基于集成模型的BOM近似度量方法

吴文李[1],范小朋[2],周庚申[1],黄 羿[1],曹 旸[1],林桂婵[1]
(1. 中国长城科技集团股份有限公司,深圳广东 518000;2. 中国科学院深圳先进技术研究院,深圳广东 518000)

摘 要: 为满足多品种小批次、大规模定制模式下有效划分产品族的需求,全面分析BOM(Bill of Materials,物料清单)所包含的特征,概括已有结构近似方法并提出内容近似度量模型,在此基础上提出组合两者的集成模型. 结构近似模型方面,以包含BOM层次结构和物料数量的相邻矩阵表示BOM,利用正交普氏分析法计算BOM与BOM之间的近似程度. 内容近似模型方面,从BOM文本中提取有效特征,引入逆向词频法将文本特征转换成机器可识别向量形式,采用余弦近似公式完成向量近似的计算. 集成模型提出基于基尼系数的权重分配方法集成结构和内容两种模型. 最后,提供测试框架并通过实验评价集成模型较已有方法在模型性能及训练耗时上的优劣.

关键词: 相似性度量; 物料清单; 产品族; 集成模型
中图分类号: TP391.7 **文献标识码:** A **文章编号:** 0372-2112 (2019)05-1023-06
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.05.007

A Novel BOM Similarity Metric Method Based on Ensemble Model

WU Wen-li[1], FAN Xiao-peng[2], ZHOU Geng-shen[1], HUANG Yi[1], CAO Yang[1], LIN Gui-chan[1]
(1. *China Greatwall Technology Group Co., Ltd, Shenzhen, Guangdong* 518000, *China*;
2. *Shenzhen Institutes of Advanced Technology, China Academy of Sciences, Shenzhen, Guangdong* 518000, *China*)

Abstract: In order to meet the requirements of grouping product families for advanced manufacturing modes such as mass customization, the features in BOM (Bill of Materials) are comprehensively analyzed, and a concept of BOM structure-based similarity metric model, a content-based similarity metric model, and an ensemble model combined with both are proposed. In the structure-based model, BOMs are represented by adjacent matrixes, including the relationships between materials and the quantity of materials, and the Orthogonal Procrustes Analysis is implemented to measure the similarity among BOMs. While in content-based model, effective text features are extracted from BOMs, being transformed to vectors by TFIDF(Term Frequency-Inverse Document Frequency), and finally being inputted into cosine approximation formula for similarity value. To obtain more accuracy and performance, a weight distribution method based on the Gini coefficient is proposed for the ensemble model. Finally, a test framework is provided and all models are in evaluated experimentally in accuracy and performance.

Key words: similarity metric; BOM(Bill of Materials); product family; ensemble model

1 引言

产品近似度量问题是产品族划分的分支,是工业界亟待解决的问题之一. 研究表明,BOM[1~7]是刻画产品特征的信息源头和划分产品类别的重要依据. 部分文献[3~5]认为BOM是无序的,采用无序树和相邻矩阵来描述和存储BOM及其结构[7],在此之上引入一系列诸如奇异分解的数学办法获取BOM间的近似度[5]. 但是,仅依据BOM的几何结构来度量两个BOM间的近似程度,忽略了物料名称、型号、性能、评价指标等诸多因素,难以有效划分产品族. 为此,文献[4~7]分别在BOM结构信息的基础上引入个人经验、使用频率以及生命周期等其它因素,形成不同的解决办法. 但从BOM本身来看,个人经验过于主观,子BOM的使用频率或产品生命周期因素均属于BOM的衍生属性,不能有效表征BOM的特性,若抛开应用场景直接使用,易导致近似

收稿日期:2018-08-22;修回日期:2019-01-02;责任编辑:马兰英

度量值出现偏差.

