Journal of Computer Applications
计算机应用,2017,37(4):1038-1043,1050
ISSN 1001-9081
CODEN JYIIDU
2017-04-10
http://www.joca.cn

文章编号:1001-9081(2017)04-1038-06
DOI:10.11772/j.issn.1001-9081.2017.04.1038

基于跨境电商可控关联性大数据的出口产品销量动态预测模型

王雪蓉,万年红*
(浙江东方职业技术学院 信息传媒与自动化学院,浙江 温州 325011)
(*通信作者电子邮箱 wnhhong@126.com)

摘　要:目前流行的外贸产品销量预测方法单纯地分别从第三方平台或大数据角度研究预测问题,对互联网平台、跨境电商、大数据融合应用于产品销量动态演化预测的考虑不足。为提高出口产品销量预测效果,实现预测系统的伸缩性和动态演化性,基于研究"互联网+外贸"环境下跨境电商出口产品销量可控关联性大数据挖掘、个性化预测机制、智慧预测算法,改进分布式定量、集中式定性计算等相应算法,提出一个"互联网+外贸"驱动下基于跨境电商可控关联性大数据的出口产品销量动态预测模型,并进行了应用实验,对各种模型的实验结果进行对比分析。实验结果表明,该模型充分融合了"互联网+"的开放性、可延伸性和大数据动态预测优势,实现了"互联网+外贸"环境下基于跨境电商可控关联性大数据的出口产品销量动态、智慧、定量定性预测。该模型综合预测效果明显优于传统模型,具有较强的动态演化性和较高的实用价值。

关键词:互联网+外贸;跨境电商;可控关联性大数据;出口产品销量;动态预测
中图分类号:TP311.7; TP391　　**文献标志码:**A

Dynamic prediction model on export sales based on controllable relevance big data of cross-border e-commerce

WANG Xuerong, WAN Nianhong*
(*School of Information Media and Automation, Zhejiang Dongfang Vocational and Technical College, Wenzhou Zhejiang* 325011, *China*)

Abstract: Current popular prediction methods of foreign trade product sales only respectively study prediction problems from angles of the third party platform or big data, lacking consideration of dynamic evolution prediction on product sales based on Internet platform, big data and cross-border e-commerce. To improve the efficiency of export sales prediction, to achieve scalability and dynamic evolution of prediction systems, with mining controllable relevance big data of cross-border e-commerce export sale based on "Internet + foreign trade" surroundings, personalized prediction mechanism and smart prediction algorithms, improving corresponding algorithms such as distributed quantitative calculation and centralized qualitative calculation, a dynamic prediction model on export sales based on "Internet + foreign trade"-driven controllable relevance big data of cross-border e-commerce was proposed. Finally, this model was verified and analyzed. The performance analysis results show that the model integrates fully openness and extensibility of "Internet + " and dynamic prediction advantages of big data, achieving dynamic, smart, quantitative, and qualitative prediction on export sales based on "Internet + foreign trade"-driven controllable relevance big data of cross-border e-commerce. The comprehensive prediction efficiency of the proposed model is obviously better than those of traditional models, and it has stronger dynamic evolution and higher utility.

Key words: Internet + foreign trade; cross-border e-commerce; controllable relevance big data; export sale number; dynamic prediction

0　引言

"互联网+外贸"环境下跨境电商活动比较复杂,出口产品销量预测受到需求、关税、物流、风险等多种因素的影响[1],但"互联网+外贸"跨境电商的核心是具有预测优势的大数据,这使得其出口产品销量预测相对容易,因此,设计一种"互联网+外贸"驱动下准确、安全、高效的基于跨境电商可控关联性大数据的出口产品销量预测模型已经成为备受瞩目的热门课题。目前,对产品销量预测的研究,文献[2-5]分别基于大数据分类方法、相关性分组规则、在线聚类方法以及互联网大数据匹配原则、语义分析、行为分析方法挖掘第三方平台中海量的跨境电商数据并建立了产品需求回归预测模型;文献[6-10]分别利用跨境电商历史销量数据建立产品销量预测模型;文献[11-15]通过定量和定性分析方法根据历史数据、流行度、新产品客户价值提出了基于大数据的预测模型。以上研究具有借鉴作用,但由于"互联网+"战略提出时间较短,"互联网+"强大的大数据在线预测优势没有显现,以上研究对于如何在"互联网+外贸"环境下融合运用大数据到跨境电商出口产品销量动态智慧的预测中目前并没有一个比较深入、有效的研究。

本文针对目前研究现状,从"互联网+外贸"环境下跨境电商出口产品销量可控关联性大数据挖掘、个性化预测机制、

收稿日期:2016-07-29;**修回日期:**2016-10-20。　**基金项目:**浙江省社会科学界联合会研究课题成果(2017Z03)。
作者简介:王雪蓉(1981—),女,浙江平阳人,副教授,硕士,主要研究方向:跨境电商、大数据;　万年红(1977—),男,江西南昌人,副教授,硕士,主要研究方向:互联网+、大数据、跨境电商。

智慧预测算法等角度,尝试设计一个可量化、动态、智慧的"互联网 + 外贸"驱动下基于跨境电商可控关联性大数据的出口产品销量动态预测模型(Dynamic Prediction Model on Export Sales based on controllable relevance big data of cross-border e-commerce, DPMES),着重实现出口产品销量的动态预测目标,以便更好地指导外贸企业营销和优化库存策略,同时也促进互联网 +、大数据、跨境电商技术的创新研究与应用发展。

