业务流程模型抽象中最优子流程数的确定

孙善武 王 楠
(吉林财经大学管理科学与信息工程学院 长春 130117)
(吉林财经大学物流产业经济与智能物流吉林省重点实验室 长春 130117)
(吉林财经大学吉林省互联网金融重点实验室 长春 130117)

摘 要 根据业务流程模型的特征,基于笔者前期工作中给出的两个不同约束条件下的受限 k-means 行为聚类算法,提出确定最优子流程数的方法。基于对流程结构的假设,同时结合行为语义的经验阈值限定,给出了确定子流程数恰当上限值的方法,以达到减少循环次数的目的。根据 k 值的变化,分别基于子流程结构紧密性特征和流程结构树,在循环过程中设计增量式方法,对簇中心进行简便的递增;设计合理的有效性指标,对抽象结果模型进行评估,进而生成最佳子流程数;利用真实的流程模型库对设计的方法进行实验验证,得到的最优子流程数与人工设计的结果非常接近。

关键词 业务流程模型抽象,最优子流程数,行为文档
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.10.044

Determining Optimal Number of Subprocesses in Business Process Model Abstraction

SUN Shan-wu WANG Nan
(College of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117, China)
(Laboratory of Logistics Industry Economy and Intelligent Logistics, Jilin University of Finance and Economics, Changchun 130117, China)
(Jilin Province Key Laboratory of Internet Finance, Jilin University of Finance and Economics, Changchun 130117, China)

Abstract According to the characteristics of the business process model, this paper proposed a method to determine the optimal number of subprocesses based on the k-means activity clustering algorithm with two different constraints given in the previous work. Combining the assumption for the process structure with the threshold restriction of activity semantics, the method of determining the appropriate upper bound of the number of subprocesses is given in order to reduce the number of iterations. According to the change of k value, based on the characteristics of structural compactness of the subprocesses and the refined process structure tree, an incremental approach is designed to simplify the incremental of the cluster centers. A reasonable index is designed to evaluate the abstract result model, and then the optimal number of subprocesses is generated. The proposed method is applied to a process model repository in use, and the number of the optimal subprocesses is very close to the result given by the modelers involved.

Keywords Business process model abstraction, Optimal number of subprocesses, Behavioral profiles

1 引言

众多文献提出了业务流程模型抽象(Business Process Model Abstraction, BPMA)中的行为聚类方法[1-2],类似现有的绝大部分聚类算法通常需要事先给定聚类数,已有的行为聚类方法也都假设可以事先确定子流程数 k。但在实际应用中,子流程数 k 很难根据建模者的经验来准确获得。

当聚类数目未知时,如何确定数据集的聚类数目是聚类分析研究中的一个基础性难题[3]。通常采用聚类算法和内部有效性指标相结合的方式,使用一种迭代的"trial-and-error"过程,通过设定不同的聚类数条件来运行聚类算法,采用内部有效性指标来评估多次聚类结果的质量,以确定数据集的最佳聚类数,具体过程如图1[4]所示。

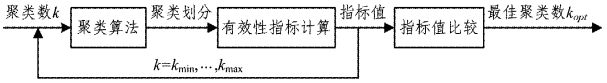

图1 数据集的最佳聚类数确定方法示意图

图1给出的最佳聚类数的确定过程为:给定 k 的权值范围$[k_{min},k_{max}]$,对数据集使用不同的聚类数 k 运行同一聚类算法从而得到一系列聚类结果,计算每个结果的有效性指标值,并对其进行比较,将对应最佳指标值的聚类数 k_i 作为最佳聚类数 k_{opt}。

本文根据业务流程模型的特征,基于笔者前期工作中给出的两个不同约束条件下的受限 k-means 行为聚类算法[5-6],

到稿日期:2016-09-26 返修日期:2017-03-21 本文受国家自然科学基金(61402193,61702213),吉林省教育厅"十三五"科学技术研究项目(2016105),吉林省教育科学"十二五/十三五"规划课题(GH150285,GH16249)资助。
孙善武(1969-),男,硕士,副教授,主要研究方向为建模与抽象、传感器网络、网络安全;王 楠(1980-),女,博士,副教授,主要研究方向为业务流程模型抽象、基于模型的诊断、自动推理,E-mail:ctuwangnan@126.com。