针对于此,本文不仅考虑 BOM 结构,还考虑了物料数量、物料型号、物料描述等诸多特征,用平衡的视角看待结构和其他特征信息,提出内容近似度量模型,同时将仅考虑结构近似的已有办法[5,7]概括为结构近似度量模型,然后在两种模型的基础上提出组合两者的集成模型. 结构近似方面,本文以文献[5]为蓝本,主要考虑的特征是 BOM 的结构和物料数量. 内容近似方面,基于 BOM 的文本进行特征提取,引入逆向词频法[8]对文本进行处理,并在此之上完成模型的相似度计算. 集成模型综合了结构模型和内容模型,在考虑 BOM 特征方面更为全面. 为获得更好的模型性能,本文在集成权重的分配问题上提出比等权重更有优势的基于基尼系数的分配方法. 最后,本文提供框架并以实验对算法进行评价. 实验结果显示,本文集成模型在不降低效率的情况下,在模型性能上较已有办法更有优势.

2 实体解释及问题定义

BOM 内含零部件、半成品和成品三种实体. 本节对三种实体在不同模型中的体现作出解释并给出本文问题的定义.

BOM 树和树节点:BOM 树的叶子节点、中间节点和根节点分别对应零部件、半成品和成品. 每个节点包含标签 *lbl* 和该节点相应物料数量 *qty*. 例如,表 1 是一份原始的 BOM,其对应的 BOM 无序树如图 1(*a*)所示.

表 1 产品 A 的 BOM

标签	项目	物料号	描述	数量	单位	策略
B	L	3145436	主板机插组件 OCZ-ZT550W-2 REV:001 A T26	1	EA	2
C	L	3146728	小板机插组件 OCZ-ZT700W-1 REV:009 E T26	1	EA	2
D	L	3153354	散热片组件 12-326	1	EA	2
E	L	1224220	电阻 MF 510K 1/10W F0603	2	EA	2
F	L	1221725	电感 CI ZL-0420-008 T DIP TIMO	8	EA	2
G	L	1221432	电感 CI ZL-0420-008 T DIP NEO	3	G	2
I	L	1221361	绝缘粒 6.3 * 3 * 4.55 白色 UL94V0 BN	2	EA	2
J	L	1220766	三极管 KSC2383OTA 160V1A TO92L FSC	2	EA	2
K	L	1220709	三极管 MMBT2907A 50V0.6A SOT23 TC	2	EA	2
M	L	1220708	思维特电子灌封胶 JG-910-B 胶白 GP	8.5	G	2
O	L	1217809	锡条 Sn 99.7 Ag0.3 RJ GP	1.4	G	2
P	L	1217453	锡丝 Sn 0.3Ag 0.7Cu 1.2 GP	2	G	2
R	L	1217291	磁珠 3 * 1.2 * 3 BULK MX	2	EA	2
S	L	1216727	三刀卡 ESP 150W V54 180℃B = C 550 * 110 GP	2	EA	2
T	L	1216481	螺钉 十字盘头组合 GB90748-88 M3 * 7 ROHS	2	EA	2

表 2 产品 A"的 BOM

标签	项目	物料号	描述	数量	单位	策略
B	L	3145987	主板机插组件 GA150S REV:S2 B T26	1	EA	2
C	L	3176584	小板机插组件 GW-4000(85 +)-2 V01 B T26	1	EA	2
D	L	3153678	散热片组件 22-048	1	EA	2
E	L	1234229	电阻 TR 1kΩ J 1/8W 0805 REEL	2	EA	2
F	L	1231726	电感 MA BL-1205A-015 T DIP FS	8	EA	2
G	L	1231452	电感 MA BL-1205A-015 T DIP FS	3	G	2
I	L	1231351	绝缘粒 6.3 * 3 * 4.6 白色 UL94V0	2	EA	2
J	L	1220776	三极管 KTB772-Y 30V-3A TO126 KEC	2	EA	2
K	L	1220739	三极管 2SC2235-Y 120V-0.8A TO92L TSB	2	EA	2
M	L	1220708	思维特电子灌封胶 JG-910-B 胶白 GP	8.5	G	2
O	L	1227809	锡条 Sn 99.7 Ag0.3 RJ GP	1.4	G	2
P	L	1227453	锡丝 Sn 0.3Ag 0.7Cu 1.2 GP	2	G	2
R	L	1227291	磁珠 3 * 1.2 * 3 BULK QHS	2	EA	2
S	L	1216727	三刀卡 ESP 150W V54 180℃ B = C 550 * 110 GP	2	EA	2
T	L	1216491	螺钉 十字槽 K 头 M3 * 11 M 削尾 镀镍	2	EA	2

