(Q_r, 0.2)$$

$$\bar{opv} = \text{avg}(Q_r - \text{cut}(\text{Max}(Q_r), \text{Min}(Q_r)))$$

于是, W 的原始项目集合 U 的评价结果可进行如下描述:

$$U_c = \{\bar{opv}_1, \bar{opv}_2, \cdots, \bar{opv}_n\}$$

利用特征选择技术发现敏感特征影响因子,剔除评估结果完全无关项,获得最能代表问题空间的特征子集。从特征相关性和冗余性定义出发[6],采用 CF-ISF(Characteristic Frequency, Inverse Sample Frequency)权重计算方法,则某影响因子的权重为:

$$\xi_k = opf_k * \log(N/n_k + 0.01)$$

其中, opf_k 为特征项 op_k 在样本集中出现的次数; N 为全部训练集的样本数; n_k 为训练样本中出现特征项 op_k 的次数。

考虑到样本长度对权值的影响,对 ξ_k 做归一化处理,将各项的权值规范到[0,1]之间:

$$\xi_k = \frac{opf_k * \log(N/n_k + 0.01)}{\sqrt{\sum_{k=1}^{N} [opf_k * \log(N/n_k + 0.01)]^2}} \quad (1)$$

于是,在剔除评估结果完全无关项以后,得到异常预警目标 W 的粗糙原始项目:

$$U_{\text{rough}} = \{op_{r1}, op_{r2}, \cdots, op_{rk} \mid k < n\}$$

2.2 影响因子的逻辑转化

定义影响因子的逻辑表达结构,是影响因子进行抽象描述和符号转换,以及后续数据处理的重要步骤,也是特征模式提取、数据挖掘和应用推理的必备过程。结合混合智能方法的提出思想,影响因子的表达不仅需要包括支持知识推理、迭代学习的相关主要属性,同时也要考虑保证在 Hebbian 学习、置信危机消解过程中的副属性描述[7]。定义 $C-E$ 结构,具体描述如下:

定义影响因子可由表示直接支持其有效化计算及特征提取的属性和方法的集合 C(核心集)与具有其他表征意义或具有辅助作用的属性和方法的集合 E(扩展集)表示。其一般形式: $R = \{C, E \mid C \neq \varnothing\}$。其中核心集 C 可表示为:

$$C = \{\hat{op}, \mu_c, \varphi_c \mid \text{count}(\hat{op}) \geqslant 1\}$$

其中, $\hat{op}$ 表示构成该影响因子的项目集合。通过对生产数据分析可知,通常单一项目可以完整地表述某影响因子,即影响因子与项目存在一一映射关系。为提高 $C-E$ 结构的可扩展性,在此提出可包含多种映射关系的项目集合[8]。μ_c 表示针对项目数据的处理方法,如:时序处理法、频分法、傅里叶变换等。φ_c 表示影响因子有效化处理方法。

扩展集 E 可表示为: $E = \{D_{op}\}$。其中, D_{op} 表示该影响因子的项目集合所包含的数据信息。由业务数据特点可知,原始数据多数以不同粒度的时间序列进行存储,这一特点是选择原始项目数据处理方法 μ_c 的重要依据。

由此完成对 $C-E$ 结构的设计和描述。$C-E$ 结构不仅实现将离散的、模糊的信息抽象化和结构化,同时,结构松散的设计思路满足了混合智能方法的需要。

2.3 建立影响因子与粗糙原始项目的映射关系

在完成获取粗糙原始项目集,并定义 $C-E$ 结构描述影响因子后,需要获取与生产异常相关的全部原始影响因子。于是,依据业务要求和生产异常预警理论,为实现建立自然语言描述的影响因子与数据体内数据实体的映射关系,引入 ND 模型。具体定义如下:

定义 1:包含影响因子的自然语言形式 op,直接描述 op 的数据实体 du 及映射关系函数 F 的闭包结构成为 ND 模型。其一般表示形式为:

$$\text{ND} = \{\text{op}, \text{du}, F \mid \text{du} \neq \varnothing, \text{op} \in U_c\}$$

其中, du 为数据实体,实例化后为数据体内的数据单项; U_c 为由专家组提供的原始影响因子集; F 为映射关系函数,在 op 可直述时, F 可为空,当 op 不可直述时, du 由 F 进行计算获得。