提出确定最优子流程数的方法。本文第 2 节引入保序业务流程模型及流程结构分解的相关概念;第 3 节给出了子流程数搜索范围的求解方法;第 4 节基于笔者前期工作中设计的两种求解初始簇中心的方法,提出了求解最优子流程数的循环过程中初始簇中心的增量式确定方法;第 5 节设计了流程抽象结果模型的新的有效性指标;第 6 节给出确定最佳子流程数的算法并用真实的流程模型库对实验结果进行分析;最后总结全文。

2 相关定义

本节引入文献[7]中的一些定义以便后续章节使用,首先给出保序业务流程模型的相关概念。

定义 1(业务流程模型)[7] 称元组 $PM=(A,G,F,t,s,e)$ 为一个业务流程模型。其中,A 为行为的有限非空集合;G 为 *gateway* 的有限集合;$N=A\cup G$ 为节点的有限集合,且 $A\cap G=\emptyset$;$F\subseteq N\times N$ 表示流关系,使得 (N,F) 是一个连接图;每个行为至多有一个入边,至多有一个出边;s 是唯一没有入边的行为,即起始行为,e 是唯一没有出边的行为,即终止行为;$t:N_G\rightarrow\{and,xor\}$ 是为每个 *gateway* 分配控制流构件的函数,每个 *gateway* 表示 *split* 或 *join*,其中 *splits* 只有一个入边和至少两个出边,*joins* 有至少两个入边和一个出边。

定义 2(弱序关系)[7] 令 $PM=(A,G,F,t,s,e)$ 为一个流程模型,$\mathcal{T}_{PM}$ 是其轨迹集合。弱序关系 $\succ_{PM}\subseteq(A\times A)$ 包含所有这样的行为对 (a,b):在 $\mathcal{T}_{PM}$ 中存在一个轨迹 $\sigma=n_1,\cdots,n_l$,使得 $j\in\{1,\cdots,l-1\}$,$j<k\leqslant l$,$n_j=a$ 且 $n_k=b$ 成立。

定义 3(行为文档)[7] 令 $PM=(A,G,F,t,s,e)$ 为一个流程模型,行为对 $(a,b)\in(A\times A)$ 是以下关系之一:

1) *strict* 顺序关系 $\rightsquigarrow_{PM}$,如果 $a\succ_{PM}b$ 且 $a\not\succ_{PM}b$;

2) *exclusiveness* 关系 $+_{PM}$,如果 $a\not\succ_{PM}b$ 且 $b\not\succ_{PM}a$;

3) *interleaving* 顺序关系 $\|_{PM}$,如果 $a\succ_{PM}b$ 且 $b\succ_{PM}a$。

所有这 3 种关系 $BP=\{\rightsquigarrow_{PM},+_{PM},\|_{PM}\}$ 的集合即为 PM 的行为文档。$a\not\succ_{PM}b$ 表示从 a 到 b 没有弱序关系。行为文档中的 3 种关系与 *strict* 顺序的逆关系 $\rightsquigarrow^{-1}=\{(a,b)\in(A\times A)\mid(b,a)\in\rightsquigarrow\}$ 共同划分行为集合的笛卡尔积。

定义 4(聚合函数 aggregate)[7] 令 $PM=(A,G,F,t,s,e)$ 为一个流程模型,$PM_a=(A_a,G_a,F_a,t_a,s_a,e_a)$ 为 PM 对应的抽象模型。函数 $aggregate:A_a\rightarrow(P(A)\setminus\emptyset)$ 确定了 PM_a 中的一个行为与 PM 中的行为集合之间的对应关系。

定义 5(保序的业务流程模型抽象)[7] 令 $PM=(A,G,F,t,s,e)$ 为一个流程模型,业务流程模型抽象 α 将 PM 映射到 $PM_a=(A_a,G_a,F_a,t_a,s_a,e_a)$,即 $\alpha:(PM,activity\ groups)\rightarrow PM_a$,$PM_a$ 中的行为是抽象对象。令函数 *aggregate* 用于建立 PM 与 PM_a 中的行为之间的关联关系,则运算 α 是保序业务流程模型抽象,当且仅当对于 $\forall x,y\in A_a$,$x\neq y$,$\forall a,b\in A$,$a\in aggregate(x)$ 且 $b\in aggregate(y)$,有以下结果成立:1) $a\rightsquigarrow_{PM}b\Rightarrow x\rightsquigarrow_{PM_a}y$;2) $a\rightsquigarrow^{-1}_{PM}b\Rightarrow x\rightsquigarrow^{-1}_{PM_a}y$;3) $a+_{PM}b\Rightarrow x+_{PM_a}y$;4) $a\|_{PM}b\Rightarrow x\|_{PM_a}y$。