BOM 树的相邻矩阵:BOM 树的存储结构是相邻矩阵 $\boldsymbol{M}$,矩阵中的数字表示物料用量 *qty*. 图 1(*a*)对应的

相邻矩阵如图 2 所示.

BOM 词:表示为 *wrd*. *wrd* 表示 BOM 中任意一个以空格作为分词标志的词. 比如,表 1“描述”列的第一行包含“主板机插组件”、“OCZ-ZT550W-2”等五个词.

BOM 向量:表示为 ***BOMV***. ***BOMV*** = $\{f(wrd_1),f(wrd_2),\cdots,f(wrd_n)\}$是一系列用于刻画 BOM 特征的词通过函数 $y=f(x)$变换后形成的向量. 例如,用 BOM 向量表示表 1 时,有 ***BOMV*** = $\{f(\text{“}L\text{”}),f(\text{“}20\text{”}),\cdots,f(\text{“}M3*7\text{”})\}$.

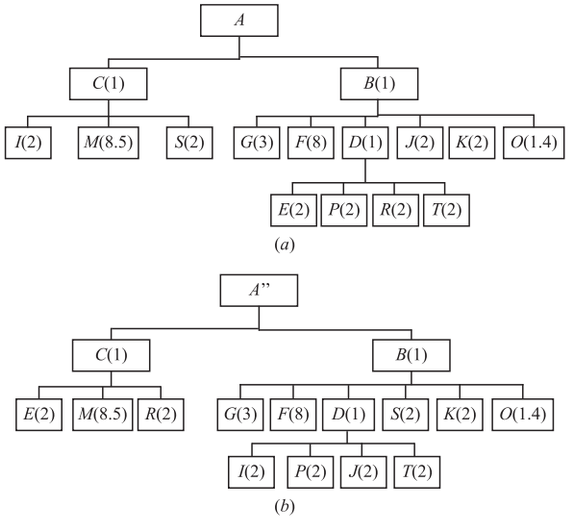

图1　产品A和A”的无序树

问题定义:给定两个产品 Pro_1 和 Pro_2,产品的 BOM 分别是 BOM_{pro1} 和 BOM_{pro2},求 BOM_{pro1} 和 BOM_{pro2} 的近似程度 *Sim*.

	A	*C*	*B*	*I*	*M*	*S*	*G*	*F*	*D*	*J*	*K*	*O*	*E*	*P*	*R*	*T*
A	**1**	**1**	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	**1**	0	**2**	**8.5**	**2**	0	0	0	0	0	0	0	0	0	0
B	0	0	**1**	0	0	0	**3**	**8**	**1**	**2**	**2**	**1.4**	0	0	0	0
I	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	**1**	**2**	0	0	**2**	**2**	0	**2**
J	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**

图2　产品A的相邻矩阵

3　度量模型

3.1　结构近似度量模型

根据文献[4]的结论,结构近似度量模型采用无序树描述 BOM 结构并将计算无序树之间的近似程度转换成计算树相邻矩阵之间的相似度. 两个相邻矩阵 $\boldsymbol{M}_X$ 和 $\boldsymbol{M}_Y$的近似程度被定义为:

$$Sim_{struct}=\frac{\{trace\{\boldsymbol{M}'_X\boldsymbol{M}_YT\}\}^2}{trace\{(\boldsymbol{M}_YT)'(\boldsymbol{M}_YT)\}trace\{\boldsymbol{M}'_X\boldsymbol{M}_X\}} \quad (1)$$

