基于动态抽样的图分类算法

尹婷婷,刘俊焱,周溜溜,业　宁,尹佟明

(南京林业大学信息科学技术学院,江苏 南京 210037)

[**摘要**]　传统的图分类算法由于支持度阈值选择过低导致频繁子模式规模过大,进而造成效率过低,阈值选择过高导致重要模式丢失而造成分类精度下降,如 FSG 和 CEP 方法. 针对这些问题,提出将动态抽样策略引入图分类领域,在保持分类准确率的前提下通过顶点平均度的计算抽样选取代表性子模式,结合 CEP 所给出的频繁闭显露模型,设计出一种新的图特征(分类规则)提取方法,解决了 CEP 算法由于支持度阈值设置过低而导致的无法计算现象,大大提高了分类效率;并通过实验证明本文算法优于现有的一些主流算法.

[**关键词**]　图分类,动态抽样,顶点平均度,代表子模式

[**中图分类号**]TP311.13　[**文献标志码**]A　[**文章编号**]1001-4616(2015)01-0113-06

Graph Classification Algorithm Based on Dynamic Sampling

Yin Tingting, Liu Junyan, Zhou Liuliu, Ye Ning, Yin Tongming

(College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China)

Abstract: Support threshold of traditional graph classification algorithm like FSG or CEP is too low that may cause oversized frequent subschema and low efficiency, while too high may lead to the loss of important models or accuracy drop. To solve these problems, we introduce the strategy of dynamic sampling into the graph classification and design a new pattern feature extraction method, by which we get the representative sub-model by calculating every graph's average vertex degree and use the frequent closed revealed model from CEP, under the premise of classification accuracy. The new method settled the problem unable to be calculated due to support threshold chosen too low in CEP and greatly improved the classification efficiency. Experiments showed that the new method surpassed a series of mainstream algorithm in this field.

Key words: graph classification, dynamic sampling, average vertex degree, representative sub-model

近年来,随着图结构在生物信息学、化学、社会网络分析等领域的广泛应用,图挖掘和图数据管理逐渐成为了热点研究内容[1,2]. 因图模型能够准确地表述科学与工程领域中数据的关键特征,所以很多基于图的数据挖掘算法被提出. 利用图结构化形式对大数据进行表现与阐述,既可达到降低信息损耗的目的,又可更方便地利用图结构对其进行刻画[3].

图分类策略主要分为两种:一种是基于频繁子图挖掘的图特征提取方法,其中以 FSG 算法[4]为主要代表;另一种是基于图核函数[5-7]的方法,其中以文献[5]中的 CPK 分类算法为主. 但是这两种策略都具有局限性和瓶颈,图核函数的算法大多基于一个假设,即图数据中的环结构数受限于某一固定常数,且只能使用于具有天然性图结构的数据中,如果只具有少量的环结构甚至无环时该方法不可用,缺乏可扩展性,而基于频繁子图策略的算法不可避免地涉及到了子图规模的选择以及子图同构的麻烦. 如果在做频繁子图挖掘步骤时选择的支持度阈值过低,则会导致分类模型无法在有效时间内形成,反之如果阈值设定过高,又会导致区分度较高的有效模式流失现象,降低模型的精确度;同时子图同构是一个经典的 NP 完全问题,因此如何有效减少图同构计算规模与次数也成为了一条提高图数据管理效率的可行之路. 文献[2]给出的 CEP 学习框架便是绕开了传统的普通频繁子图方式,利用频繁闭图数量远远小于频繁子图数(如在 NCI/CA[8]数据中:当支

收稿日期:2014-08-16.

基金项目:国家 973 项目(2012CB114505)、国家杰青项目(31125008)、江苏省自然科学基金(BK2012815)、江苏省青蓝工程项目、江苏省六大人才高峰项目、江苏省 2013 年度普通高校研究生科研创新计划项目(CXZZ13_0538).

