基于拓扑信息加速马尔科夫毯学习

傅顺开 苏致祯 Sein Minn 吕天依

(华侨大学计算机科学与技术学院 厦门 361021)

摘 要 目标变量的马尔科夫毯(MB)是用于预测其状态的最优特征子集。提出一种新的约束学习类MB推导算法FSMB,它遵循后向选择的搜索策略,并依赖条件独立(CI)测试删除任意结点对之间的伪连接。与传统约束学习类算法不同,FSMB能从已执行的CI测试推导出不同结点扮演d-分割(d-separation)结点的优先等级;而后基于该信息在未来优先执行条件集中包含高优先级结点的CI测试,从而更快速地判断并删除伪连接边。该策略可帮助快速缩小搜索空间,从而大大提升学习效率。基于仿真网络的实验研究显示,FSMB在计算效率上较经典的PCMB和IPC-MB有显著的提升,而学习效果相当;在面对较大网络结构时(比如100和200个结点),甚至比公认最快速的IAMB还节省近40%的计算量,但学习效果要远优于IAMB。基于16个UCI数据集和4个经典的分类模型的实验显示,基于FSMB输出的特征集合所训练模型的分类准确率普遍接近或高于基于原有特征全集训练所得模型。因此,FSMB是快速且有效的MB推导算法。

关键词 马尔科夫毯,贝叶斯网络,局部搜索,结构学习,约束学习,条件独立测试

中图法分类号 TP391 **文献标识码** A

Accelerating the Recovery of Markov Blanket Using Topology Information

FU Shun-kai SU Zhi-zhen Sein Minn LV Tian-yi

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract Markov blanket(MB) has been known as the optimal feature subset for prediction, and there exist fertile works to induce MB by local search since 1996. A novel one called FSMB was proposed which heavily relies on conditional independence(CI) test to determine the existence of connection between nodes, so it is kind of constraint-based learning as well. However, it differs from previous works by treating candidate CI tests unfairly. FSMB extracts critical d-separation topology information from conducted CI tests, and applies them to sort and perform those more likely to uncover independent relations with priority. Search space therefore is expected to shrink quickly in a more efficient manner. Experimental studies indicate that FSMB achieves tremendous improvement over state-of-art works PCMB and IPC-MB in term of time efficiency, but with no sacrifice on learning quality. When given large networks(e. g. 100 and 200 nodes), FSMB runs even more efficiently than IAMB which is recognized as the fastest algorithm by now, requiring up to 40% fewer CI tests, and produces much higher quality of results. Experiments with UCI data sets and four classical classification models indicate that the classification accuracy of the models trained on the output of FSMB are close to or exceed performance achieved by models trained on all features, hence FSMB is an effective feature subset selector.

Keywords Markov blanket, Bayesian network, Local search, Structure learning, Constraint-based learning, Conditional independence test

1 引言

贝叶斯网络(Bayesian Network, BN)提供了一种有效且可紧凑表达联合分布的图建模工具,在过去30年被成功应用到机器学习和数据挖掘的诸多领域中。1988年,Peal在关于BN研究的专著[1]里定义和讨论了马尔科夫毯(Markov Blanket, MB)。关于BN中的任意结点T,它的MB是唯一的,包括T的所有父、子和配偶(和T有共同子结点的)3类结点(见图1)。MB构成的集合可完全屏蔽掉其他变量对T的"影响",即可完全确定T的值。虽然研究人员较早揭示了该性质,但直到1996年Koller和Sahami(K&S)两位斯坦福大学的学者才将MB和特征子集选择(Feature Subset Selec-

本文受国家自然科学基金资助项目(61305058,61300139),福建省自然科学基金(2014J05074),厦门科技计划基金资助项目(3505Z20133027),华侨大学科研基金资助项目(11Y0274,12HJY18),中央高校基本科研基金资助项目(11J0263)资助。

傅顺开(1978—),男,博士,讲师,主要研究方向为数据挖掘、机器学习,E-mail:fusk@hqu.edu.cn;**苏致祯**(1994—),男,主要研究方向为数据挖掘及其应用;**Sein Minn**(1990—),男,硕士生,主要研究方向为数据挖掘及其应用,E-mail:1525441329@qq.com;**吕天依**(1994—),女,主要研究方向为数据挖掘及其应用。

tion,FSS)关联起来[2]。K&S从信息论角度首次证明了MB是预测目标变量T的最优子集,即不包含无关(Irrelevant)和冗余(Redundant)变量。FSS是数据挖掘和机器学习的重要预处理步骤,可有效降低学习和推导复杂度,还有助于避免过拟合现象。同时,降维也可为数据采集、存储、传输和预处理带来可观的成本节约。理想的FSS方案在兼顾以上目标的同时不以牺牲模型的预测能力为代价。