1　DPMES 总体框架

1.1　可控关联性销量大数据定义

定义 1　可控关联性。

可控关联性是指影响研究结果的多个现象之间的可以控制的相互关联的性质[16]。算法如下:假设 u、v 分别为两类产品,论域 $I_{uv} = I_u \cup I_v$ 是"互联网 + 外贸" 定量空间,作为现象的集合;空间中的数据源矢量 $\boldsymbol{D} = (I_{uv}, I_u, I_v)$;$I = \{I_1, I_2, \cdots, I_n\}$ 是对象项目集合;u_i 和 $v_i \in I_{uv}$ 是定性概念 u_i 和 v_i 的一次定量信任约束;$m \times n$ 阶矩阵 $\boldsymbol{R}$ 是基本用户的评分矩阵;u_i 和 v_i 的确定度 $\mu(u_i)$ 和 $\mu(v_i) \in [0,1]$ 是有稳定倾向的随机数;给定目标用户 a_i 及其评分向量 $\boldsymbol{A}(1,n)$;$\mu: I_{uv} \to [0,n]$,$\forall u_i \in I_{uv}, u_i \to \mu(u_i)$;对于 $\forall i \in I_{uv}$,假设定量信任约束 u_i 和 v_i 之间的属性信任为 $S(u_i, v_i)$,将 $S(u_i, v_i)$ 最大的 n 个基本数据组成集合。输入两个对象消息对 m_1 和 m_2,对非 False 公钥,若 $m_1 \wedge m_2$ 返回值为 False,则 m_1 和 m_2 可控关联无效;若返回值为非 True,则 m_1 和 m_2 存在可控关联。

定义 2　可控关联性销量大数据。

可控关联性销量大数据指满足以上算法的大数据,其算法如下:假设产品 u 和目标产品 v 的评分大数据项集合分别为 $I_u = \{u_i \mid i \in \mathbf{N}_+\}$ 和 $I_v = \{v_i \mid i \in \mathbf{N}\}$,若 $I_v \leqslant I_u$,即对于 $\forall i \in I_v$,都有 $i \in I_u$ 成立,定量值 $u_i, v_i \in I_{uv}$ 是定性概念的一次随机实现,则产品 v 的所有评分大数据项都已被产品 u 评价过,因此 v 不可能向 u 推荐可控关联;若 $I_v > I_u$,即对于 $\forall i \in I_u$,都有 $i \in I_v$ 成立,且当 $\mu: I_{uv} \to [0,1]$,有 $\forall u_i \in I_{uv}, u_i \to \mu(u_i)$,则产品 u 的所有评分大数据项都已被产品 v 评价过,因此产品 v 的大数据必定与 u 可控关联。

1.2　跨境电商出口产品销量可控关联性大数据挖掘

"互联网 + 外贸"环境下基于跨境电商的出口产品大数据挖掘首先基于"互联网 + 外贸"环境平台(本平台由对外贸易经济合作部牵头,集中部署在国家互联网中心),挖掘政策、产品种类、客户总产品需求、交易群体、客户购买心理、支付、报价、关税、库存、物流、订单、合同以及信誉、商品质量风险、退货或换货率、假冒伪劣产品、虚假宣传等数据。需要挖掘影响出口产品销量预测的关键因素和可控关联性大数据。挖掘模型如图 1 所示。

具体挖掘流程和方法如下:

步骤 1　设计基于预测行为的大数据在线分类与估计函数,对影响预测的可控关联性指标大数据进行在线估计和归类,确定大数据挖掘方向。文献[2]已有大数据分类相关方法的运用,但没有明确刻画数据的模糊现象,且没有融入互联网思维,不能实现互联网大数据在线分类功能,在此需要进行改进,提出基于预测行为的大数据在线分类与估计方法。

假设:在 1.1 节算法基础上,给定数据挖掘论域 C 及其非空子集 A,$(C,A) = \{(C_i, A_i) \mid i \in \mathbf{N}_+\}$ 定义为可信性测度集合,$S = \{S_i = (S_{i1}, S_{i2}, S_{i3}) \mid i \in \mathbf{N}_+\}$ 表示"互联网 + 外贸"环境下第三方平台上文档和日志的集合,包括 n 个元素;$SAM = \{SAM_i \mid i \in \mathbf{N}_+\}$ 和 $SAB = \{SAB_i \mid i \in \mathbf{N}_+\}$ 分别表示数据模糊现象和不确定性现象集合;分类主体与客体映射函数为 $F1 = (FS_i \to FC_i)$。若 $\exists C_i \in C \wedge (C_i, A_i) \wedge (S_{i1}, S_{i2},$

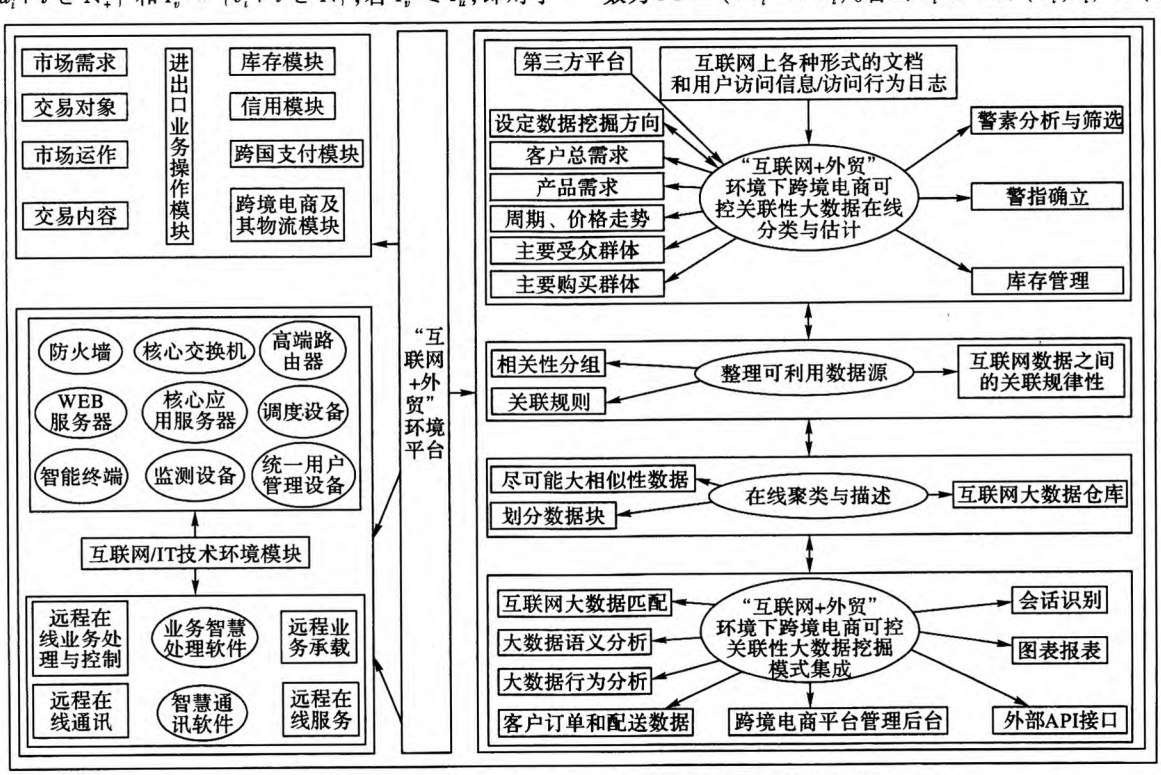

图 1　"互联网 + 外贸"环境下跨境电商出口产品销量可控关联性大数据挖掘模型

Fig. 1　Mining model of controllable relevance big data of cross-border ecommerce export sales based on "Internet + foreign trade" surroundings