通讯联系人:业宁,教授,研究方向:数据挖掘,特征信息. E-mail:yening@ njfu.edu.cn

持度阈值设为5%时,频繁图可达到100万张,而闭图尚不足2 000张)并且能够保证继承其分类作用的特性,通过对频繁闭图挖掘,并利用显露模式概念初步过滤冗余闭图获得显露模式集,再进行分类特征规则提取获得可计算的分类规则,最后利用规则匹配方法实现对一个新的图数据分类.

CEP 学习框架在一定程度上利用对闭图的挖掘而摒弃对频繁子图的挖掘增强了模式的可计算性,较 FSG 算法提升了性能,但该方法仍然具有以下两点缺陷:一是其所选用的阈值参数设置最低仅为 2%,仍然遗漏掉了很多的有效模式,若将该参数调低,则算法无法在有效时间内完成;二是无法作用于大规模数据库,不能满足实际应用需求. 本文基于以上两点,在继承 CEP 学习框架优势的前提下,将结合点平均度的计算抽样策略引入了图分类领域;可以在不损失分类正确率的前提下,使算法进一步调低阈值,保留有效模式同时确保了模式可计算性,设计了 DPS-CEP(Dynamic Probability Sample based on CEP)算法,提升了图分类效率. 经实验证明,DPS-CEP 优于目前主流图分类算法,理论上可以适用于任意大规模图数据集.

在文章的剩余部分,第二节介绍了相关定义和概念,第三节给出了抽样规则和算法主体思想及步骤,在第四节通过相关实验比较证明了 DPS-CEP 算法的优势,最后是对本文工作的总结和未来工作的展望.

1　背景知识

本文中所处理的数据为普通无向标号图. 在介绍具体算法之前,需要给出有关基本概念和问题的解释,其中部分与文献[2]一致.

定义 1　标号图:任何一个标号图 G 均可以用一个五元组来表示,$G=\{V,E,\sum V,\sum E,l\}$,其中 V 为图的顶点集合,E 为边集合,$\sum V$ 为顶点标号集,$\sum E$ 为边标号集,l 为标号映射函数.

定义 2　支持度:给定一组图数据 $GD=\{G_i \mid i=1,2,\cdots,n\}$,图模式 p 的支持度 θ 为 GD 中所有包含 p 的图数量与$|GD|$的比例;当 θ 大于某一给定的支持度阈值 θ_0 时,则 p 为频繁模式.

定义 3　频繁闭图:在频繁模式集 FM 中,若存在支持度相等的子图 p 和 p_1,若 $p\subset p_1$,且 p_1 不再是其他相等支持度模式的子图,那么 p_1 便被称作频繁闭图(以下简称为闭图).

定义 4　显露比:设模式 p 在 2 个图数据库 GD_1 和 GD_2 中的支持度分别为 θ_1 和 θ_2,则称 $\frac{\theta_1}{\theta_2}$ 为模式 p 从 GD_1 到 GD_2 的显露比,记为 $emr(p,GD_1,GD_2)$;当 $emr(p,GD_1,GD_2)$ 大于某一设定阈值时,则称 p 为 GD_1 中的显露模式.

定义 5　图的环模式:首先定义在一个图结构中任意两顶点距离为其之间所连接的最少边数,所有与顶点 v_0 距离为 n 的顶点集合称为 v_0 的第 n 个环.

图分类问题可以类比与数据挖掘中的关系数据的分类问题,其主要目的也是将每一张图数据按照一定的规则判定为某一特定类别,而这些规则类比于传统的分类器. 文中所给出的 DPS-CEP 算法就是利用已有的图信息(训练集)学习得出一系列判定规则,进而对未知信息图进行类别归纳. 为方便起见,本文中所涉及的分类问题均为二元分类,只要将 DPS-CEP 稍加改动,也同样适用于多元分类.

2　DPS-CEP 算法

我们设计了 DPS-CEP(Dynamic Probability Sample based on CEP)算法,在不损失分类正确率的前提下,进一步调低阈值,保留有效模式的同时确保了模式可计算性,提升了图分类效率. 鉴于 DPS-CEP 是在 CEP 算法的基础上改进,在介绍 DPS-CEP 之前先对 CEP 学习框架和动态抽样作一些简单的说明.