接下来介绍具有单入节点和单出节点的流程片段分解,即 RPST(the Refined Process Structure Tree)分解。根据文献[7],在流程建模上下文中,流程片段可以作为“自包含”的流程部分。由于这种片段只有一个单入节点和一个单出节点,因此可以从结构上将其独立为一个子流程,且这种分解是唯一的。我们能够在相对于流程模型中包含节点数的线性时间内构造 RPST,详见文献[8]。

定义 6(流程片段)[7] 令 $PM=(A,G,F,t,s,e)$ 为一个流程模型。流程模型 PM 的片段是一个元组 $f=(A_f,G_f,F_f,t_f)$,其中 $(A_f\cup G_f,F_f)$ 是图 $(A\cup G,F)$ 的一个连接子图,函数 t_f 是对 PM 中的 t 的约束,使得生成 G_f。

定义 7(边界节点)[7] 令 $PM=(A,G,F,t,s,e)$ 为一个流程模型,其包含一个流程模型片段 $PMF=(A_{PMF},G_{PMF},F_{PMF},t_{PMF})$。节点 $n\in N_{PMF}$ 是 PMF 的一个边界节点,如果 $\exists e\in in(n)\cup out(n)$,则函数 $in(n)$ 和 $out(n)$ 分别表示节点 n 的入边集合和出边集合。如果 n 是一个边界节点,那么若 $in(n)\cap F_{PMF}=\emptyset$,则 n 是 PMF 的一个入口。一个节点 n 是 PMF 的一个出口,如果它是 PMF 的边界节点并且 $out(n)\cap F_{PMF}=\emptyset$。

定义 8(部件)[7] 令 $PM=(A,G,F,t,s,e)$ 为一个流程模型,其包含一个流程模型片段 $PMF=(A_{PMF},G_{PMF},F_{PMF},t_{PMF})$。片段 PMF 是一个部件,如果它恰好具有两个边界节点:一个入口节点和一个出口节点。

令 F 是流程模型 PM 中的所有部件集合。

定义 9(规范部件)[7] 一个部件 $PMF=(A_{PMF},G_{PMF},F_{PMF},t_{PMF})$ 是规范的,如果 $\forall PMF'\in F:PMF\neq PMF'\Rightarrow(F_{PMF}\cap F_{PMF'}=\emptyset\vee(F_{PMF}\subset F_{PMF'})\vee(F_{PMF'}\subset F_{PMF}))$。

定义 10(RPST)[7] 令 $PM=(A,G,F,t,s,e)$ 为一个流程模型。流程模型 PM 的 $RPST$ 是一个树状图 $RPST_{PM}=(\Omega,r,\chi)$,使得:1)$\Omega$ 是 PM 中的所有规范部件集合;2)R 是树中根节点对应的部件;3)$\chi\subseteq\Omega\times\Omega$ 是部件和其孩子部件之间的关系。

3 子流程数的搜索范围

确定聚类数的搜索范围 $[k_{\min},k_{\max}]$ 等同于是确定 $k_{\min}$ 和 $k_{\max}$。其中 $k_{\min}=1$ 指样本均匀分布且无明显差异,通常聚类数最小取 2,即 $k_{\min}=2$,本文也将子流程数的搜索下限设为 2。对于如何确定聚类数的上限(即 $k_{\max}$),目前仍然没有明确的理论指导,很多学者认为可以使用经验规则:$k_{\max}\leqslant\sqrt{n}$。文献[9]对此进行了说明,该结论是以不确定性函数 $f(x)=x^{-1}$ 为前提的,此前提不是充分条件;文献[10]对此进行了证明,其结论是以样本空间具有分形几何特征为前提来推导的,结论不具有一般性。另外,文献[9]中所有数据集的样本数和实际类数不具有这样的性质,文献[11]中部分数据集的样本数和实际类数也不具有这样的性质,因此 $k_{\max}\leqslant\sqrt{n}$ 仅仅是一种经验规则,不具有普遍性和一般性。