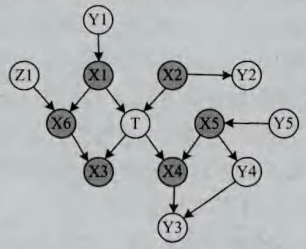

图1 贝叶斯网络和结点T的马尔科夫毯(灰色节点)示例

若已知BN(的结构),根据定义可容易"读取"任一结点T的MB。然而,BN的结构学习属于NP难题[3],一般应用规模限于数十个变量[4]。为了提高学习效率,自1996年起陆续出现的相关学习算法都尽力避免全局学习。截止2014年10月,基于Google Scholar搜索结果的整理和统计显示至少有超过100篇在国内外(主要是国外)重要期刊和会议上发表的相关论文,这些工作可分成以下3大类:

①基于完整的MB集合的条件独立特性。该类算法根据式(2)来设计,包含增长和裁剪两个子步骤。增长阶段里,首先初始化候选MB集合CMB为ϕ;之后将满足关于CMB与T条件相关的变量X持续添加进CMB,直到没有更多的变量可以被添加。由于此时的CMB包含假正(False Positive)元素,需要额外的裁剪过程——从CMB中识别并删除与T关于CMB条件独立的变量X。该类算法的典型代表是I-AMB[5]及其诸多衍生版本[6,7]。它们的最大优势是计算复杂度低,但由于条件独立(Conditional Independence,CI)测试所涉及的条件集合可能为MB甚至更大的集合,实际应用需要大规模的训练样本才能获得较好的学习效果。

②基于有向图中T和MB集合内3类角色结点的连接关系。优先推导和T直接相邻的父、子结点;而后推导和T不直接相邻但与T已知相邻的结点相邻的配偶结点。相比类①中的算法,此类算法虽然也高度依赖CI测试,不过涉及的条件集通常不大(考虑到实际的BN的连接密度一般不高),因此实现同等学习效果所需的样本规模要少得多。代表性算法有MMPC/MB[8],HITON-PC/MB[9],PCMB[10],IPC-MB[11]、MBOR[12]等。

③基于搜索-评分的学习。此类算法视推导任务为模型选择问题,通过事先定义的评分函数和搜索停止策略决定是否继续往前搜索;如果停止条件不满足,则基于当前状态选取可带来局部最优的操作之一:添加、删除或反转边。除了无法达到全局最优,由于尚并不清楚全局网络和局部结构之间的关联(例如全局搜索过程中添加一条边对全局评分有益,但并不直接促进局部推导目标),目前还无法实现较高效的搜索。此类仅有的DMB和RPDMB两个算法在时间效率上并不占优[13],也并非推导通用目标的算法,而是以利于分类任务为目标。

以上3类算法中,目前已知综合性能占优的是第②类[14],本文提出的算法亦属于该类。

2 局部搜索策略和图拓扑信息结合用于马尔科夫毯推导的基本原理

2.1 有向图模型的基本概念和理论

给定问题域$U=\{X_1,\cdots,X_n\}$上的联合分布P,对应的BN是一个二元组,$B=\langle G,\Theta\rangle$;其中$G=\langle V,A\rangle$是关于BN模型的定性描述:$V$(同变量集合$U$)和$A$分别对应有向无环图(DAG)的结点和边集合。对任意节点$X\in V$,G中与其相邻的包括其父、子结点,分别用$Pa(X)$和$Ch(X)$表示;与X有共同子结点的称为X的配偶结点,记为$Sp(X)$;从X出发存在有向路径可达的结点集合为X的后代结点,记为$Des(X)$(包含$Ch(X)$);G中$Des(X)$以外的结点集合统称为非后代结点,记为$NonDes(X)$(包含$Pa(X)$和$Sp(X)$)。以图1中的BN为例,$Pa(T)=\{X_1,X_2\}$,$Ch(T)=\{X_3,X_4\}$,$Sp(T)=\{X_5,X_6\}$,$Des(T)=\{X_3,X_4,Y_3\}$,除$Des(T)$以外的结点均属$NonDes(T)$。Θ为参数集合,包含了如$\Theta_{x|Pa(x)}=P(x\mid Pa(x))$这样的条件概率分布,它定量地给出当父结点集合取值为$Pa(x)$时,它们的共同子节点X取x的条件概率。若$Pa(x)=\emptyset$,参数为先验概率$P(x)$。