2.1　CEP 学习框架

CEP 学习框架主要分为 4 个模块:首先利用频繁闭图挖掘算法 CloseGraph[8] 分别对训练集中正例和反例进行操作,分别获得闭图集 PCS 合 NCS;其次再对获得的 2 个闭图集进行显露模式过滤从而得出 2 个显露模式集 PES 和 NES;再次利用覆盖方法[9,10] 对显露模式集实现再甄选,获得有限个便于计算且信息含量丰富的元素组合成分类规则,最后对测试集中的每一个样本图数据进行预测分类. 下面给出 CEP 算法的伪代码,如算法 1 所示:

算法 1. CEP.

输入:两个已知类别的图集 PGD,NGD.

输出:一系列分类规则$\{R_1,\cdots,R_n\}$

```
[PCS,NCS] = CloseGraph[PGD,NGD]
    for pc ∈ PCS, do
        计算 pc 的显露 emr_pc
            if emr_pc大于显露比阈值 emr_o
                将 pc 放入显露模式集 PES
            else delete pc
        end
    对 NCS 中每个元素做如上同样操作以获得 NES
从 PES 和 NES 中选出有效模式,分别组成正例分类规则 PER 和反例分类规则 NER
    由 PER 和 NER 构成{R_1,…,R_n}.
```

2.2 抽样策略

正如上述对目前利用频繁子模式进行图分类的算法所分析的那样,此类算法所面临最难解决的问题是对于是否属于频繁具有决定“话语权”的阈值大小的设定问题. 如果阈值设定过大,则很明显将会遗漏部分对分类结果起重要影响的子图;反之设定过低,就会获得数量巨大的子图,虽然理论上可以提升模型的分类精度,但是在模型构造阶段会消耗大量的时间,甚至无法在有效时间内构造成功. 观察后发现,此类算法大多是在通过阈值筛选出频繁模式后便全部用于模型的塑造;我们分析,如果在已获得的频繁模式集中再一次按规则进行筛选抽样,便可以获得一个可接受易计算的适当规模频繁集,最后利用这一频繁集进行模型塑造. 在第四节的实验部分证明了该策略的有效性,两种策略的主要异同点如图 1 所示如下.

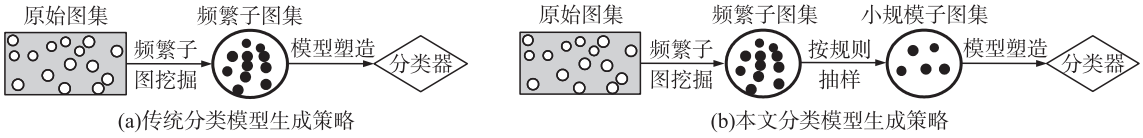

图 1 两种图分类模型生成策略示意图

Fig. 1 Two kinds of graph classification models generation strategy

2.2.1 抽样原则

抽样规则需要遵循以下两个原则:

(1)抽取的样本容量不能太小,如果规模太大则体现不出抽样优势,取值太小容易忽略等价类子图.

(2)所抽取出的每一个样本要具有代表性,可以尽量覆盖等价类子图的结构信息,对分类模型的效果起积极作用.

2.2.2 图的点平均度

我们知道图是一种包含丰富结构信息的数据. 那么在一个边数确定的普通无向图中,是否无论顶点如何组织都含有等量的结构信息呢,答案显然是否定的. 例如图 2 所给出的两种不同结构的三边图以及图 3 的两个五边图所包含的信息便是不相等的. 根据图论的观点,结构越复杂的图所包含的结构化信息也就越丰富,那么利用一个什么样的标准来衡量一幅图是否足够复杂,便是一个值得探讨的问题;在本文中,我们采用图顶点的平均度量方法,即 $\theta=\sum_{i=0}^{|V|-1} de(v_i)\Big/|V|$,其中 $de(v_i)$ 表示顶点 v_i 的度.