ND 模型建立了自然语言与逻辑语言间的映射关系,并将因子间相互独立,可以清晰地描述其抽象结构,提高 U_c 集的松散度,易于分析和计算。

以 ND 模型进行 U_c 集的逻辑转化,得到原始闭包集 FU_c。其一般表述形式为:

$$FU_c = \{FC_1, FC_2, \cdots, FC_n \mid n = \text{len}(U_c)\}$$

其中, FU_c 集维度与 U_c 集维度相同,并存在一一对应关系。

受到专家不确定性经验及定性化知识的影响, FU_c 集往往包含真实集 R_s,即 $R_s \subset FU_c$。为了进一步提高 FU_c 集的有效性,提出一种基于粒度分析的数据处理和清洗方法,去除 FU_c 集内无效元素,降低模式维度,防止维灾。

3 基于粒度分析的数据处理

数据粒度是数据仓库中数据的细化和综合程度。一般情况下,根据数据粒度划分标准,可以将数据仓库中的数据划分为:详细数据、轻度总结、高度总结。数据信息细化程度越高,粒度越小;细化程度越低,粒度

越大。粒度的选取原则是使其处于一个合适的级别,既不能太高也不能太低。低的粒度级别能提供详尽的数据信息,但要占用较多的存储空间和需要较长的查询时间。高的粒度级别能快速方便地进行查询,但不能提供过细的数据信息。

数据粒度的确定实质上是业务决策分析、硬件、软件和数据仓库使用方法的综合考虑。从生产异常动态分析需求的角度看,希望数据能以最原始的、细节化的状态保存,使得分析的结论最可靠[9]。但是,过低的粒度、过大的数据规模,在分析过程中给系统的 CPU 和 I/O 通道增加过大的负担,从而降低了系统效率。同时根据业务特点可知,研究异常事件周期时间内影响因子数据变化规律是发现异常特征的最优方法。由于影响因子存在连续性、周期性和时序性等特点,并且影响因子的时间粒度受关注度影响,因此,结合业务数据的特点[10],借鉴同一模式多重粒度的思想[11],通过以下方式确定合理的粒度值。

引入时间序列,使用低粒度数据保存近期的生产数据和汇总数据,对时间较久远的生产数据只保留粒度较大的汇总数据。这样既可以对生产异常近况进行细节分析,又可以利用汇总数据对生产异常规律进行分析。数据处理具体算法如下:

Start:生产异常预警目标 W 触发。

Step1:给定时间序列将数据实体 FC_n 划分为 m 段:$T=\{t_1,t_2,\cdots,t_m\}$ 。

Step2:定义 $t_m=\{t_{m1},t_{m2},\cdots,t_{ms}\}$ 内数据集合为 $d_m=\{d_{m1},d_{m2},\cdots,d_{ms}\}$ 。

Step3:若 d_m 原始数据长度 $s>0$,计算 d_m 原始数据均值。

$$\bar{du}_m=\frac{1}{s}\sum_{i=0}^{s}d_{mi},s=DT_{length}(d_m) \tag{2}$$

Step4:将 d_m 原始数据处理为局部距离数据。

$$dk_m=\{(d_{m1}-\bar{du}_m),(d_{m2}-\bar{du}_m),\cdots,(d_{ms}-\bar{du}_m)\}$$

Step5:取局部距离的标准差,得到数据集 d_m 的离散程度。

$$\sigma(d_m)=\frac{1}{s}\left[\sum_{i=0}^{s}dk_{mi}^2\right]^{1/2} \tag{3}$$

Step6:FC_n 的数据处理结果为:

$D=\{T,\sigma\},T=\{t_1,t_2,\cdots,t_m\},\sigma=\{\sigma_1,\sigma_2,\cdots,\sigma_m\}$

End

4　影响因子的有效化

FC_n 数据处理结果 $D=\{T,\sigma\}$ 结合 Carlson 定理(柯西定理针对 $m\times n$ 矩阵的一般推广形式)及切比雪夫变形[12]得到激剧判定函数:

$$F(\sigma)=\frac{1}{m-1}\sum_{1}^{m}\left(\frac{1}{s^2}\left(\sum_{i=0}^{s}\left(d_{mi}-\frac{1}{s}\sum_{i=0}^{s}d_{mi}\right)\right)\left(\sum_{i=0}^{s}\left(d_{(m-1)i}-\frac{1}{s}\sum_{i=0}^{s}d_{(m-1)i}\right)\right)\right)^{1/2}$$

其中,d_{ms} 为分段内原始数据;s 为分段内数据长度;m 为分段量。

将函数整理得:

$$F(\sigma)=\frac{1}{m-1}\sum_{1}^{m}\left(\prod_{m-1}^{m}\sigma\right)^{1/2} \tag{4}$$