事实上,在对业务流程模型进行抽象时,聚类的数据集是流程的构成行为,行为之间具有语义相似性。针对该特点,我们在文献[5]和文献[12]中通过引入虚拟文档的概念将每个行为转化为一个多维向量表示,以满足一般数据集的样本类型。但在业务流程中,行为之间除了具有语义相似性之外,还在流程结构的控制流顺序上存在相互制约的关系,因此在确定子流程数的上限值时,仅仅凭借上述研究中对一般数据集提出的经验规则是不可靠的。本文充分利用业务流程的控制

流保序结构特征和子流程内行为的语义相似性特征，使得该上限值更接近同时满足流程的语义和结构要求。

通过对已经包含大量人工设计子流程的流程模型集合进行统计(这里对随机抽取的 150 个模型进行分析，模型数据的来源与文献[5-6]相同)发现，所有模型对应的 RPST 中，同一规范部件中包含的行为基本上都属于同一子流程或者都不属于任何子流程。因此假设："流程模型对应的 RPST 中规范部件属于同一子流程的概率很大"。该假设仅仅从流程结构上依据经验得出了可能构成同一子流程的行为集合，但除了考虑结构之外，还应该考虑同一子流程中行为之间的语义是否足够相似，或者该流程片段中的行为与流程片段中心的距离是否足够小。因此，对于同一规范部件中的所有行为，计算每个行为与该规范部件对应的流程片段中心的相似度，若得到的所有相似度都小于某个阈值 ω，则说明这些行为虽然在结构上隶属于同一个规范部件，有可能属于同一子流程，但是由于语义上不能构成有业务意义的子流程，因此这种情况的规范部件也不计入可能的子流程数上限。

阈值 ω 确定了特征值的下限，如果一个行为对于某个子流程的相似性小于 ω，则表示该行为属于该子流程的可能性较小。阈值 ω 可以由建模者事先根据经验给出，也可以利用真实的流程模型库统计获得。本文为了简化，直接由参与实验的多名工作人员根据经验事先给出 ω，并取其平均值。

根据以上假设和分析，设计如下过程求解 k_{max} 值。

Step1　对于待抽象的业务流程模型 PM，生成其对应的 RPST，T_{PM}。

Step2　循环执行以下步骤：

Step2.1　若 T_{PM} 中的每个规范部件 C 不包含其他行为数大于 1 的规范部件，则计算 C 中的所有行为与 C 对应的簇中心之间的相似度；

Step2.2　若 Step2.1 中计算的所有相似度值都大于 ω，则认为 C 有可能构成一个合理的子流程，因此 k_{max} 值累加 1，并标记 C。

Step3　将 T_{PM} 中所有不包含在被标记的规范部件中的单个行为个数累计入 k_{max}。

Step4　输出 k_{max} 作为待抽象流程模型 PM 的最佳子流程数上限。

4　初始簇中心的增量式确定

本节利用提出的两种生成初始簇中心的方法[5-6]，在生成最佳子流程数的循环过程中，随着 k 值的累加，设计增量式方法实现每次初始簇的确定。

4.1　基于子流程结构紧密性特征的增量式初始簇确定方法

文献[5]介绍了基于子流程结构紧密性特征确定初始簇的方法，该方法选取 k 个聚类中心的基本思想是：取距离尽可能远的对象作为聚类中心，避免了初始选取时可能出现的初始聚类中心过于临近的情况。根据矩阵 D 优先选择两个距离最远的行为作为初始聚类中心，而不是随机选择一个行为。

在求解最佳子流程数的循环过程中，k 值每次累加 1，传统的 k-means 算法求解最佳聚类数时采用随机方法重新生成 k 个初始簇，本文则在生成新的 k 时，采用之前产生的初始簇集合。设所有行为的集合为 A，初始簇的集合为 S，具体方法描述如下。

Step1　初始化 k；//新一轮循环的聚类(子流程)个数

Step2　若 k=2，选择矩阵 D 中最大值对应的两个行为 a^1，a^2，并令 S←{a^1，a^2}，j←2；

Step3　否则，在 A−S 中选择与 S 距离最远的行为 a，S←S+a。

如前文所述，Step3 通过求解如下最优化问题来确定与行为集合 S 距离最远的行为 a^j，即最优函数：$\max\limits_{a^j\in A-S}\min\limits_{a^i\in S}D(a^j,a^i)$。该最优函数表示：求解集合 $A-S$ 中的每一个行为 a^t($1\leqslant t\leqslant|A-S|$，$|A-S|$ 表示集合 $A-S$ 中的行为个数)到 S 中所有行为的最近距离 d_t，则 a^j 是与集合 S 距离最远的行为，当 $d_j=\max\limits_{1\leqslant t\leqslant|A-S|}\{d_t\}$。