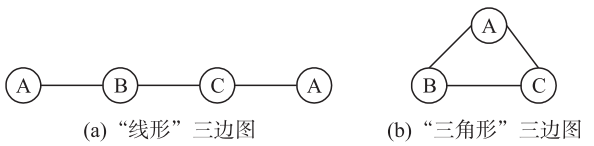

图 2 两种不同结构的三边图

Fig. 2 Three-eage graph of different structures

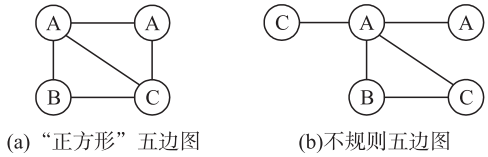

图 3 两种不同结构的五边图

Fig. 3 Five-eage graph of different structures

显然图 2(a)所示结构三边图的顶点平均度为(1+2+2+1)/4=6/4=1.5,图 2(b)的顶点平均度为(2+

2+2)/3=6/3=2;图 3(a)的顶点平均度为(3+2+2+3)/4=10/4=2.5,图 3(b)的顶点平均度为(1+4+1+2+2)/5=10/5=2. 所以若按支持度以及显露比判定后可以得出图 2(a)和 2(b)、图 3(a)和 3(b)属于等价类子图结论;则可以判定图 2(b)的结构化信息要比图 2(a)丰富,图 3(a)的结构化信息要比图 3(b)丰富.

2.2.3　抽样规则

在本文的第二节定义 5 中给出了图的环模式定义. 由定义可知,任意一个属于 k 阶图 G_k 的顶点 v 的环都是一个顶点集合,且该顶点最多 v 具有 $k-1$ 个环;假设这 $k-1$ 个环的元素个数分别为$(x_1,x_2,\cdots x_{k-1})$,构成一个向量. 称这样的一个向量为顶点 $\boldsymbol{v}$ 的一个环分布,那么所有可能的环分布将组成一个环分布矩阵[11],用 $\boldsymbol{W}$ 表示;且 $\boldsymbol{W}$ 的元素个数为 $N*(k-1)$,其中 $N=2^{k-2}$. 例如,一个三边连通图的环分布矩阵为:$\boldsymbol{W}=\begin{pmatrix}2 & 0\\ 1 & 1\end{pmatrix}$.

其中 W_{ij} 表示顶点 v 的第 i 类环分布第 j 个环的元素个数. 由图 2,3 的例子可以观察到所有 k 阶图的顶点数不一定相等,且不同结构 k 阶图的顶点平均度(即结构复杂度)也是不同的. 假设现存在一张 k 阶图 G_k,其顶点个数为 n 个,则 G_k 的每一个顶点 v_i 按照上述环分布介绍都可以对应一个环分布矩阵 $\boldsymbol{W}_i$. 为了利用这样的环分布矩阵刻画一个图的紧密程度以及其结构复杂程度,我们定义如下目标函数:

$$Com(G)=\sum_{r=1}^{2^{|E|-2}} w_r. \tag{1}$$

$$\text{其中}\quad w=\sum_{j=1}^{|E|-1}\frac{\boldsymbol{W}_{ij}}{j},\quad \text{且}\quad W=\sum_{i=1}^{|V|}\boldsymbol{W}_i. \tag{2}$$

上式的 w_r 表示向量 $\boldsymbol{w}$ 的第 r 个元素取值;$\boldsymbol{W}_{ij}$ 表示矩阵 $\boldsymbol{W}$ 的第 j 列的列向量,$\boldsymbol{W}_i$ 表示图的顶点 v_i 所对应环分布矩阵,$Com(G)$ 表示了图 G 的紧密程度以及其结构复杂程度. 其中 $\boldsymbol{W}_{ij}/j$ 反映了在统计环分布矩阵元素作用程度时,距离被选择中心点越远则其重要程度越低的规律. 有了这样一个量化指标函数,便可以按照此规则为每一类等价类子图中元素计算出一个目标函数值 $Com(G_i)$,然后在此类集中只要按需求取出 m 张取值为 *top-m* 的对应子图即可,最后利用这些小规模子模式集进行模型塑造.