$S_{i3}) \wedge SAM \wedge SAB \neq \text{False}$，则基于预测行为的大数据在线分类与估计方法可用如式(1)所示的函数 $CF(i)$ 表示：

$$CF(i) = \frac{A(1,n) \cdot |S_i| \cdot |FS_i| \cdot I_{uv} \cdot \mu(v_i) \cdot \sum_{i=1}^{n} SAM_i}{(I_{uv}, I_u, I_v) \cdot \mu(u_i) \cdot |C_i| \cdot |FC_i| \cdot |A_i| \cdot \sum_{i=1}^{n} SAB_i} \quad (1)$$

改进后的 $CF(i)$ 将模糊现象和不确定性现象各自映射到不同的 (C_i, A_i) 中。这比改进前的方法更能刻画互联网大数据在线分类的模糊和不确定性。

步骤2 设计关联规则函数，找出这些可控关联性大数据之间的关联规律性进而整理可利用数据源。文献[3]提出了一个相关性分组规则，但该规则仅仅是对数据初步的相关，关联精度低，因此需要改进，提出更精确的关联规则函数。

在式(1)基础上，设 $K1$、$K2$ 分别是可能性空间、必然性测度空间；S_i 在 A 集中约束时间为 t_i，ξ 为定义在 $K1$ 上的模糊变量；当前关联度 J_i 与不确定性变量 x 的相关性期望值 $E(x) = b$，阈值为 ω_0；满足映射“G:$K1 + K2 \rightarrow K1 \cdot K2$”的约束时间表示一个最佳搜索时间 T_i；关联度误差记为：$e(S_i, S_j) = b - \sum_{x=i}^{j} E(x)$，则改进的关联规则可用式(2)所示的函数表示：

$$CV(S_i, S_j) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} e(S_i, S_j) \cdot \xi \cdot (|CF(i) - K1|)}{\omega_0 \cdot \prod_{i=1}^{n} (K2 \cdot t_i \cdot T_i \cdot J_i)} \quad (2)$$

改进后式(2)的作用就在于准确实现了可控关联性出口产品销量大数据之间的关联规律性。

步骤3 设计聚类功能更强的在线 k-Means 聚类与描述函数，使已归类可控关联性大数据相似性尽可能大，划分数据块。文献[4]提出了一种在线聚类方法，然而该方法划分的数据块明显存在越界现象。由于 k-Means 聚类方法具有关联聚类的功能，因此本文结合这两种方法进行改进，设计一个在线 k-Means 聚类与描述方法。

在式(2)定义基础上，假设预测主体、客体推荐的权重分别为 ω_1, ω_2；聚类推荐集合为 $TJ = \{tj_1, tj_2, \cdots, tj_n\}$，相关联描述的期望值 $Qx(tj_i, tj_j) = TS(SAM_{ii}, SAB_i) \cdot e(S_i, S_j)$，熵值 $Qn(tj_i, tj_j) = TS(SAM_{ii}, SAB_i) + e(S_i, S_j)$，超熵值 $Qe(tj_i, tj_j) = TS(SAM_{ii}, SAB_i)/e(S_i, S_j)$，则改进的在线 k-Means 聚类与描述方法可用式(3)所示的函数动态表示：

$$CVV(S_i, S_j) = \frac{Qx(tj_i, tj_j) \cdot \sum_{i=1}^{m}\sum_{j=1}^{n} (tj_i \cdot tj_j \cdot |\omega_i \cdot TJ|)}{Qn(tj_i, tj_j) \cdot Qe(tj_i, tj_j) \cdot CV(S_i, S_j)} \quad (3)$$

改进式(3)的作用就在于可以尽可能有效地划分相似性“互联网 + 外贸”跨境电商产品销量数据块。

步骤4 运用文献[5]提出的互联网大数据匹配原则、语义分析、行为分析方法，针对主要用户合理匹配保留可控关联性销量大数据的高度演化特征，通过线上或线下方式将影响预测的关键的可控关联性大数据导入，集成到互联网大数据仓库、跨境电商平台管理后台和外部应用程序接口，实现关键因子的集成，较好地解决了主体客体属性混淆、语义控制矩阵体现预测行为域间映射的问题。

1.3 个性化预测机制

根据图1所示模型，构建如下个性化预测机制：

1)“增量演化-集成”式预测机制。

通过神经网络方法，实现增量式的动态演化集成的准确预测。其中增量演化因子属性可表示为 B_n^2 元组 $Q = (B_{n-1}$ 与 B_n 的数学组合)，组合数 $M_1 = \omega_i CER_n^2$。

2)“随机分布-关联”式预测机制。

通过机器学习，将随机分布的产品销量大数据关联起来进行预测。其中随机分布因子属性可表示为 B_n^2 元组 $Q = (B_{n-1}$ 与 B_n 的数学组合)，组合数 $M_2 = CER_n^2$。

为定量定性实现个性化预测，需以数学形式来表达这两种机制。C&M-CVPDSS(Case-based and Multiplicative analytic hierarchy process-based Customer Value Prediction Decision Support System)[11]较好表达了这种机制，但其对新产品关键数据在随机分布、关联、演化、集成的指标化和相似度评估方面表现不足，因此，下面在1.1节可控关联性大数据定义和式(1)～(3)的基础上，对 C&M-CVPDSS 进行改进。