4.2　基于 RPST 的增量式初始簇确定方法

文献[6]设计了选择初始簇的方法，该方法基于业务流程模型对应的 RPST，优先选择由多个单个行为或原子部件(这里将行为数为 1 的规范部件称为原子部件)构成的规范部件作为初始簇，其次选择分散的单个行为，最后随机选择剩余单个行为以完成初始簇构造。

本节基于该过程，利用已经生成的初始簇集合 $S=(S_1,\cdots,S_k)(k\geqslant 2)$，在上一次选择第 k 个初始簇时设计一个标志变量 Tag 进行标记，以便选择第 $k+1$ 个初始簇时直接进入到相应的顺序步骤。初始时 $Tag=1$，表示从 Step1 顺序选择。设已有初始簇集合 $S=(S_1,\cdots,S_k)(k\geqslant 2)$，在 $k=k+1$ 时进入到下一轮选择初始簇集合，执行以下步骤。

Step1　如果 Tag=1，若 T 中未被标记的规范部件 C 由超过一个单个行为或原子部件构成，则 S_k←C，并标记 C；若 RPST 中未标记部件中不存在这样的规范部件，则 Tag=2。

Step2　如果 Tag=2，若 T 中仍未被选择节点的层中包含叶子节点，则随机选择一个未标记叶子节点对应的行为作为种子 S_k，并标记该叶子节点。若 RPST 中不存在这样的未标记行为，则 Tag=3。

Step3　如果 Tag=3，随机选择一个与已被选节点不直接相邻的未标记行为作为种子 S_k，并标记该行为；若 RPST 中不存在这样的未标记行为，则 Tag=4。

Step4　如果 Tag=4，随机选择一个未标记单个行为作为种子 S_k。

5　流程抽象模型的有效性指标

本文处理的聚类样本是业务流程模型中的行为集合，行为之间具有控制流顺序的约束关系，因此利用保序流程抽象的要求，根据模型抽象结果对控制流顺序的改变程度设计流程抽象模型的有效性指标。

基于初始模型对应的行为文档来计算抽象行为间的控制流关系[13]。抽象行为之间的关系同时会映射到对应原始模型中的细节行为，从而导致原始流程模型的控制流顺序发生冲突。为了评估抽象结果对初始模型控制流的影响程度，本文引入了行为的 m^3 相似性度量[14]，该度量方法基于行为文档计算两个流程的相似度。

设抽象结果模型 PM_a 中的抽象行为和原始模型 PM 中的细节行为之间的映射关系为 φ，抽象行为(子流程)个数 m 通常远小于细节行为个数 n，即 $m\ll n$。利用文献[14]中的算法，在 $O(m^2)$ 时间内生成抽象行为间的控制流关系，并且构造抽象模型 PM_a 对应的行为文档 BP_{PM_a}。

基于行为文档 BP_{PM_a}，可以推导出初始模型中细节行为之间的新关系(见算法 1)，其时间复杂度不超过 $O(n^2)$。

算法 1 Transform(BP_{PM},BP_{PM_a},φ)

//计算原始模型 PM 的新行为文档 $BP^*_{PM_a}$,该行为文档反映了抽象模型 PM_a 对 PM 中行为控制流关系的反向影响。

输入:原始模型 PM 对应的行为文档 BP_{PM},抽象模型 PM_a 对应的行为文档 BP_{PM_a},抽象模型 PM_a 中抽象行为与原始模型 PM 中细节行为之间的映射关系 φ

输出:原始模型 PM 的新行为文档 BP^*_{PM}

BP^*_{PM}←BP_{PM}

对于 BP_{PM_a} 中的每对行为 x 和 y(设 $BP_{PM_a}(x,y)=R^*$)

根据 φ,对于任意 a∈aggregate(x)并且 b∈aggregate(y)