2.3　DPS-CEP 算法步骤

经过以上讨论与分析,可以给出 DPS-CEP 算法的主要运行步骤如下:

1　将已知类别的图集分类为正类 PGD 和负类 NGD,
2　分别对 PGD、NGD 做频繁子图挖掘获得正类频繁子图 PFD、负类频繁子图 NFD,
3　从两类频繁子图集分别获得闭图集 PCD 和 NCD,
4　将 PCD 和 NCD 中的所有闭图按边的条数分类,获得各类 PCD_i 和 NCD_i,
5　将每一类 k 阶闭图按元素求得每一张闭图的 $Com(G)$ 值,并按降序排列,
6　按照需求规模在每一等价类闭图集中取出 *top-m* 的 $Com(G)$ 值对应闭图构成新的闭图集 PNCD 和 NNCD,
7　对 PNCD 和 NNCD 做算法 1 中的 2-10 步骤运算,
8　最后获得基于正负两类不同的分类规则集合,形成分类模型,
9　对新的未知类图数据进行判定.

3　实验与分析

实验运行环境为双 CPU PC 机,CPU 是主频为 3.2G 的 Intel Pentium 4 多线程处理器,内存为 1G,操作系统是 Windows XP Professional SP3. 实验的一切算法实现均利用 Matlab 完成,由于本文的部分实验比较需要对频繁闭图进行挖掘,所以也实现了 CloseGraph[8] 算法. 实验共分为两部分,一部分是本文算法与利用闭图构造分类器的 CEP 算法进行比较,主要是从模型构造时间以及最终的模型分类准确度两方面进行衡量;另一部分是 DPS-CEP 与 FSG 算法的两方面性能比较. 需要指出的是:在本节实验中,始终将抽样提取后所获得数据集规模控制在 3 000,本身数量不足 3 000 的取其实际规模(如支持度阈值为 5%时 NCI/CA 数据中闭图数量约为 2000). 算法本身要求抽样后的数据集规模设置为某个特定的常数值,我们将其控制在 3 000 的原因有两点:一是 FSG 算法在本文所采用的支持度阈值背景下均达到了数万计的频繁子结构,为了体现 DPS-CEP 算法的有效性和高效性,我们决定将抽样规模控制在以千为单位的数量级(如果采取以百为单位数量级,则不可避免地丢失重要结构信息);二是 CEP 算法中支持度阈值为 5%时 NCI/CA

数据中闭图数量约为 2 000,阈值为 3%时闭图数量约为 15 000,为了使得规模远低于 15 000,同时方便与 2 000 做比较,我们选取了以 2 000 为基准并向上扩张 50%的规模.

3.1 与 CEP 算法的比较

在本文的前面部分已介绍到,CEP 算法为了解决 FSG 算法中的频繁子图集规模太大这一缺陷,采用了规模相对较小的闭图进行模型构造,一定程度上解决了这一问题. 但若将 CEP 算法使用在一个更大规模数据集上的话,那么势必会导致该数据集的闭图集也会随之增大,这样便会再次出现 FSG 算法的弊端;同样倘若为获得高分类准确率而将阈值调低的话,也会出现这种现象. 而我们所给出的 DPS-CEP 算法正是针对这两点考虑,所以在下面的实验数据展示部分,我们主要通过调节阈值来观察两种算法的实验结果. 为了方便与 CEP 算法比较,我们采用了与文献[2]相同的数据库:NCI-HIV 化合物数据集(共 43 905 个),该数据集可以从 http://dtp.nci.nih.gov/下载获得. 该数据集根据其中每一种化合物对人体 CEM 细胞保护的强度又具体可以分为 3 个子数据库:CA,CM,CI. 其中 CA 有 422 个,CM 有 1081 个,其余的均属于 CI. 同样为了便于比较,在此我们只考虑以下两种二元分类情况:CA 与 CM 的分类以及 CA 与 CI 的分类(假定前者为正类,后者为负类). 具体实验结果由图 4 和图 5 所示,其中图 4(b),5(b)采用的训练集是通过随机抽取 80%正负数据集所得,余下的 20%作为测试集用于验证分类效果.