式(1)右边的 $\boldsymbol{M}'_X$、$(\boldsymbol{M}_YT)'$分别是 $\boldsymbol{M}_X$和 $\boldsymbol{M}_YT$ 的转置矩阵,$trace\{A\}$表示矩阵 $\boldsymbol{A}$ 的对角线元素之和. 其中,矩阵 $\boldsymbol{M}_X$、矩阵 $\boldsymbol{M}_Y$已知,矩阵 $\boldsymbol{T}$ 来自于 $\boldsymbol{M}_X$和 $\boldsymbol{M}_Y$的正交普氏分析的结果. 正交普氏分析过程中,$\boldsymbol{M}_X$是基准矩阵,$\boldsymbol{M}_Y$是用于匹配的矩阵,$\boldsymbol{T}$ 是自转正交矩阵. $\boldsymbol{T}$ 的求解过程如公式(2),(3),(4)所示.

$$\begin{aligned}L&=\|\boldsymbol{M}_X-\boldsymbol{M}_Y\boldsymbol{T}\|\\&=trace\{(\boldsymbol{M}_X-\boldsymbol{M}_Y\boldsymbol{T})'(\boldsymbol{M}_X-\boldsymbol{M}_Y\boldsymbol{T})\}\\&=\min \quad (2)\end{aligned}$$

$$\boldsymbol{T}'\boldsymbol{T}=1 \quad (3)$$

$$\boldsymbol{T}=u*v \quad (4)$$

$$[u,\Sigma,v]=SVD(\boldsymbol{M}'_Y\boldsymbol{M}_X) \quad (5)$$

式(2)隐含欧几里德范数 $\|A\|=trace\{A'A\}=\sum_i^m\sum_j^m a_{ij}^2$,采用经典的最小二乘法来求出使 L 取得最小值的 $\boldsymbol{T}$ 矩阵. 式(3)把问题限定到正交普氏分析的范畴. 式(4)的 u 和 v 来自于矩阵 $\boldsymbol{M}'_Y\boldsymbol{M}_X$的奇异值分解(式(5))[9].

综上所述,公式(1)到公式(5)是一个自上而下的推到过程. 因此,逆向代入数值则可求解两个 BOM 的 Sim_{struct},图 1(a)和图 1(b)的近似程度结果如下:

$$\begin{aligned}Sim_{struct}&=\frac{\{trace\{\boldsymbol{M}'_A\boldsymbol{M}''_A\boldsymbol{T}\}\}^2}{trace\{(\boldsymbol{M}''_A\boldsymbol{T})'(\boldsymbol{M}''_A\boldsymbol{T})\}trace\{\boldsymbol{M}'_A\boldsymbol{M}_A\}}\\&=\frac{(197.54)^2}{198.21*198.21}=0.9933\end{aligned}$$

3.2　内容近似度量模型

假设二维空间中有两个 BOM,分别是由 $x_{wrd11}=f(wrd_{11})$和 $y_{wrd12}=f(wrd_{12})$组成的 $\boldsymbol{BOMV}_1=(x_{wrd11},y_{wrd12})$和由 $x_{wrd21}=f(wrd_{21})$和 $y_{wrd22}=f(wrd_{22})$组成的 $\boldsymbol{BOMV}_2=(x_{wrd22},y_{wrd22})$,那么引入向量空间余弦相似度[10]后,两者的近似度量公式是:

$$\begin{aligned}Sim_{content}&=\cos\theta(\boldsymbol{BOMV}_1,\boldsymbol{BOMV}_2)\\&=\frac{x_{wrd11}*x_{wrd21}+y_{wrd12}*y_{wrd22}}{\sqrt{x_{wrd11}^2+y_{wrd12}^2}+\sqrt{x_{wrd21}^2+y_{wrd22}^2}} \quad (6)\end{aligned}$$