$BP^*_{PM}(a,b)$←R^*

计算 BP_{PM} 和 BP^*_{PM} 的 m^3 相似度[14],其值越大,说明抽象模型对初始模型行为间控制流顺序的改变越小,亦即引起的冲突越小。因此,设计一个新的评价抽象结果模型的有效性指标如下:

$$I^*(k)=m^3-SIM^{(k)}(BP_{PM},BP^*_{PM}) \tag{1}$$

$$k_{opt}=\max_{k_{min}\leqslant k\leqslant k_{max}} I^*(k) \tag{2}$$

$m^3-SIM^{(k)}(BP_{PM},BP^*_{PM})$表示抽象结果对原始模型控制流改变程度的指标,其值越大,说明抽象模型对初始模型控制流的改变越小,聚类效果越好;k_{opt} 表示最佳子流程数。

6 求解最佳子流程数的算法及实验结果

综上,利用提出的子流程数搜索范围确定方法、初始簇中心的增量式确定方法和式(1)定义的抽象结果有效性指标,结合文献[5-6]中设计的基于两种不同约束条件下的受限的 k-means行为聚类算法,本文给出确定最佳子流程数的算法,如算法 2 所示。

算法 2 KM-kopt

Step1 选择聚类数的搜索范围[k_{min},k_{max}]。

Step2 For k=k_{min} to k_{max}

Step2.1 调用受限的行为聚类算法[5-6];

Step2.2 利用式(1)计算抽象结果的有效性指标值。

Step3 利用式(2)计算最佳聚类数。

Step4 输出最佳聚类数、有效性指标值和行为聚类结果。

令 $A=\{a_1,\cdots,a_n\}$ 为业务流程模型 PM 的行为集合,$D=\{d_1,\cdots,d_n\}$ 为行为对应的虚拟文档集合。$\{\mu_1,\cdots,\mu_k\}$ 表示第 4 节中初始化的簇集合$\{S_1,\cdots,S_k\}$对应的 k 个划分中心。对于每个 $a\in A$,当将其分配到簇 S_i 时,不仅考虑 a 和 μ_i 之间的语义相似性(距离),同时考虑 a 加入到 S_i 产生的可能的控制流冲突(约束函数的第二部分)。因此,文献[5]将语义相似性和控制流顺序相结合来设计约束函数限制簇的选择,即当将行为 a 分配到某一个簇时,选择使得以下目标函数最小化的簇 S_i:

$$objective(S_i,a)=dist(d,\mu_i)+conflicts^*(S_i\cup\{a\}) \tag{3}$$

其中,$dist(d,\mu_i)$表示行为 a 与簇 S_i 中心的距离;$conflicts^*(S_i\cup\{a\})$表示将行为 a 归类到 S_i 时引起的可能的控制流顺序冲突,冲突的计算利用了行为文档的概念(如定义 3 所示),具体计算过程详见文献[6]。

文献[6]对该目标函数进行了改进,为约束函数中的每个项添加了权重值,如式(4)所示:

$$objective1(S_i,a)=w_1 dist(d,\mu_i)+w_2 conflicts^*(S_i\cup\{a\}) \tag{4}$$

其中,w_1 和 w_2($0\leqslant w_1,w_2\leqslant 1$)的值可以隐含设计者的抽象重点;若 $w_1=1$ 并且 $w_2=0$,则说明分类仅仅基于行为的业务语义;若 $w_1=0$ 并且 $w_2=1$,则表示分类仅考虑控制流顺序保序要求。

分两种方式获得 w_1 和 w_2 的值:1)用户根据其抽象所强调的重点,明确地指定这两个参数的值;2)采取一种非确定方式,从包含丰富子流程关系的流程数据库中挖掘得到两个参数的值,详见文献[6]。

对文献[6]中的实验流程模型 M_1-M_{40}(模型属性如表 1 所列)进行算法的执行,其中分别利用本文给出的两种生成初始簇的方法以及文献[5-6]中设计的基于两种不同约束条件的受限的 k-means 行为聚类算法。

实验首先根据流程模型 M_1-M_{40} 生成其对应的展开模型,然后对每个模型运行算法 KM-kopt,得到最佳的子流程数 k_{opt}。两种生成初始簇的方法与两种受限的 k-means 算法交替组合,运行算法 KM-kopt 得到 k_{opt} 与原始模型中的子流程数 k 的对比结果,如表 2 所列。为了简化,表 2 中只给出对 40 个模型求得的 k_{opt} 与 k 的平均值,即 k^*_{opt} 和 k^*。