从图 4 所仿真的实验结果可以看出两点,首先本文算法在通过引入抽样策略缩小频繁集规模后所构造模型的分类精度与 CEP 算法高度可比,甚至在某些时刻还要优于 CEP;其次在模型构造时间方面的优势体现地非常明显,这也是我们设计算法的出发点,尤其当支持度阈值设置较低时(如 1%),这也印证了此前关于算法的理论分析. 为了令算法的比较更具有说服力,以下将会利用规模更大的 CI 数据集对算法进行验证,如图 5 所示. 伴随着数据集规模的扩大,DPS-CEP 的性能优势体现地更为突出. 如图 5(a)所示,当频繁支持度阈值位于 5%~10%时,此时的效率提升程度要远高于图 4(a)的结果,这说明了新算法对大规模数据集的操作更为有效. 且从图 5(b)也可以看出在大规模数据集上,新算法所构造模型分类精度的稳定性要强于 CEP 方法. 需要指出的是:实验中所设定正类图集到负类图集的显露比阈值为 2.

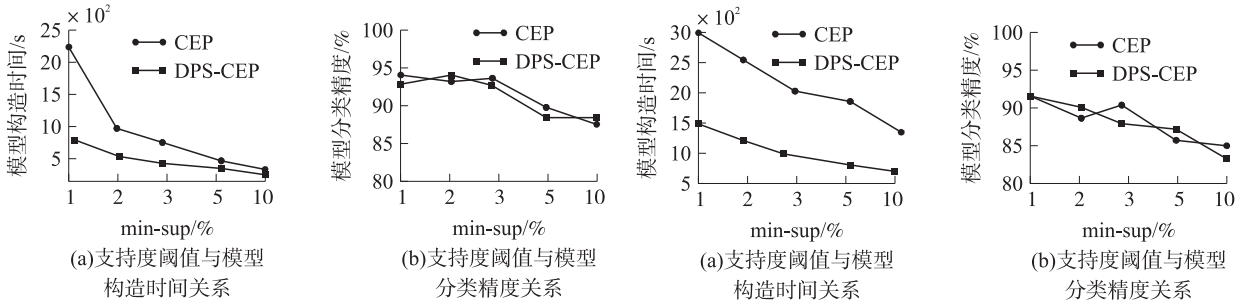

图 4 CA 与 CM 的二元分类实验结果

Fig. 4 Binary classification results of CA and CM

图 5 CA 与 CI 的二元分类实验结果

Fig. 5 Binary classification results of CA and CI

3.2 与 FSG 算法的比较

在本部分实验中之所以没有将 FSG 算法一同纳入图 4~5 中与上述两种算法作比较,主要是因为 FSG 算法所采用的频繁模式集是频繁子图而非频繁闭图. 而我们所设计的 DPS-CEP 算法可以适用于任何一种频繁模式,所以在此部分为了方便比较,我们选择在频繁子图集中进行抽样提取操作[12,13]. 实验数据结果如表 1~4 所示(若运行时间大于 5 000 s 时,用 N/A 表示).

表 1 集中列出了 DPS-CEP 算法与 FSG 算法作用于 CA 与 CM 数据集时构造模型的时间消耗,可以观察出时间随着支持度阈值的增大所造成的频繁集规模减小而减少. 很明显,频繁子图的规模往往大于频繁闭图的规模. 由于 FSG 算法既没有抽样策略的作用也没有采用闭图模式,所以算法的模型构造时间消耗较大. 而文中所设计的新算法由于可以控制频繁集规模大小而使得效率有了很大程度地提升. 由于实验采取了固定规模策略,所以新算法的时间消耗增加主要是由抽样过程导致.

表 2 集中列出了 DPS-CEP 算法与 FSG 算法构造出模型作用于 CA 与 CM 数据集的分类精度. 新算法的分类精度要略低于 FSG 算法但高度可比,且要比 FSG 算法更加稳定,不会由于频繁集规模的减小而造成较大程度地下降.

表 3 的结果与表 1 类似,但是由于 CI 数据集的规模比较大,所以导致 FSG 算法在低支持度阈值情况下无法在可接受时间内(5 000 s)完成模型的塑造. 同时伴随着频繁集规模的增大,使得 DPS-CEP 抽样过程更加复杂,故而令时间消耗增多.