对应于BN的联合分布(基于局部结构)可因式分解为:

$$P(X_1,\cdots,X_n)=\prod_{i=1}^{n}P(X_i\mid Pa(X_i)) \tag{1}$$

定义1(忠实(Faithfulness)条件) 一个贝叶斯网络B和一个联合分布P是相互忠实的,当且仅当任意它们蕴含完全一致的条件独立关系,即任意的$X\perp_G Y|Z$有相应的$X\perp_P Y|Z$。

定义2(马尔科夫条件) 若已知父结点$Pa(X)$,BN中的任何结点X条件独立于它的所有非后代结点,即$X\perp_G NonDes(X)|Pa(X)$。当忠实条件满足时,$X\perp_P NonDes(X)|Pa(X)$。

定义3(V-结构) 给定3个结点X、Y和Z,若存在边$X\rightarrow Z$和$Y\rightarrow Z$,但X和Y之间没有边,则称它们构成一个V-结构$X\rightarrow Z\leftarrow Y$。

V-结构$X\rightarrow Z\leftarrow Y$中的结点$Z$称作碰撞(Collider)或汇聚结点。

定义4(封堵的路径) 连接X和Y的(有向)路径称为被集合Z封堵,如果以下任一条件满足:(1)Z中包含该路径上的非碰撞结点;(2)若路径上存在碰撞结点Z,Z以及它的后代结点都不包含在Z中。否则,连接X和Y的(有向)路径为非封堵或开放的。

定义5(d-分割(dependence-separate,or d-separate)) 若结点X和Y之间所有的有向路径都被结点集合Z封堵,则称X和Y被Z所d-分割。相应的Z称为关于X和Y的d-分割集[15]。

定义6(马尔科夫毯) 一个变量$T\in U$的马尔科夫毯记为$MB(T)$,它是满足以下条件的任意变量集合:

$$\forall X\in U\backslash MB(T)\backslash\{T\},T\perp X|MB(T) \tag{2}$$

满足条件的最小马尔科夫毯又称马尔科夫边界(Markov boundary)。当忠实条件满足时,$MB(T)$是唯一的,即亦是马尔科夫边界,包含$Pa(T)$、$Ch(T)$和$Sp(T)$。

d-分割概念桥接了有向无环图概率和图视图下的条件独立关系:(1)$X\perp_P Y|Z$同$X\perp_G Y|Z$是一一对应的;(2)$MB(T)$是T与$U\backslash MB(T)\backslash\{T\}$之间的最小$d$-分割集。

定理 1 给定忠实假设，X 和 Y 相邻当且仅当不存在 Z 使得 $X\perp_P Y|Z$ 且 $X,Y\notin Z$。

即 X 和 Y 相邻当且仅当不存在关于它们的 d-分割集。从定理 1 容易推出如下的引理 1。

引理 1 给定忠实假设，如果存在一个不包含 X 和 Y 的集合 Z 使得 $X\perp_P Y|Z$，那么 X 和 Y 不相邻，记为 $X\neq Y$。

定理 2 给定忠实条件和任意 3 个结点 X、Y 和 Z，如果 X 和 Y 分别和 Z 相邻，但 X 和 Y 不相邻，那么存在 V-结构 $X\rightarrow Z\leftarrow Y$，当且仅当 X 和 Y 关于任意包含 Z 的集合条件相关。

2.2 基于拓扑信息设计的广度优先搜索策略

忠实条件允许进一步将 $MB(T)$ 内的结点区分为：(1)和 T 直接连接的结点($Pa(T)$ 和 $Ch(T)$)；(2)与 T 不直接连接的结点($Sp(T)$)。MMPC/MB[8] 和 HITON-PCM/MB 算法[9]发现并利用了该拓扑信息，提出了比更早出现的 IAMB 更精细的推导策略：

第 1 步：推导和 T 相邻的结点($Pa(T)$ 和 $Ch(T)$，但无法进一步区分它们)；

第 2 步：针对 $\forall X\in Pa(T)\cup Ch(T)$，分别推导与其相邻的结点集合($Pa(X)$ 和 $Ch(X)$)，而配偶结点包含在新发现的结点集合中；