假设个性化的机器学习、神经网络的参照样本库并集为 $mint = \{CER(C_1C_1), CER(C_2C_2), \cdots, CER(C_iC_j)\}$，某一个由市场需求、交易对象、市场运作、交易内容集成问题组成的随机分布、关联、演化问题集为 $AAT = \{AA_i \mid i \in \mathbf{N}_+\}$，数据分配函数[12]为 $G(C_i, C_j) = (CER(C_iC_j)/n, m)$，根据增量演化因子属性和随机分布因子属性元组，则“增量演化-集成”式预测机制和“随机分布-关联”式预测机制可分别用式(4)、(5)所示的数学函数来表达：

$$DT = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} (|CER(C_iC_j)| \cdot |B_j| / G(C_i, C_j))}{Q \cdot B_n^2 \cdot AA_i \cdot CVV(S_i, S_j)} \quad (4)$$

$$DTT = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} (|AA_i \cdot CVV(S_i, S_j) \cdot G(C_i, C_j)| / B_j)}{DT \cdot Q \cdot B_n^2 \cdot CER(C_iC_j)} \quad (5)$$

其中：DT 表示自动实现从复杂的增量演化、集成指标到具体的机制转换；DTT 表示自动实现从抽象的随机分布、关联指标和相似度评估到具体的模式转换。式(4)、(5)自动实现从抽象的随机分布、关联、演化、集成指标和相似度评估到具体的机制、模式转换。

1.4 DPMES 的智慧预测算法

步骤1 按式(1)对 $CF(i)$ 求解，按预测方向进行操作，当 $CF(i)$ 值域不为空集时，则“互联网 + 外贸”环境下可控关联性出口产品销量大数据预测资源规划设计与集成策略如式(6)所示：

$$(C_i, A_i) : \{(C, A)\} \xleftrightarrow{CF(i)} FS_i : \{FC_i \mid i \in \mathbf{N}_+\} \quad (6)$$

利用式(6)提取可控关联性预测影响因素，在线智慧分类，使预测实体间具有严格的物理映射关系。

步骤2 设计动态预测智慧集成策略，对预测构件进行动态地加入或删除操作。按式(2)计算模糊变量的定义域，抽取出 DT 的最大值 DT_{max} 和 DTT 的最小值 DTT_{min}，选取若干个满足取值范围为 $[DTT_{min}, DT_{max}]$ 的预测构件。集成策略可用如式(7)所示的约束系数 λ 表示：

$$\lambda = \frac{(C_i, A_i)}{FS_i : FC_i} \quad (7)$$

步骤3 按照式(7)，抽取 $K1$ 和 $K2$ 中所有属于 (C, A) 的

任意满足阈值 ω_0 的 ξ,寻找最佳搜索时间 T_i。通过式(2) ~ (3) 发现数据可控关联性和规律性。基于产品标识失效预测方法[14],设计可控关联性出口产品销量大数据动态预测评估集成策略,其函数如式(8)所示,即计算 $CVV(S_i,S_j)$ 对 $Qn(tj_i,tj_j)$ 的动态优化集成的隶属度 $M(S_i,S_j)$:

$$M(S_i,S_j) = CVV(S_i,S_j) \cdot \lambda \cdot DT \cdot DTT \tag{8}$$

步骤 4　根据式(4) ~ (5)增量演化因子属性和随机分布因子属性划分预测行为类属 BT,目标预测评分与实际评分之间的偏差为 MAE,并使得簇之间的相似度达到最小值,而 $\mu(x)$ 和 $\mu(y)$ 之间的相似度达到最大值,将具备不同评分特征的若干目标预测行为划入隶属度不同的行为子集中,实现协同过滤推荐。

步骤 5　设计作为本文算法最关键算法的分布式定量、集中式定性等动态优化预测方法。文献[11 - 15]运用定量定性方法来解决动态优化问题,但没有较好解决分布式和集中式优化计算问题。而多元线性回归方法[2-5,12]根据关键因子采用二维表,能较好地解决数据的分布式、集中式、相关性预测问题,因此本文基于此方法对定量和定性方法进行如下改进:根据 $Qe(tj_i,tj_j)$ 计算所有的 $M(S_i,S_j)$,各节点根据其所有邻居节点当前位置,动态地选择下一个簇头节点。

为此按顺序分别令集群、分割、孤立点为:

$$\begin{cases} HQ = \dfrac{(CV(S_i,S_j)^{T_i})^{-1/4}(DT^{T_i})^{-1/2}(CF(i)^{T_i})^{-1/3}}{M(S_i,S_j)} \\ HF = (FS_i^{T_i})^{-1/3}(CVV(S_i,S_j)^{T_i})^{-1/2}(DTT^{T_i})^{-1/4} \\ HG = CVV(S_i,S_j)^{\lambda/2}CV(S_i,S_j)^{\lambda}(DTT^{\lambda})^{-1/3} \end{cases} \tag{9}$$

则分布式定量计算、集中式定性计算函数分别如式(10)、(11)所示,定量、定性预测出口产品销量。

$$TCL(S_i) = \frac{HG \cdot |CF(i)| \cdot \sum_{x=1}^{n}(CV(S_i,S_j) \cdot DTT)}{HF \cdot FS_i \cdot DT \cdot \prod_{x=1}^{n}(\lambda \cdot CVV(S_i,S_j))} \tag{10}$$

$$TCX(S_i) = \frac{HQ \cdot |M(S_i,S_j)| \cdot FS_i \cdot \sum_{x=1}^{n}CV(S_i,S_j)}{CVV(S_i,S_j) \cdot DTT \cdot \prod_{x=1}^{n}(\lambda \cdot TCL(S_i))} \tag{11}$$

步骤 6　设 $(C_iC_j)_{max}$ 和 $(C_iC_j)_{min}$ 分别表示 C_iC_j 的最大值和最小值。重复步骤 1 ~ 5,将 C_i 的代码特征 C_iC_j 固定在阈值区间 $[\omega_i,\omega_j]$,判断 $(C_iC_j)_{max} > \omega_i \wedge (C_iC_j)_{min} < \omega_j$ 是否成立,若成立采用式(10)从海量大数据中查找产品正常的可控关联性大数据,构建分布式定量计算系统,得出定量预测结果;若不成立,采用式(11)从时空上将被求解的问题集群,建立集中式定性智慧预测模型。