表 4 的结果与表 2 类似,可以得出结论即使对大规模数据集进行操作,本文 DPS-CEP 算法的分类性能也与 FSG 算法高度可比. 综合表 3 的结果,新算法的综合性能优于 FSG.

表 1　两种算法的模型构造时间(CA 与 CM 数据集比较结果)

Table 1　Model structure time of two algorithms(CA and CM)

支持度阈值/%	FSG 算法/s	DPS-CEP/s
1	N/A	1 424
2	N/A	926
3	4 768	788
5	2 982	554
10	1 746	388

表 2　两种算法的模型分类精度(CA 与 CM 数据集比较结果)

Table 2　Model classification accuracy of two algorithms(CA and CM)

支持度阈值/%	FSG 算法/%	DPS-CEP/%
1	89.6	82.1
2	87.2	81.6
3	88.1	79.9
5	82.4	80.2
10	79.4	76.9

表 3　两种算法的模型构造时间(CA 与 CI 数据集比较结果)

Table 3　Model structure time of two algorithms(CA and CI)

支持度阈值/%	FSG 算法/s	DPS-CEP/s
1	N/A	1 914
2	N/A	1 236
3	N/A	996
5	N/A	754
10	2 146	531

表 4　两种算法的模型分类精度(CA 与 CI 数据集比较结果)

Table 4　Model classification accuracy of two algorithms(CA and CI)

支持度阈值(%)	FSG 算法/%	DPS-CEP/%
1	89.4	83.5
2	88.1	83.4
3	85.2	83.1
5	84.6	80.2
10	84.3	77.1

4　结论与展望

本文在分析了一系列主流的传统图分类算法各自所存在缺陷的基础上,提出了通过计算顶点平均度而采用抽样方法,进而选取代表性的子模式用于构造分类模型,给出了对应的 DPS-CEP 算法. 克服了由于频繁模式集规模过大所造成的模型塑造时间过长问题,使得算法理论上可以适用于任何大规模的原始图集操作,不再受限于初始支持度阈值的选择. 同时因为抽样选取的模式含有丰富的结构化信息,所以新算法的分类性能与其他算法高度可比,甚至某些时刻优于目前精度最高的算法. 而如何进一步选取那些更具有代表性信息的模式集则是我们接下来要做的工作之一,这也是提升算法分类性能的主要出发点;同时如何利用图的几何结构信息来提升我们的算法性能也是一个值得开展的工作.

[参考文献]

[1] 汪卫,周皓峰,袁晴晴,等. 基于图论的频繁模式挖掘[J]. 计算机研究与发展,2005,42(2):230-235.

[2] 刘勇,李建中,朱敬华. 一种新的基于频繁闭显露模式的图分类方法[J]. 计算机研究与发展,2007,44(7):1169-1176.

[3] 周溜溜. 基于图结构的数据挖掘研究及应用[D]. 南京:南京林业大学信息科学技术学院,2013.

[4] Deshpande M, Kuramochi M, Karypis G. Frequent substructure based approaches for classifying chemical compounds[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(8): 1 036-1 050.

[5] Horvath T, Gartner T, Wrobel S. Cyclic pattern kernels for predictive graph mining[C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD). Washington DC, USA: ACM, 2004: 158-167.

[6] Kashima H, Tsuda K, Inokuchi A. Marginalized kernels between labeled graphs[C]//Proceedings of the 20th International Conference on Machine Learning. Washington DC, USA: ICML, 2003.

[7] Borgwardt K M, Kriegel H P. Shortest-path kernels on graphs[C]//Proceedings of the 5th IEEE International Conference on Data Mining(ICDM). Houston, Texas, USA: IEEE Computer Society, 2005: 74-81.

[8] Yan X, Han J. Closegraph: Mining closed frequent graph patterns[C]//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD). Washington DC, USA: ACM, 2003: 286-295.