第 3 步：通过一定的 CI 测试识别真正的配偶结点，从而最终完成 MB 的推导。

它们的推导过程应用到了两个经典学习策略：

①分而治之。分别学习 $Pa(T)\cup Ch(T)$ 和 $Sp(T)$，由于 d-分割 T 与 $Y\in Pa(T)\cup Ch(T)$ 或 $Y\in Sp(T)$ 所需的结点集合一般远小于 d-分割 T 与 $U\backslash MB(T)\backslash\{T\}$ 所需的结点集合(后者正是第①类算法(IAMB 等)的判断策略)，学习过程中所依赖的统计测试在基于同样样本大小时能获得更可靠的结果，从而让整个统计学习过程受益。实验结果的确验证了这一点[5]。

②广度优先。由近即远逐层迭代搜索，并将搜索层级限制在两层，规避了全局搜索而不牺牲完整性。实验显示该策略较传统基于 PC 算法学习全局 BN 而后抽取局部 MB 的途径节省了可观的计算[10,11]。

虽然 MMPC/MB 和 HITON-PC/MB 两个算法首先提出并实践了该学习策略，但由于设计存在漏洞并不能保证恒输出正确的结果[8,9]。PCMB[10] 是基于该学习框架的第一个可被证明为正确的算法，而几乎同时出现的 IPC-MB 算法的计算效率较 PCMB 有显著优势[11]。本文提出的算法也将基于该搜索框架以避免全局搜索。

2.3 基于条件独立测试推导关键拓扑信息

遵循 2.2 节介绍的框架所设计的现有算法普遍基于统计测试验证任意变量对是否条件独立，从而决定它们是否相连(相邻)。MMPC/MB、HITON-PC/MB 和 PCMB 基于定理 1 而设计，故需要遍历所有可能的 CI 测试才能确定是否不存在让两个变量条件独立的集合。该策略有两个明显的缺点：(1)容易遗漏(该执行却没有执行的统计测试)，MMPC/MB 和 HITON-PC/MB 在部分场景下无法保证输出正确的结果；(2)计算量大，PCMB 算法虽然完善了 MMPC/MB 和 HITON-PC/MB 算法设计上的漏洞，但过度谨慎的处理导致计算量显著增加。

IPC-MB 算法意识到 PCMB 等算法的缺点，选择基于引理 1 进行设计。首先假设 T 和任意的 $X\in U\backslash T$ 相邻；然后寻找是否存在可以让 T 和 X 条件独立(d-分割)的集合 Z，若存在，则删除边 T-X。与验证“不存在性”任务相比，“存在性”验证相对要容易许多：只需找到一个反例即可。考虑到实际应用中 BN 通常较为稀疏[11]，找到相应 d-分割集的代价不会太高。IPC-MB 同样可被证明是正确的，而解决同样问题时它所需的 CI 测试较 PCMB 显著地减少。

以上两类算法有 3 个共同点：(1)依赖拓扑信息构建学习策略；(2)强依赖 CI 测试；(3)不加区分地对待所有候选 CI 测试，以随机顺序逐个执行以寻找 d-分割集。此处将通过一个简单例子揭示不同结点封堵路径的能力不同，相应地基于不同结点构成的结点集合的 d-分割能力也将不同，量化这个能力差异并加以利用将有助于搜索。

如图 2 所示，若要寻找 X 和 Y 之间是否存在包含一个变量的 d-分割集，候选集合有 $\{Z\}$、$\{M\}$、$\{O\}$、$\{P\}$、$\{N\}$ 和 $\{Q\}$。现有基于约束学习的 MB 推导算法“平等”对待这些集合，极端情况下 $\{Z\}$ 可能在最末个被选择(作为 CI 测试的条件集)，这导致所执行的 6 个 CI 测试中有 5 个是“徒劳”的。如果 $\{Z\}$ 能在首个被选中，通过执行一个 CI 测试即可确定 X 和 Y 之间存在 d-分割集，意味着能节省 83%的计算。图 2 中，$\{Z\}$ 不但是 X 和 Y 的 d-分割集，还是 X 和 M、X 和 N、X 和 O 等的 d-分割集。相比之下，M、Y 和 N 只能分别 d-分割 X 到 O、P 和 Q 的有向路径；而 O、P 和 Q 则没有机会 d-分割任何有向路径。