然而,物料清单 BOM 包含了物料的名称、型号、数量、评价指标、性能等多个维度的信息,即上述的 ***BOMV*** 必然是一个包含 n 维词汇信息的向量 $\boldsymbol{BOMV}_1(x_{wrd11},x_{wrd12},x_{wrd13},\cdots,x_{wrd1n})$和 $\boldsymbol{BOMV}_2(y_{wrd11},y_{wrd12},y_{wrd13},\cdots,y_{wrd1n})$. 因此,本文对式(6)进行推广,有:

$$\begin{aligned}Sim_{content}&=\cos\theta(\boldsymbol{BOMV}_1,\boldsymbol{BOMV}_2)\\&=\frac{\sum_{k=1}^n x_{wrd1k}y_{wrd1k}}{\sqrt{\sum_{k=1}^n x_{wrd1k}^2}+\sqrt{\sum_{k=1}^n y_{wrd1k}^2}} \quad (7)\end{aligned}$$

其中,BOM 词 wrd_i到向量元素 x_{wrdi}的转换过程 $f(wrd_i)$

采用修订后的“词频-逆向文件频率”法—BOM 词频-逆向文件频率法[11],公式如下:

$$f(wrd) = TFIDF(wrd, d_{\text{BOM}}, D) = TF(wrd, d_{\text{BOM}}) * IDF(wrd, D) \quad (8)$$

$$IDF(wrd, D) = \log \frac{|D| + 1}{DF(wrd) + 1} \quad (9)$$

这里,$TF(wrd, d_{\text{BOM}})$是 wrd 在文档 d_{BOM}中的出现次数,$DF(wrd, d_{\text{BOM}})$指包含了词汇 wrd 的物料清单数,$|D|$是语料库中的文档总数.

若将表 1 和表 2 的 BOM 词的分别代入式(8)和式(9),有:

$$f(\text{"OCZ-ZT550W-2"}) = TFIDF(\text{"OCZ-ZT550W-2"}, d_{BOM_A}, D) = 1 * \log \frac{3+1}{1+1} = 1$$

$$\vdots$$

$$f(\text{"ROHS"}) = TFIDF(\text{"ROHS"}, d_{BOM_B}, D) = 4 * \log \frac{3+1}{2+1} = 0.4998$$

综上所述,$\boldsymbol{BOM}_A$ 和 $\boldsymbol{BOM}_A$" 对应的向量如表 4 所示.

表 3 $\boldsymbol{BOM}_A$和 $\boldsymbol{BOM}''_A$向量

	L	20	……	OCZ-ZT550W-2	……	T26	……	ROHS	$M3*7$
$\boldsymbol{BOM}_A$	0	1	……	1	……	0.2500	……	0.1249	0.1249
$\boldsymbol{BOM}_A$"	0	0	……	0	……	0.2500	……	0	0

所以,有:

$$Sim_{content} = \cos\theta(\boldsymbol{BOM}_A, \boldsymbol{BOM}_A'') = \frac{0.4902}{2.2608 + 2.1525} = 0.1111$$

3.3 集成模型

根据文献[11]的解释,本节集成了 BOM 结构近似度量模型和内容近似度量模型. 集成办法分别有:(1)等权重集成(Equal Weighted Ensemble,ENS1);(2)基于基尼系数集成(Performance Based Ensemble,ENS2). 等权重集法定义如下:给定结构近似模型的近似度 Sim_{struct}和内容近似模型的近似度 $Sim_{content}$,$ENS1$ 定义如下:

$$Sim_{ens1} = w_{eq1} * Sim_{struct} + w_{eq2} * Sim_{content} \quad (10)$$

其中,$w_{eq1} = w_{eq2} = 1/2$.