借鉴特征选择方法的思想给出全局阈值系数 ξ 作为有效权重[13],于是得到阈值函数:

$$\mu_r(\sigma)=\xi\mu(\sigma)$$

取 $\mu(\sigma)=\max(\sigma)-\min(\sigma)$,根据激剧判定函数 $F(\sigma)$,阈值函数 $\mu_r(\sigma)$,给出如下判定方法。

定义 2:在 FU_c 集内元素 FC_n ,以原始数据作为计算样本,当 $F(\sigma)>\mu_r(\sigma)$,则认为 FC_n 发生了激剧变化,且判定元素 FC_n 是 FU_c 集的有效元素。

利用阈值函数 $\mu_r(\sigma)$ 控制数据实体对特定指标的影响程度,通过对 FU_c 集内元素的判定,逐一认定数据实体 FC_n 的有效化,去除无效元素,过滤噪声数据,降低 FU_c 集的维度,减少输入变量,简化网络结构,达到在有限数据下缩短训练周期,提高泛化能力的目的。最终得到有效 FU_c 集。

5　油田生产异常预警有效影响因子的获取

在油田生产领域,影响因子的有效性越高,异常预警的准确率就越高,这为安全生产以及生产效率提供了保障。在对油田生产开发的现有数据组成和特点分析后发现,故障发生的历史数据与生产数据的原始项目基数极大,全域内所有原始项目分析难度极大。为了提高实例效果的直观性和分析效率,缩小专家组的界定范围,通过与 8 位聚驱区块压裂作业工程师及 2 位石油勘探领域专家组成的专家小组的交流,选定生产异常预警目标 W 为压裂增油量[14]。大庆油田某采油厂所处聚驱区块为具体样本采集区块,界定针对 W 的粗糙原始项目集 U ,并对 U 中元素进行评价,得到原始项目的打分(0-1)情况,如表 2 所示。

表 2　粗糙原始项目专家打分情况

原始项目(名称代码)	自然语言评价	打分
采聚浓度(cjnd)	A little(一点)	0.30
破裂压力(plyl)	More or less(或多或少)	0.65
压裂液类型(ylye)	Somewhat(略微)	0.50
…	…	…
砂岩厚度(syhd)	Indeed(的确)	0.95
平均加砂比(pjjsb)	Very(非常)	0.75

通过模糊限制词的定量化映射,得到专家小组对

于 U 的量化结果：

$$\boldsymbol{Q}=\begin{vmatrix} \text{cyl} & \text{cmd} & \text{0.5ylye} & \cdots & \text{0.9syhd} & \text{yxhd} & \cdots & \text{0.8yxstl} \\ 0.5 & 0.3 & 0.6 & \cdots & 0.85 & 0.7 & \cdots & 0.7 \\ 0.3 & 0.15 & 0.55 & \cdots & 0.75 & 0.95 & \cdots & 0.55 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0.35 & 0.25 & 0.8 & \cdots & 0.9 & 0.9 & \cdots & 0.8 \end{vmatrix}$$

经筛选得到具有典型特征的影响因子op 80 余项。通过 TRIMMEAN(Q_r,percent) 内均法对 op 进行评估，剔除 20% 数据点，筛选得到影响因子 $\overline{\text{opv}}$ 60 余项，如表 3 所示。

表 3 粗糙原始项目集的影响因子选定

影响因子 $\overline{\text{opv}}$	粒度程度	与数据实体 du 映射关系 F
产油量/t	日/月	直接映射
压裂液类型	日/月	直接映射
破裂压力/MPa	日/月	直接映射
采聚浓度/(mg/L)	日/月	通过聚合物采出质量与体积之比计算
	…	…
平均加砂比/%	月	通过日产油量数据相减计算
砂岩厚度/m	月	通过层数据累加计算
压裂有效厚度/m	月	通过有效层数据累加计算

根据影响因子 $\overline{\text{opv}}$ 与数据实体 du 映射关系 F，截取时间序列 $T=\{t_1,t_2,\cdots,t_m\}$，选定 $T=1\ 000d$。结合压裂作业生产特点，依据上文提出的最佳粒度选取办法，使用低粒度保存近期的生产数据和汇总数据，对时间较久远的生产数据保留粒度较大的汇总数据。将时间粒度选定为：若 $T_c=400d_{\text{recent}}$，则 $m=400$；若 $T_c=600d_{\text{pre}}$，则 $m=20$。基于最佳数据粒度对数据进行处理，然后对处理后的数据依据激剧函数进行有效化判定。