表 1 实验流程模型 M_1-M_{40} 的相关属性[6]

	行为	子流程	子流程中的行为
平均	94.10	7.97	7.52
最大	127.00	20.00	10.50
最小	59.00	3.00	4.20

表 2 对 M_1-M_{40} 运行算法 KM-kopt 求得的最佳子流程数

使用的 k-means 方法 / 生成初始簇的方法	k^*_{opt} 受限的 k-means 行为聚类算法(约束条件 1[5])	k^*_{opt} 受限的 k-means 行为聚类算法(约束条件 2[6])	k^*
基于行为连接紧密性	6.21	6.33	7.52
基于 RPST	6.78	7	

事实上,经过统计,在对这 40 个模型运行算法 KM-kopt 的过程中,基于 RPST 生成初始簇,并使用文献[6]提出的 k-means算法进行行为聚类,有 80%以上的模型求得的 k_{opt} 值与原始模型中的实际子流程数之间的差距都在 2%之内。

结束语 本文主要探讨如何在进行业务流程抽象的行为聚类之前,预先生成比较合理的最佳子流程数。基于笔者在前期工作中提出的两种不同约束条件下的受限 k-means 行为聚类算法,设计了生成最佳子流程数的算法,并进行了实验结果分析。

k-means 聚类是一个数据集的硬划分方法,即每个行为必须归类至某一个子流程。但实际上,在建模者手工进行子流程划分时,存在大量不属于任何子流程的行为,有些与某个子流程距离较近的行为甚至会被人工分配到其他距离较远的子流程中。

这些局限给了我们未来研究的方向,比如可以应用和改进软聚类技术(如 FCM(Fuzzy C-Means)聚类)来代替 k-means 聚类,从而更灵活地对行为进行子流程归类。

参 考 文 献

[1] SMIRNOV S,REIJERS H A,WESKE M. A Semantic Approach for Business Process Model Abstraction[C]//Proceedings of the CAiSE 2011. Springer,2011:497-511.

(下转第 275 页)

2015,42(9):263-267. (in Chinese)
明洁,张贵军,刘玉栋. 多模式公交组合调度优化模型[J]. 计算机科学,2015,42(9):263-267.

[3] CEDER A. Designing Transit Short-Turn Trips with the Elimination of Imbalanced Loads [M]. Computer-Aided Transit Scheduling,1988:321-326.

[4] AICHONG S,MARK H. The Real-Time Stop-skipping Problem [J]. Journal of Intelligent Transportation Systems,2005,9(2):91-109.

[5] MEKKAOUI O,DE PALMA A,Lindscy R,et al. Optional bus timetables and trip timing preferences[C]// The 8th International Conference on Computer aided Scheduling of Public Transport. 2000:356-364.

[6] 杨兆升. 城市智能公共交通系统理论与方法[M]. 北京:中国铁道出版社,2004.

[7] WU L R. Real-time Tour Planning for Flexible Route Bus Considering Passengers' Waiting Behavior[D]. Dalian:Dalian University of Technology,2014. (in Chinese)
吴丽荣. 考虑乘客等待行为的柔性路径公交车实时调度方法[D]. 大连:大连理工大学,2014.

[8] NIU X Q,CHEN Q,WANG W. Optimal model of urban bus frequency determination[J]. Journal of Traffic and Transportation Engineering,2003,3(4):68-72. (in Chinese)
牛学勤,陈茜,王炜. 城市公交线路调度发车频率优化模型[J]. 交通运输工程学报,2003,3(4):68-72.

[9] 余志生. 汽车理论[M]. 北京:机械工业出版社,2009.

[10] TIAN R,SUN L F. Multi-unloading Packing Problem Model Research Based on MMAS[J]. Journal of Chongqing Jiaotong University(Natural Science Edition),2016,35(2):156-162. (in Chinese)
田冉,孙林夫. 基于最大最小蚁群算法的多卸载点车载装箱模型研究[J]. 重庆交通大学学报(自然科学版),2016,35(2):156-162.

[11] GE B,HAN J H. Dynamic Adaptive Ant Colony Optimization Algorithm for Min-Max Vehicle Routing Problem[J]. Pattern Recognition and Artificial Intelligence,2015,28(10):930-938. (in Chinese)
葛斌,韩江洪. 最小最大车辆路径问题的动态自适应蚁群优化算法[J]. 模式识别与人工智能,2015,28(10):930-938.