步骤 7　设计如式(12)所示的并行式综合预测函数公式:

$$TCZ(S_i) = TCX(S_i)^{\lambda} \cdot TCL(S_i)^{T_i} \tag{12}$$

至此,算法结束。

2　DPMES 动态预测模型构建

通过上述个性化预测机制和智慧预测算法,运用决策树构建 DPMES 动态预测模型,如图 2 所示。

构建路径如下:

首先,根据智慧预测算法步骤 1,确定预测目标及选择用于建模的数据样本范围,并筛选、过滤得到具有预测特征和能力的若干因子。

其次,根据智慧预测算法步骤 2 ~ 3,归纳可控关联性关键因子,利用"增量演化-集成"和"随机分布-关联"式预测机制验证关键特征数据序列一致性、可控关联性和规律性,量化各种随机分布的出口产品销量关键影响因素间的增量演化、集成关系,并利用 λ 动态地加入或删除预测构件。整合关键因子,预备、估算、清洗、非线性变换和校验数据。

然后,根据算法步骤 4 实现协同过滤推荐,划分预测行为子集。

再次,根据式(10) ~ (12)集中实时跟踪可控关联性大数据流,合成、错位对齐互联网大数据搜索指数,选出具有最大搜索指数的关键数据作为基准指数,构建模型,在线重配、赋予并行式大数据权重和相关系数,集中式定性预测哪些潜在客户最可能成为消费者和交易者,并对可能的交易线索进行显著性检验,分布式定量预测下一周期的销售量,实时预测出口产品未来销量结构走势。

最后,模型评价与应用。对每种预测方法的预测结果进行误差分析,如果方法综合时对前几期预测结果的平均误差越大,那么综合预测时应该使该方法对综合预测结果的影响程度越小。决策树根据筛选过滤出的权重和相关系数进一步划分出叶节点,待模型稳定后即可得到销量的综合预测值,并据此实现库存策略的优化,实现模型应用。

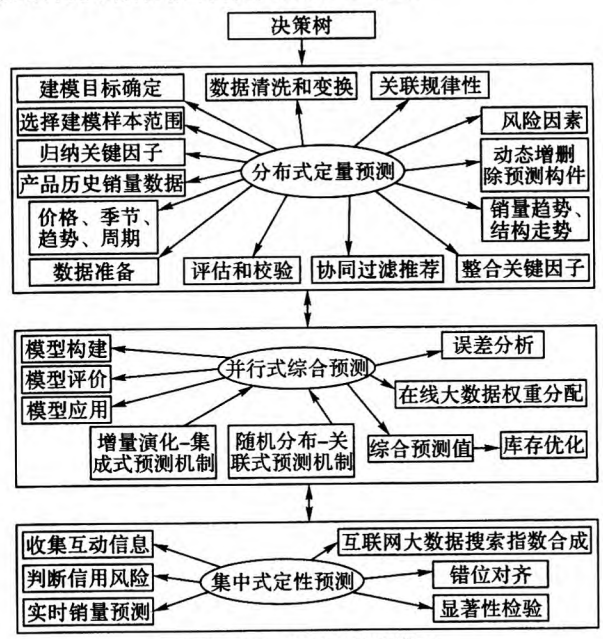

图 2　DPMES 动态预测模型

Fig. 2　DPMES dynamic prediction model

3　算法模型应用实例

根据图 1 ~ 2,前端采用 Java 语言(或 C#/C +),后台采用开源 PHP 和 SQL Server 构建动态预测系统。

3.1　实验数据收集

本文所采用的数据均来自 2015 年 1 月至 2016 年 7 月阿里巴巴全球速卖通平台、对外贸易经济合作部"互联网 + 外贸"平台、中国跨境电商综合服务平台、国家统计局固定产品

销量指标统计平台、浙江省跨境电商公共服务平台这5个平台内关于皮鞋、机械、电器等制造业出口产品的10000条固定数据以及平台外的若干实际动态数据。根据1.2节大数据挖掘模型,采用关键词挖掘工具从“销量关注度”和“因子关注度”两个维度去描述关键词,并以一周为一个更新周期。数据分析如表1所示。

表1 实验数据分析

Tab. 1 Experimental data analysis

可控关联性指标(关键词)(范围:10~15)	销量关注度	因子关注度
产品品类	9.976670	主营销售产品占比
客户总需求	13.557320	主要受众群体比率、产品需求量
客户购买心理	13.642350	兴趣爱好、购买能力和行为比率
产品周期	10.547280	生产制造、销售周期
库存	10.469230	采购量、积压量、库存量、成本、补货出货量、库存均衡与调拨、存储位置
价格	14.682140	跨国支付、报价、投保、关税、成本、利润
物流	10.467822	仓储、运输配送、供应链成本
风险	14.973471	商品质量与性能、退货或换货率、信用等级

3.2 实验结果对比分析

实验1 用以上10000条固定数据,验证改进方法及基于改进方法的个性化预测机制和智慧预测算法的合理性。过程如下:

1)验证产品 v 的大数据必与产品 u 可控关联,$E(x)=b$ 按 $CF(i)$ 进行约束,计算隶属度、关键数据随机分布、关联、演化、集成的指标化和相似度评估,验证式(1)~(9)有效性。

2)验证式(10)~(12)分布式定量、集中式定性预测、并行式综合预测函数。

验证指标包括重配误差(出口产品销量大数据间存在的重新匹配差异)、特征点误差(出口产品销量可控关联性特征匹配的差异)、误配率(出口产品销量可控关联性大数据发生错误匹配的比率),计算公式参见文献[13-15]的互联网搜索误差均方。为更好地表达误差关系,对该误差均方添加 ω_0 线性参数为 $(b_1,b_2,\cdots,b_n)$,则重配误差、特征点误差、误配率公式分别用如式(13)所示的 $CPW(v,u)$、$APW(v,u)$、$WPR(v,u)$ 函数表示:

$$\begin{cases} CPW(v,u)=\dfrac{M(S_i,S_j)^{\lambda}\cdot\sqrt{CV(S_i,S_j)^{\omega 0}+TCX(S_i)^2}}{CVV(S_i,S_j)^{T_i}\otimes(TCZ(S_i)\cdot TCL(S_i))} \\ APW(v,u)=\dfrac{\prod_{i=1}^{n}TCX(S_i)^3\cdot M(S_i,S_j)^{\omega 0/2}\cdot CPW(v,u)}{DT^{T_i}\cdot TCZ(S_i)^{\lambda/3}\cdot TCL(S_i)} \\ WPR(v,u)=\dfrac{\sqrt[T_i]{TCL(S_i)^3\cdot TCX(S_i)}}{APW(v,u)\cdot DTT^{\omega 0}\cdot TCZ(S_i)^{b_n/4}} \end{cases} \tag{13}$$