图 2 和图 3 分别为压裂层段砂岩厚度以及含水分级对预警目标 W 的影响关系图。

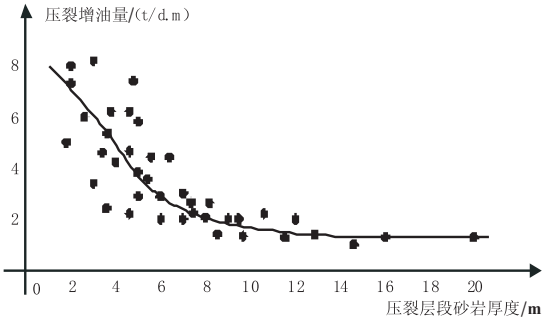

图 2 油井压裂厚度增油与压裂层段厚度关系曲线

由图 2 可以看出，压裂井的平均每米压裂砂岩厚度的增油量随组成压裂层段厚度的减少而增加，当砂岩厚度在 2 ~ 6 m 时对压裂增油量的影响程度非常大。

由图 3 可以看出，在油井自喷开采条件下不可忽视压裂井含水的高低(即层间干扰的作用)对压裂效果的影响。一般说来，油井含水低有利于压裂效果的发挥[15]，但是在油井转抽以后，油井含水对压裂效果的影响程度相对减小。

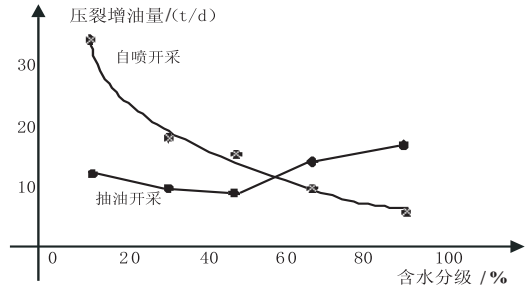

图 3 压裂增油量与含水关系曲线

经过上述步骤，最终得到针对生产异常预警目标 W(压裂增油量)的包含 12 项有效影响因子的数据集：

$\text{FU}_c=\{$压裂井点所处砂体部位，储油砂体的沉积环境，液量含水比，压裂液类型，支撑剂粒度，破裂压力，压裂时间，砂岩厚度，压裂有效厚度，层措施位平均渗透率，措施层位有效渗透率，压裂层平均加砂比$\}$

6 结束语

文中提出了时序化生产预警有效影响因子的获取方法。通过建立自然语言描述的影响因子与数据体内数据实体的映射关系，结合模糊综合评价法量化专家对影响因子的模糊语义描述，构建影响因子逻辑表达结构。采用 CF-ISF 权重算法基于特征选择技术挖掘时序化数据的敏感特征因子，利用 TRIMMEAN 内均法及均方差收敛计算等方法过滤噪声数据，同时根据激剧判定函数实现对数据的有效化判定，从而获取时序化生产预警的有效影响因子，以达到辅助生产异常动态分析、提高异常预警准确率的目的。

参考文献：

[1] Zhang Jian, Huang Kun. Research on early-warning method and its application of complex system of circular economy for oil and gas exploitation[J]. Energy Procedia, 2011, 5: 2040-2047.

[2] 王 添，姜 麟，米允龙. 海量数据下不完备信息系统的知识约简算法[J]. 计算机技术与发展，2015，25(1)：137-142.

[3] 苏新宁，杨建林，江念南，等. 数据仓库和数据挖掘[M]. 北京：清华大学出版社，2006.

[4] 王 虹，张文修，李鸿儒. 粗糙模糊集的不确定性度量[J]. 计算机工程与应用，2005，41(2)：51-52.

(下转第 132 页)

4 问题总结

在云平台的搭建过程中也遇到了一些问题,下面列出比较典型的几个,以便在以后的平台搭建过程中引起注意和提供参考。

(1)权限问题。

ssh 免密码配置后各节点间数据访问还是需要输入密码,将防火墙关闭也不行。最后将. ssh 目录下的所有文件权限设置为 600(chmod 600 ~/. ssh/ *),问题就解决了。

(2)防火墙的问题。

Hadoop 配置完成后,用 jps 命令查看发现 slave 节点中没有 nodemanager 进程,禁用防火墙后 slave 节点中出现 nodemanager 进程。

(3)格式化 HDFS 的问题。

重新格式化 HDFS 文件系统之后,发现 DataNode 无法启动。原因是每次格式化 HDFS 文件系统会重新创建一个 namenodeId,而目录"tmp/dfs/data"下包含了上次格式化后的 id,格式化 HDFS 文件系统清空了 NameNode 下的数据,但没有清空 DataNode 下的数据,所以导致启动时失败。所要做的就是每次格式化 HDFS 文件系统前,先清空 tmp 文件夹下的所有目录。