基于基尼系数[10]的权重决策策略定义如下:

$$Sim_{ens2} = w_{gini1} * Sim_{struct} + w_{gini2} * Sim_{content} \quad (11)$$

$$w_{gini} = g_s/(g_s + g_c) \quad (12)$$

$$g = \sum_{i=1}^{n} gini_i \quad (13)$$

其中,$i, j \in [1, m]$,g 表示同一个模型下多个基尼系数之和,下标 c 和 s 用于区分内容度量模型和结构度量模型,$gini$ 是基尼系数. 假设有 K 个类别,第 k 个类别的发生概率为 p_k,那么式(14)成立. 然而,近似度量模型不具备分类能力. 为此,本文引入 K-Means 聚类算法来确定 g_i值,继而求出两个模型的集成比例.

$$gini = \sum_{k=1}^{K} p_k(1 - p_k) = 1 - \sum_{k}^{K} p_i^2 \quad (14)$$

因此,A 和 A''在两种集成模型中的近似程度为:

$$Sim_{ens1} = \frac{1}{2} * Sim_{struct} + \frac{1}{2} * Sim_{content} = 0.5 * 0.9933 + 0.5 * 0.1111 = 0.5522$$

$$Sim_{ens2} = 0.4107 * Sim_{struct} + 0.5839 * Sim_{content} = 0.4107 * 0.9933 + 0.5839 * 0.1111 = 0.4734$$

4 实验框架

本节实现了模型本身的近似度量办法以及构建在该近似度量办法之上的分类算法,分别如算法 1 和算法 2 所示.

算法 1 frm-Similarity($\boldsymbol{M}$,BOM)

输入:相邻矩阵 $\boldsymbol{M}$ 集,BOM 文本集
输出:$SimDF['Sim_{struct}', 'Sim_{content}', 'Sim_{ens1}', 'Sim_{ens2}']$ //DataFrame 格式
开始
1. for all $M_i, M_j \in \boldsymbol{M}$ do
2. $MM_k \leftarrow M_i * M_j$
3. $[u_k, sigma_k, v_k] \leftarrow SVD(MM_k)$
4. $T_k \leftarrow u_k * v_k$
5. $Sim_{struct_i} = trace\{M_s' M_A'' T\}^2/(trace\{(M_A'' T)'(M_A'' T)\} trace\{M_s' M_s\})$
6. end for all
7. $SimDF['Sim_{struct}'] = Sim_{struct}$
8. BOMs = $Call\ Model_{TFIDF}$ //调用 TFIDF 模型并返回 BOM 文本向量集 BOMs
9. for all $\boldsymbol{BOM}_i, \boldsymbol{BOM}_j \in$ BOMs do
10. $Sim_{content_k} = \cos(BOM_i, BOM_j)$
11. end for all
12. $SimDF['Sim_{content}'] = Sim_{content}$
13. $SimDF['Sim_{ens1}'] = 0.5 * SimDF['Sim_{struct}'] + 0.5 * SimDF['Sim_{content}']$
14. $SimDF['Sim_{ens2}'] = 0.3891 * SimDF['Sim_{struct}'] + 0.6109 * SimDF['Sim_{content}']$
15. Return $SimDF$
结束

算法 2　frm-ModelCapability (SimDF,k)

输入:近似集 *SimDF*,可能存在的类别数 k
输出:*Score*　//模型得分
开始
1. for all $Sim \in SimDF$ do　//*Sim* 是 *SimDF* 中的列
2.　　初始化常数 k
3.　　随机选取初始点为质心
4.　　while 质心不再改变 do
5.　　以 *Sim* 作为距离,扫描样本与每个质心之间的数值,将样本归类到最相似的类中
6.　　重新计算质心
7.　　end while
8. End for all
9. Return K-Means. Score　// Score 中包含模型的 *F*1 值、Recall 值和 Precision 值

算法 1 是近似度量模型的训练过程. 第 1 ~ 7 行属于结构近似模型的部分. 第 8 ~ 12 行是内容近似的计算过程. 13 行和 14 行分别是等权重集成模型和基于基尼系数集成模型. 算法 2 实现了 K-Means 聚类算法,是四个模型的性能测试框架.