(下转第 127 页)

[8] Kalyanmoy D,Samir A,Amrit P,et al. A fast elitist non-dominate sorting genetic algorithm for multi-objective optimization: NSGA-II[J]. Transactions on Evolutionary Computation,2002,6(2):182-197.
[9] Zitzler E,Laumanns M,Thiele L. SPEA2:Improving the strength pareto evolutionary algorithm,technical report TIK-Report 103[R]. Swiss:Swiss Federal Institute of Technology Zurich(ETH),2001.
[10] 王宇平,焦永昌,张福顺. 解多目标优化的均匀正交遗传算法[J]. 系统工程学报,2003,13:481-486.
[11] 曾三友,魏巍,康立山. 基于正交设计的多目标演化算法[J]. 计算机学报,2005,28(7):1 153-1 162.
[12] 关世华,寇纪淞,李敏强. 基于 ε-约束方法的 Lagrangian 多目标协同进化算法[J]. 系统工程与电子技术,2002,24(9):33-37.
[13] Shi C,Li Q Y,Shi Z Z. A quick multi-objective evolutionary algorithm based on domination tree[J]. Journal Software,2007,18(3):505-516.
[14] Gong W,Cai Z,Ling C. ODE:a fast and robust differential evolution based on orthogonal[C]//LNAI 4304:pROC of Advances in Artificial Intelligence. Berlin:Springer,2006:709-718.
[15] 公茂果,焦李成,杨咚咚,等. 进化多目标优化算法研究[J]. 软件学报,2009,2(20):271-289.
[16] 龚文引,蔡之华. 基于 ε 占优的正交多目标差分演化算法研究[J]. 计算机研究与发展,2009(4):655-666.
[17] Leung Y,Wang Y. An orthogonal genetic algorithm with quantization for global numerical optimization[J]. IEEE Transaction on Evolutionary Computation,2001,5(1):41-53.
[18] Gong W Y,Cai Z H. An improved multi-objective differential evolution based on Pareto adaptive ε-dominance and orthogonal design[J]. European Journal of Operational Research,2009,198:576-601.
[19] 罗辞勇,陈民铀,张聪誉. 采用循环拥挤排序策略的改进 NSGA-II 算法[J]. 控制与决策,2010,25(2):227-231.
[20] 陈民铀,张聪誉,罗辞勇. 自适应进化多目标粒子群算法[J]. 控制与决策,2009,24(12):1 851-1 855.
[21] 贺群,程格,安军辉,等. 基于 Pareto 的多目标克隆进化算法[J]. 计算机科学,2012,39(6A):489-492.
[22] 杨尚军,王社伟,陶军,等. 基于混合细菌觅食算法的多目标优化方法[J]. 计算机仿真,2012,29(6):218-222.
[23] Robič T,Filipič B. DEMO:Differential evolution for multiobjective optimization[C]//LNCS 3410:Proc of EMO'05. Berlin:Springer,2005:520-533.
[24] Deb K,Mohan M,Mishra S. Evaluating the ε-domination based multi-objective evolutionary algorithm for a quick computation of Pareto-optimal solutions[J]. Evolutionary Computation,2005,13(4):501-525.
[25] Zitzler E,Thiele L,Laumanns M,et al. Performance assessment of multi-objective optimizer:an analysis and review[J]. IEEE Trans on Evolutionary Computation,2003,7(2):117-132.

[责任编辑:黄　敏]

(上接第 118 页)

[9] Liu B,Hsu W,Ma Y. Integrating classification and association rule mining[C]//Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining(KDD-98). New York,USA:AAAI,1998:80-86.
[10] Silva A,Meira Jr W,Zaki M J. Structural correlation pattern mining for large graphs[C]//Proceedings of the Eighth Workshop on Mining and Learning with Graphs. USA:ACM,2010:119-126.
[11] 赵建邦,董安国,高琳. 一种用于生物网络数据的频繁模式挖掘算法[J]. 电子学报,2010,38(8):1 803-1 807.
[12] 丁悦,张阳,李战怀,等. 图数据挖掘技术的研究与进展[J]. 计算机应用,2012,32(1):182-190.
[13] 薛冰,张俊峰,郑超,等. 基于分割图集的频繁闭图挖掘算法[J]. 计算机应用研究,2011,28(1):61-64,68.

[责任编辑:黄　敏]