值得注意的是,在处理问题的过程中要善于通过查看运行日志找到问题的产生原因,并通过借助互联网寻求问题的处理办法来解决问题。

5 结束语

综合采用并行计算、分布式计算和虚拟化等技术的云计算将海量数据处理推进到一个新时代。而 Hadoop 的开源、跨平台、高容错等特点使其成为构建云计算平台的首选技术。文中详细介绍了 Hadoop 集群的搭建方法,成功地搭建了 Hadoop 云计算平台并进行了测试,有助于以后对大数据[14]的研究。最后将云平台搭建过程中遇到的问题进行了总结,以便引起注意并提供参考。下一步的工作主要是在 Hadoop 云计算平台下进行相关算法的研究和应用。

参考文献:

[1] 柯栋梁,郑 啸,李 乔. 云计算:实例研究与关键技术[J]. 小型微型计算机系统,2012,33(11):2321-2329.

[2] 林 利,石文昌. 构建云计算平台的开源软件综述[J]. 计算机科学,2012,39(11):1-7.

[3] Armbrust M,Fox A,Griffith R,et al. A view of cloud computing[J]. Communication of the ACM,2010,53(4):50-58.

[4] 张良将. 基于 Hadoop 云平台的海量数字图像数据挖掘的研究[D]. 上海:上海交通大学,2013.

[5] 王彦明,奉国和,薛 云. 近年来 Hadoop 国外研究综述[J]. 计算机系统应用,2013,22(6):1-5.

[6] Chaudhary A,Singh P. Big data - importance of Hadoop distributed filesystem[J]. International Journal of Scientific & Engineering Research,2013,4(11):234-237.

[7] 童 明. 基于 HDFS 的分布式存储研究与应用[D]. 武汉:华中科技大学,2012.

[8] 郝增勇. 基于 Hadoop 用户行为分析系统设计与实现[D]. 北京:北京交通大学,2014.

[9] Dean J,Ghemawat S. MapReduce:simplifier date processing on large clusters[J]. Communications of the ACM,2008,51(1):107-113.

[10] Berlinska J,Drozdowskib M. Scheduling divisible MapReduce computations[J]. Parallel and Distributed Computing,2011,71(3):450-459.

[11] 徐焕良,翟 璐,薛 卫,等. Hadoop 平台中 MapReduce 调度算法研究[J]. 计算机应用与软件,2015,32(5):1-6.

[12] Apache Hadoop[EB/OL]. 2015-08-07. http://hadoop.apache.org/.

[13] 王婷娟,管会生,尹 晖. DSA 与 RSA 相结合的数字签名技术[C]//全国第 19 届(CACIS)学术会议论文集(下册). 出版地不详:出版者不详,2008:1129-1133.

[14] 严霄凤,张德馨. 大数据研究[J]. 计算机技术与发展,2013,23(4):168-172.

(上接第 126 页)

[5] Negnevitsky M. 人工智能:智能系统指南[M]. 北京:机械工业出版社,2012.

[6] 王美方,刘培玉,朱振方. 一种基于 TFIDF 的特征选择方法[J]. 计算机工程与设计,2007,28(23):5795-5796.

[7] 张可佳,李春生,姜海英,等. 时间序列下模式挖掘模型设计[J]. 计算机工程与应用,2015,51(19):146-151.

[8] Yang Y,Pedersen J O. A comparative study on feature selection in text categorization[C]//Proceedings of 14th international conference on machine learning. Nashville,US:[s. n.],2007:412-420.

[9] 吕海燕,车晓伟. 数据仓库中数据粒度的划分[J]. 计算机工程与设计,2009,30(9):2323-2325.

[10] 王晓鹏,武 彤. 生产质量控制数据仓库模型设计与实现[J]. 计算机技术与发展,2015,25(6):181-184.

[11] 王 虎,丁世飞. 序列模式挖掘研究与发展[J]. 计算机科学,2009,36(12):14-17.

[12] 卓书月. 柯西不等式及其变式的应用[J]. 民营科技,2011(9):78-78.

[13] Duda R O,Hart P E,Stock D G. 模式分类[M]. 北京:机械工业出版社,2000:36-39.

[14] 高 建,侯加根,王 军,等. 聚合物驱后砂岩储层岩石物理特征变化机制[J]. 中国石油大学学报:自然科学版,2009,33(3):22-26.

[15] 徐松辽. 影响二类聚驱油层压裂效果的原因分析[J]. 黑龙江科技信息,2012(11):63-63.