5　实验

5.1　实验设置

实验设置的第一个参数是实验数据,包含 3 个数据集,均来自于 SAP 系统. 数据集的统计信息如表 4 所示.

表 4　数据集的统计信息

数据集	来源	BOM 数	数据形式	带标签
Data1	工厂 1	500	相邻矩阵和文本向量	是
Data2	工厂 2	500	相邻矩阵和文本向量	是
Data3	多个工厂	500	相邻矩阵和文本向量	是
Data4	多个工厂	5000	相邻矩阵和文本向量	是

第二个参数是实验对象,分别是结构近似度量模型 Struct-Baseline、内容近似度量模型 Content、等权重集成模型 Ens1 和基于基尼系数集成模型 Ens2. 其中,结构近似度量模型是同一研究领域中的比较对象.

第三个参数是评价指标,含 *F*1、召回率和精确率(Precision). 具体如表 5 所示:

表 5　评价指标

数据集	正样本	负样本	
分类正确	TP	FP	$Precision = TP/(TP + FP)$
分类错误	TN	FN	
	$Recall = TP/(TP + TN)$		$F1 = 2TP/(2TP + FP + TN)$

5.2　模型性能对比实验

本节实验分别在 Data1、Data2、Data3 三个数据集上用四种不同的模型调用算法 2,并得出如图 3 所示的实验结果.

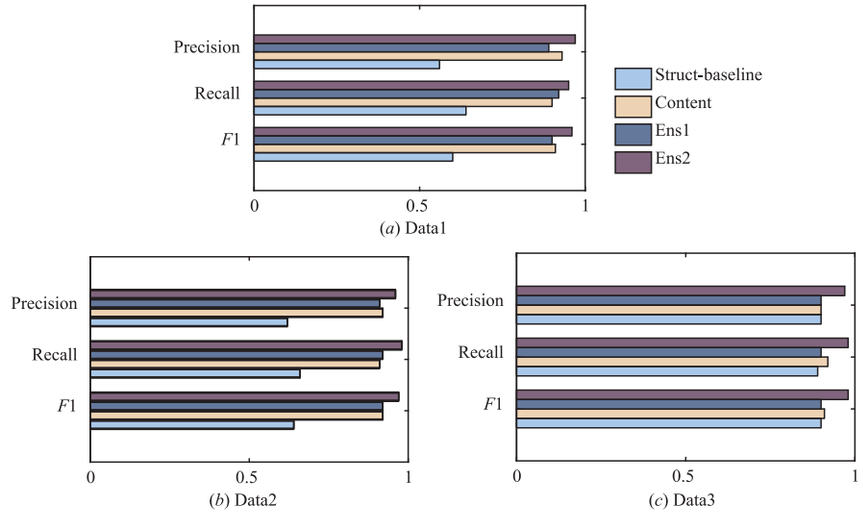

图3　模型性能对比

模型稳定性. 图 3 显示,Struct-Baseline 模型在不同数据集间波动幅度较大,其他三种模型相对稳定. 究其原因, Data3 在数据集内部 BOM 与 BOM 结构上差异较 Data1 和 Data2 的更明显,即 Stuct-baseline 更倾向于区分结构差异显著的 BOM. 因此,仅考虑 BOM 结构而忽略其他因素难以有效划分产品族.

模型性能比较. 三种指标的测试结果在图 3(*a*)、3(*b*)和 3(*c*)中显示,Ens2 整体最优,Ens1 和 Content 不相上下,Struct-Baseline 最差. 显然,本文基于基尼系数的权重分配方案较等权重方法更有效.

5.3 模型训练效率对比实验

本节从算法 1 中分别抽取 Struct-baseline 模型和 Content 模型的部分单独训练. 实验所用的数据集是 Data4. 实验结果如图 4 所示,X 轴表示不断增大的 BOM 个数(单位:个),Y 轴是模型运行时间(单位:s).