验证结果分析如图3所示。从图3可以看出:重配误差散点图、特征点误差散点图、误配率散点图均为单调、并行递增形式,当散点图平滑趋近时则说明个性化预测机制更靠近实际预测结果;当分析先行的归一化拟合数据指标对出口产品销量波动的预测效率时,以改进算法作为引导方法的预测效率最高。这说明对改进方法以及基于此的个性化预测机制和智慧预测算法的设计是科学合理的。

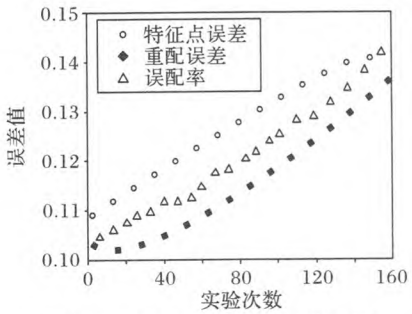

图3 改进算法验证结果直观散点图

Fig. 3 Scatter diagram for verify results of improved algorithm

实验2 基于实验1,仍然使用以上10000条固定数据进行10000次实验,实验获得的平均值作为结果数据,将DPMES与文献[2-5]所建模型及C&M-CVPDSS模型[11]、产品标识失效预测模型[14]的性能进行对比。

评价本实验结果的指标[1-15]如下:可信性测度,即表示预测结果的可信程度,参考值为16~22;不确定性区分度,即区分预测结果的多种可能状态的程度,参考值为10~15;最佳搜索时间,即衡量动态预测整体耗费的时间,参考值为10~15 s;误差系数,由重配误差、特征点误差、误配率公式综合得出,参考值为5~7;可控关联度,该指标和以上4个指标相联系,参考值为10~14。具体计算过程见文献[1-15],其公式按顺序分别用式(14)~(18)所示的函数表示:

$$KXX(x)=\frac{TCZ(S_i)\cdot\sqrt{TCL(S_i)\cdot M(S_i,S_j)^{\lambda}}}{(CPW(v,u)^{\omega 0}+TCX(S_i)^2)^{T_i}} \tag{14}$$

$$NQQ(x)=\frac{TCZ(S_i)^{T_i}\times TCL(S_i)\oplus M(S_i,S_j)^{\lambda}}{APW(v,u)^{\omega 0}\otimes TCX(S_i)^{b_n}} \tag{15}$$

$$ZJS(x)=\frac{TCZ(S_i)^{b_n}\cdot TCL(S_i)^{\omega 0}}{WPR(v,u)^{T_i}\cdot\int_0^{b_n}KXX(x)} \tag{16}$$

$$WCXS(x)=\frac{APW(v,u)\cdot TCZ(S_i)^{\omega}\cdot NQQ(x)^{b_n}}{WPR(v,u)\cdot CPW(v,u)^{T_i}\cdot\prod_{x=1}^{n}KXX(x)} \tag{17}$$

$$KKD(x)=\frac{WCXS(x)\cdot WPR(v,u)\cdot APW(v,u)}{NQQ(x)^{b_n}\cdot\sum_{x=1}^{n}(KXX(x)\cdot ZJS(x))\cdot CPW(v,u)} \tag{18}$$

实验结果如表2所示。

表2 各种模型的性能对比

Tab. 2 Performance comparison of various models

算法实例	可信性测度	不确定性区分度	最佳搜索时间/s	误差系数	可控关联度
文献[2]模型	18.6	12.7	12.1	6.8	—
文献[3]模型	17.6	12.4	12.4	6.9	—
文献[4]模型	18.3	13.3	12.8	6.6	10.1
文献[5]模型	18.4	12.1	12.6	6.3	11.3
C&M-CVPDSS模型	19.3	13.2	13.1	6.0	11.5
产品标识失效预测模型	19.5	13.1	13.2	6.1	12.2
DPMES	21.9	14.9	14.7	5.1	13.7

从表2可以发现,当固定销量指标大于期望值时,即产品销售趋热时,对增量指标的预测更接近平稳、准确的实际预测

值;反之则销量同比减少率在风险线附近徘徊。无论是产品销量期望预测指标还是实际预测指标,DPMES 的综合预测效果明显优于其他模型。

实验 3　分别通过以上 5 个平台的样本外预测和波动预测考察 DPMES 的未来预测效率。

1) 样本外分布式定量、集中式定性预测效率。

固定指标样本外预测的均方误差为:

$$MSE(S_i) = \frac{KXX(x) \cdot WCXS(x) \cdot TCL(S_i)^{T_i \cdot e}}{KKD(x) \cdot TCZ(S_i)^{\lambda} \cdot MAE} \quad (19)$$

根据 $MSE(S_i)$ 计算样本外预测的分布式定量、集中式定性误差百分比,可以采用以下方法对比预测结果:使用 $CV(S_i,S_j)$ 的多维分解方法和可联性等级,考虑 $CVV(S_i,S_j)$ 对误差关系的影响,在同比增长率偏低时下一周期预测误差达到最大;考虑 $TCZ(S_i)$ 对随机预测信任关系的影响,预测误差具有可比性。