[12] STUTZLE T,HOOS H H. MAX-MIN Ant system[J]. Future Generation Computer Systems,2000,16(9):889-914.

[13] YAO Y. Research for the Improvement of Max-Min Ant Colony Algorithm[J]. Mathematics in Practice and Theory,2014,44(15):242-247. (in Chinese)
姚艳. 一种最大最小蚂蚁系统的改进算法[J]. 数学的实践与认识,2014,44(15):242-247.

[14] DOU H L,LIU H D,YANG X G. OD matrix estimation method of public transportation flow based on passenger boarding and alighting[J]. Computer and Communication,2007,25(2):79-83. (in Chinese)
窦慧丽,刘好德,杨晓光. 基于站点上下客人数的公交客流 OD 反推方法研究[J]. 交通与计算机,2007,25(2):79-83.

[15] WANG W Q. Research on Gas-Saving technology for City CNG buses[D]. Xi'an:Chang'an University,2012. (in Chinese)
王文强. 城市 CNG 公交车辆节气技术的研究[D]. 西安:长安大学,2012.

(上接第 248 页)

[2] REIJERS H A,MENDLING J,DIJKMAN R M. On the Usefulness of Subprocesses in Business Process Models[J]. Information Systems(IS),2012,37(5):443-459.

[3] ZHOU S B. Research and application of optimal cluster number determination method in clustering analysis[D]. Wuxi:JiangNan University,2011. (in Chinese)
周世兵. 聚类分析中的最佳聚类数确定方法研究及应用[D]. 无锡:江南大学,2011.

[4] CHEN L F. Research on Clustering Methods for High Dimensional Data and Their Applications[D]. Xiamen:Xiamen University,2008. (in Chinese)
陈黎飞. 高维数据的聚类方法研究与应用[D]. 厦门:厦门大学,2008.

[5] WANG N,SUN S W. Constraint-based Activity Clustering in Business Process Model Abstraction [J] Computer Science,2017,44(1):259-263,294. (in Chinese)
王楠,孙善武. 业务流程模型抽象中基于约束的行为聚类方法研究[J]. 计算机科学,2017,44(1):259-263,294.

[6] WANG N,SUN S W,OUYANG D T. Business Process Modeling Abstraction Based on Semi-Supervised Clustering Analysis [C]//Business & Information Systems Engineering. 2016.

[7] SMIRNOV S. Business Process Model Abstraction,. Germany:University of Potsdam[OL]. http://opus. kobv. de/ubp/volltexte/2012/6025/pdf/smirnov_diss. pdf.

[8] POLYVYANYY A,VANHATOLO J,VOLZER H. Simplified Computation and Generalization of the Refined Process Structure Tree[C]// Proceedings of the WS-FM 2010. Springer,2011:25-41.

[9] FREY B J,DUECK D. Response to Comment on Clustering by Passing Messages Between Data Points[J]. Science,2008,319(5864):726.

[10] YANG S L,LI Y S,HU X X,et al. Optimization Study on k Value of K－means Algorithm K-means[J]. System Engineering Theory and Practice System EngTheorPrac,2006,26(2):97-101. (in Chinese)
杨善林,李永森,胡笑旋,等. K-means 算法中的 k 值优化问题研究[J]. 系统工程理论与实践,2006,26(2):97-101.

[11] FREY B J,DUECK D. Clustering by passing messages between data points[J]. Science,2007,315(5814):972-976.

[12] SUN S W,WANG N,OUYANG D T. Business Process Model Abstraction Based on Cluster Analysis [J]. Computer Science,2016,5(5):193-197. (in Chinese)
孙善武,王楠,欧阳丹彤. 基于聚类分析的业务流程模型抽象[J]. 计算机科学,2016,5(5):193-197.

[13] SMIRNOV S,WEIDLICH M,MENDLING J. Business Process Model Abstraction Based on Behavioral Profiles[M]// Service-Oriented Computing. Springer Berlin Heidelberg,2010:1-16.

[14] SMIRNOV S,WEIDLICH M,MENDLING J. Business Process Model Abstraction Based on Synthesis from Well-Structured Behavioral Profiles[J]. International Journal of Cooperative Information Systems,2012,21(1):55-83.