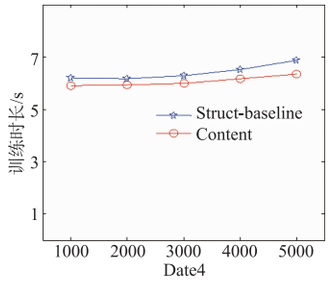

图4 耗时对比实验

图 4 中,Content 相对 Struct-baseline 模型更为高效. 分析两者的 I/O 和数学计算复杂度发现,Struct-baseline 的耗时分布均衡,Content 的更偏向于 I/O 消耗. 这反映出 TFIDF 在计算速度上的优势. 同时,本节实验证实,若采用先并行训练单模型再训练集成模型的顺序,集成模型不因在 Sturct-baseline 中引入 Content 模型而带来训练时长的剧增.

6 总结

本文解决的是 BOM 之间的近似度量问题. 创新点在于比已有办法更全面考虑物料清单的特征信息,在提出 BOM 结构近似度量模型概念、内容近似度量模型的基础上提出组合两者的集成模型. 目前,本文的解决办法被应用于智能制造信息系统的推荐功能中,为现有工业信息系统提供新的搜索功能—近似查找.

参考文献

[1] Hu X, Peng W, Jin L, Dou J, Zhong Y, Jiang R. A new product family mining method based on PLM database [J]. Journal of Central South University, 2017, 24(11): 2513 – 2523.

[2] Romanowski C J, Nagi R. On comparing bills of materials: a similarity/distance measure for unordered trees [J]. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2015, 35(2): 249 – 260.

[3] ZHU H, WANG H, ZHANG G. General bill of material reconfiguration method based on data mining [J]. Computer Integreted Manufacturing Systems. 2008, 14(2): 315 – 321.

[4] Geng J, Zhen M, Tian X, Zhang D. Product lifecycle-oriented BOM similarity metric method [J]. China Mechanical Engineering, 2008, 19(20): 2441 – 2445.

[5] Shih H M. Product structure (BOM)-based product similarity measures using orthogonal procrustes approach [J]. Computers & Industrial Engineering, 2013, 61(3): 608 – 628.

[6] Israt J C, Richi N. Identifying product families using data mining techniques in manufacturing [A]. Paradigm. Proceedings of the Twelfth Australasian Data Mining Conference [C]. Berlin: Springer's Communication in Computer and Information Science, 2014. 113 – 120.

[7] Chowdhury I J, Naya R. A novel method for finding similarities between unordered trees using matrix data model [A]. Web Information Systems Engineering-WISE [C]. Berlin: Springer's Communication in Computer and Information Science, 2013. 421 – 430.

[8] Chen C. Improved TFIDF in big news retrieval: An empirical study [J]. Pattern Recognition Letters. 2017, 93(1): 113 – 122.

[9] Shuo C, Rongxing L, Jie Z. A flexible privacy-preserving framework for singular value decomposition under internet of things environment [A]. IFIP International Conference on Trust Management [C]. Berlin: Springer, 2017. 21 – 37.

[10] Geng Z, Li Y, Han Y, Zhu Q. A novel self-organizing cosine similarity learning network [J]. Energy, 2018: 142(1): 400 – 410.

[11] Caitlin M A, Elizabeth A W, Brian G. Prediction of plant lncRNA by ensemble machine learning classifiers [J]. BMC Genomics, 2018, 19(1): 1 – 11.

作者简介

吴文李 女,1986 年生. 中国科学院深圳先进技术研究所和中国长城科技集团股份有限公司博士后实践基地联合招收博士后. 主要研究方向为机器学习、数据挖掘.
E-mail: huihuigou@whu.edu.cn

范小朋(通讯作者) 男,中国科学院深圳先进技术研究院副教授、硕士研究生导师. 主要研究方向涵盖分布式数据处理、大数据分析、移动云计算、无线网络、移动计算.
E-mail: xp.fan@siat.ac.cn