2) 并行式综合波动预测效率。

基于前期的平台内固定和平台外若干实际动态样本数据,同样利用式(10) ~ (19),对预测时间段进行波动预测以并行式综合考察 DPMES 预测效果对出口产品销量未来一年(分四个季度)波动的预测效率。评价指标有预测误差比率(总体预测结果的误差比)、置信度(表示为近期的出口产品销量波动预测精度的置信程度)、库存优化效率(根据综合预测结果的库存优化性价比)等,具体计算过程见文献[13 - 15],其计算公式按顺序分别用式(20) ~ (22) 所示的函数表示:

$$YWR(x) = \frac{WCXS(x) \oplus CPW(v,u)^{\lambda \cdot e} \cdot APW(v,u)^{T_i}}{TCZ(S_i)^{\lambda \cdot T_i} \cdot WPR(v,u) \cdot MSE(S_i)} \quad (20)$$

$$ZXCD(x) = \frac{KXX(x)/CPW(v,u)^{\lambda \cdot T_i} \cdot APW(v,u)^{T_i}}{TCZ(S_i) \otimes NQQ(x)^{\lambda \cdot e} \cdot YWR(x)} \quad (21)$$

$$KCYR(x) = \frac{YWR(x) \cdot \sum_{i=1}^{n} (TCX(S_i) \cdot TCL(S_i)^{\lambda \omega})}{APW(v,u)^{T_i} \cdot TCZ(S_i) \times ZXCD(x)^{\lambda \cdot e}} \quad (22)$$

预测结果如表 3 所示。从表 3 可看出,各季度销量预测值呈增长趋势,而预测误差比率、置信度、库存优化效率基本上在可接受的范围内。基于各季度的并行式综合预测结果,可以计算出各季度的累积增长率,这与实际的结果非常接近,因此基于平台内和平台外样本数据的 DPMES 预测结果对出口产品销量波动有较高的预测精度和库存效率。

表 3　基于平台内和平台外样本数据的 DPMES 预测结果

Tab. 3　DPMES prediction results based on sample data inside and outside the platform

季度	期望值	实际值	预测值	预测误差比率/%	置信度/%	库存优化效率/%
一	20110	20124	20103	2.30	92.50	89.40
二	22450	23451	23132	2.10	91.80	90.10
三	24563	24785	24634	2.40	90.90	90.30
四	25735	25895	25976	2.40	93.00	91.20

4　结语

基于预测行为的大数据在线分类与估计方法、关联规则函数、在线 k-Means 聚类与描述函数以及个性化预测机制、分布式定量、集中式定性、并行式综合预测方法建立起来的 DPMES 预测算法和模型具有科学性、合理性,解决了一些理论和实际问题,充分融合了“互联网 +”的开放性、可延伸性、在线化和大数据动态演化性及预测优势,实现了“互联网 + 外贸”驱动环境下基于跨境电商可控关联性大数据的出口产品销量动态、智慧、定量化、定性化预测,对外贸企业高效营销、制订高效的库存规划具有参考价值。

但是,鉴于“互联网 +”、大数据都是高度复杂的技术,本文的研究仅仅对互联网 + 和大数据技术进行应用,虽然实际过程中也对这些计算机科学技术进行了创新,但是并不容易实现,从而一定程度上降低了系统的性能,因此今后本文作者将继续对互联网 +、跨境电商、大数据技术及融合算法继续展开研究。

参考文献(References)

[1] 王翀. 跨境电商是有出有进的“互联网 + 外贸”[J]. 杭州(周刊), 2015(14): 20 - 21. (WANG C. Cross-border e-commerce is the “Internet + foreign trade” with import and export [J]. Hangzhou (Weekly), 2015(14): 20 - 21.)

[2] KULKARNI G, KANNAN P K, MOE W. Using online search data to forecast new product sales [J]. Decision Support Systems, 2012, 52(3): 604 - 611.

[3] KAWA A, ZDRENKA W. Conception of integrator in cross-border e-commerce [J]. Scientific Journal of Logistics, 2016, 12(1): 63 - 73.

[4] ASOSHEH A, SHAHIDI-NEJAD H, KHODKARI H. A model of a localized cross-border e-commerce [J]. I-Business, 2012, 4(2): 136 - 145.

[5] OKSANEN T. Use of demand forecast in operational purchasing [EB/OL]. [2016- 02- 03]. http://www.theseus.fi/bitstream/handle/10024/97344/Tuomas_Oksanen.pdf?sequence=1.

[6] 崔东佳. 大数据时代背景下的品牌汽车销量预测的实证研究——以网络搜索数据为例[D]. 开封: 河南大学, 2014: 5 - 44. (CUI D J. An empirical study of automobile sale forecast under the background of big data — based on Web search data [D]. Kaifeng: Henan University, 2014: 5 - 44.)

[7] 孔令顶. 基于互联网搜索量的大众途观汽车销量预测研究[J]. 时代金融, 2015(30): 222, 226. (KONG L D. Prediction research on the Tiguan sales based on Internet searches[J]. Times Finance, 2015(30): 222, 226.)

[8] 李铖瀚. 基于海量数据的销售预测研究与实现[D]. 杭州: 浙江理工大学, 2015: 2 - 57. (LI C H. Research and implementation of sales forecast based on massive data [D]. Hangzhou: Zhejiang Sci-Tech University, 2015: 2 - 57.)

[9] 周昊明. 销量数据挖掘技术及电子商务应用研究[D]. 广州: 广东工业大学, 2014: 3 - 67. (ZHOU H M. Research on sales data mining technology and e-commerce application [D]. Guangzhou: Guangdong University of Technology, 2014: 3 - 67.)

[10] HE Z Z, ZHANG Z F, CHEN C M, et al. E-commerce business model mining and prediction [J]. Frontiers of Information Technology and Electronic Engineering, 2015, 16(9): 707 - 719.

[11] 罗新星, 邓丽, 赵玉洁. 基于 CBR 和 MAHP 的新产品客户价值预测决策支持系统[J]. 计算机集成制造系统, 2014, 20(10): 2403 - 2410. (LUO X X, DENG L, ZHAO Y J. Decision support system for customer value prediction of new product based on CBR and MAHP [J]. Computer Integrated Manufacturing Systems, 2014, 20(10): 2403 - 2410.)

(下转第 1050 页)

ciation for Computational Linguistics, 2003, 17: 31 - 38.

[4] SASANO R, KUROHASHI S, OKUMURA M. A simple approach to unknown word processing in Japanese morphological analysis [J]. Nuclear Physics A, 2014, 21(6): 1183 - 1205.

[5] WANG A, KAN M Y. Mining informal language from Chinese microtext: joint word recognition and segmentation [EB/OL]. [2016-01-06]. http://www.aclweb.org/old_anthology/P/P13/P13-1072.pdf.

[6] SUN X, WANG H, LI W. Fast online training with frequency-adaptive learning rates for Chinese word segmentation and new word detection [C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers. Stroudsburg, PA: Association for Computational Linguistics, 2012, 1: 253 - 262.

[7] HUANG M, YE B, WANG Y, et al. New word detection for sentiment analysis [EB/OL]. [2016-01-03]. http://mirror.aclweb.org/acl2014/P14-1/pdf/P14-1050.pdf.

[8] 邢恩军, 赵富强. 基于上下文词频词汇量指标的新词发现方法[J]. 计算机应用与软件, 2016, 33(6): 64 - 67. (XING E J, ZHAO F Q. A novel approach for Chinese new word identification based on contextual word frequency-contextual word count [J]. Computer Applications and Software, 2016, 33(6): 64 - 67.)

[9] NUO M, LIU H, LONG C, et al. Tibetan unknown word identification from news corpora for supporting lexicon-based Tibetan word segmentation [EB/OL]. [2016-01-03]. http://rsr.csdb.cn/serverfiles/csdb/paper/upload/20151021/201510210132497839.pdf.

[10] 杜丽萍, 李晓戈, 于根, 等. 基于互信息改进算法的新词发现对中文分词系统改进[J]. 北京大学学报(自然科学版), 2016, 52(1): 35 - 40. (DU L P, LI X G, YU G, et al. New word detection based on an improved PMI algorithm for enhancing segmentation system [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(1): 35 - 40.)

[11] LI C, XU Y. Based on support vector and word features new word discovery research [M]// Trustworthy Computing and Services. Berlin: Springer, 2013: 287 - 294.

[12] ATTIA M, SAMIH Y, SHAALAN K, et al. The floating Arabic dictionary: an automatic method for updating a lexical database through the detection and lemmatization of unknown words [EB/OL]. [2016-01-03]. http://www.aclweb.org/anthology/C12-1006.

[13] FRANTZI K, ANANIADOU S, MIMA H. Automatic recognition of multi-word terms: the C-value/NC-value method [J]. International Journal on Digital Libraries, 2000, 3(2): 115 - 130.

[14] HUANG J H, POWERS D. Chinese word segmentation based on contextual entropy [EB/OL]. [2016-01-06]. http://www.aclweb.org/website/old_anthology/Y/Y03/Y03-1017.pdf.

[15] YE Y, WU Q, LI Y, et al. Unknown Chinese word extraction based on variety of overlapping strings [J]. Information Processing and Management, 2013, 49(2): 497 - 512.

[16] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2001: 282 - 289.

[17] LI H, HUANG C, GAO J, et al. The use of SVM for Chinese new word identification [C]// Proceedings of the 1st International Joint Conference on Natural Language Processing. Berlin: Springer, 2004: 723 - 732.

[18] XIA F. The segmentation guidelines for the PENN Chinese treebank (3.0) [EB/OL]. [2016-01-07]. http://repository.upenn.edu/cgi/viewcontent.cgi?article=1038&context=ircs_reports.

This work is partially supported by National Natural Science Foundation of China (61370130, 61473294), the Fundamental Research Funds for the Central Universities (2014RC040), the International Science and Technology Cooperation Program of China (2014DFA11350).

ZHOU Shuangshuang, born in 1991, M. S. candidate. Her research interests include natural language processing, information extraction.

XU Jin'an, born in 1970, Ph. D., associate professor. His research interests include natural language processing, machine translation.

CHEN Yufeng, born in 1981, Ph. D., associate professor. Her research interests include natural language processing, artificial intelligence.

ZHANG Yujie, born in 1961, Ph. D., professor. Her research interests include natural language processing, machine translation.

(上接第 1043 页)

[12] 杨波, 刘勇, 牟少敏, 等. 大数据背景下山东省二代玉米螟发生程度预测模型的构建[J]. 计算机研究与发展, 2014, 51(增刊2): 160 - 165. (YANG B, LIU Y, MU S M, et al. Based on big data: the establishment of meteorological forecast model for the occurrence degree of the second generation of corn borer in Shandong [J]. Journal of Computer Research and Development, 2014, 51(Suppl2): 160 - 165.)

[13] 孔庆超, 毛文吉. 基于动态演化的讨论帖流行度预测[J]. 软件学报, 2014, 25(12): 2767 - 2776. (KONG Q C, MAO W J. Predicting popularity of forum threads based on dynamic evolution [J]. Journal of Software, 2014, 25(12): 2767 - 2776.)

[14] 王健, 何卫平, 李夏霜, 等. 基于制造历史数据的产品标识失效预测与补救方法[J]. 计算机集成制造系统, 2015, 21(9): 2494 - 2503. (WANG J, HE W P, LI X S, et al. Prediction and remediation of failed product identification based on manufacturing history data [J]. Computer Integrated Manufacturing Systems, 2015, 21(9): 2494 - 2503.)

[15] 王炼, 贾建民. 基于网络搜索的票房预测模型——来自中国电影市场的证据[J]. 系统工程理论与实践, 2014, 34(12): 3079 - 3090. (WANG L, JIA J M. Forecasting box office performance based on online search: evidence from Chinese movie industry [J]. Systems Engineering — Theory and Practice, 2014, 34(12): 3079 - 3090.)

[16] 岳笑含, 周福才, 林慕清, 等. 面向可信移动平台具有用户可控关联性的匿名证明方案[J]. 计算机学报, 2013, 36(7): 1434 - 1447. (YUE X H, ZHOU F C, LIN M Q, et al. Anonymous attestation scheme with user-controlled-linkability for trusted mobile platform [J]. Chinese Journal of Computers, 2013, 36(7): 1434 - 1447.)

This paper is supported by Research Projects of Zhejiang Federation of Humanities and Social Sciences Circles (2017Z03).

WANG Xuerong, born in 1981, M. S., associate professor. Her research interests include cross-border e-commerce, big data.

WAN Nianhong, born in 1977, M. S., associate professor. His research interests include Internet +, big data, cross-border e-commerce.