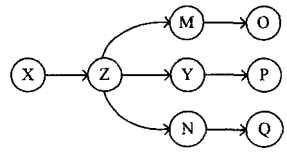

图 2 有向图、d-分割集和所包含的结点示例

倘若一个结点有较多机会独立或参与 d-分割有向路径，这样的结点在未来构建新的 d-分割集时有理由优先被选中。为描述方便，将结点(集)的这个能力定义为 d-分割力，故学习任务演变为(1)量化不同结点的 d-分割力，记为 DSA_X；(2)量化不同结点集的 d-分割力，记为 DSA。任务(1)通过以下公式解决：

$$DSA_X=\sum_{DS}\frac{1}{|DS|} \tag{3}$$

其中，DS 表示有 X 参与的 CI 测试且最终确定相应的条件集 DS 为 d-分割集。具体更新规则为(1)一旦通过 CI 测试发现新的 d-分割集，将基于式(3)分别更新集合中每个结点 X 的 d-分割力；(2)同一个 d-分割集中的结点将被平等对待，分别获得 $1/|DS|$ 的值。可见，同一个结点出现在小分割集中时获得的 d-分割力增值将多于出现在大分割集中获得的增值；一个结点的绝对 d-分割力大小取决于所累计获得的能力值。

任务(2)则基于以下公式来解决：

$$DSA=\sum_{X\in DS}DSA_X \tag{4}$$

即通过累加 d-分割集 DS 中各结点目前的 d-分割力值获得 DS 的 d-分割力。

3 基于局部与自适应搜索的马尔科夫毯推导算法描述

根据 2.2 节和 2.3 节中描述的启发式策略，提出可快速

推导马尔科夫毯的新算法 FSMB(Fast Search of Markov Blanket,见算法 1)。该算法继承了 PCMB 和 IPC-MB 的优点,基于已知的拓扑信息将搜索范围严格限制而规避了全局搜索。此外,FSMB 在搜索过程中利用已执行的 CI 测试所找到的 d-分割集动态更新各结点的 d-分割力(算法 2 的第 10 行),并根据各结点的 d-分割力量化候选 d-分割集的 d-分割力(算法 3);高 d-分割力的集合将在算法 2 的第 6 行优先被考虑(由于排在算法 3 返回的优先队列的前部)。

算法 1 FSMB

输入:T:目标变量
D:数据集/训练集
ε:显著性门限值
输出:MB(T)
1. CADJ←U\{T};//候选相邻集
2. PCD(T)←RecognizePCD(T,CADJ,D,ε);//候选父子
3. MB←Ø;//马尔科夫集
4. FOR(∀X∈PCD(T)) DO
5. CADJ←U\{X}
6. PCD(X)←RecognizePCD(X,CADJ,D,ε);
7. IF(T∈PCD(X))THEN //发现真父子结点
8. MB←MB∪{X};
9. FOR(∀Y∈PCD(X) AND Y∉MB) DO
10. IF(I_D(T,Y|$DS_{(T,Y)}$∪{X})>ε) THEN
11. MB←MB∪{Y};//发现新配偶结点
12. RETURN MB;

算法 2 RecognizePCD

输入:T:目标变量
CADJ:候选邻居
D:数据集/训练集
ε:显著性门限值
输出:T 的候选父子集合(可能包含后代结点)
1. NotPC←Ø;
2. dss←0;//初始化 d-seperator size
3. DO
4. FOR(∀X∈CADJ) DO
5. SS=GenerateSS(CADJ\{X},dss,DSA);
6. FOR(∀S∈SS) DO
7. IF(I_D(T,X|S)≤ε) THEN
8. NotPC←NotPC∪{X};
9. $DS_{(T,X)}$←S;//缓存供算法 1 中使用
10. ∀Y∈S,DSA_Y←DSA_Y+1/|S|;
11. BREAK;//跳转到 4
12. IF(|NotPC|>0) THEN
13. CADJ←CADJ\NotPC;
14. NotPC←Ø;
15. dss←dss+1;
16. WHILE(|CADJ|>dss)
17. RETURN CADJ;

算法 3 GenerateSS

输入:X:结点集合
dss:候选分割集大小
DSA:结点出现在分割集中的统计频次
输出:X 中大小为 dss 的所有候选集的有序队列
1. PRQ←Ø; //初始化空优先队列
2. FOR(∀S⊆X AND |S|=dss) DO
3. 计算$\sum_{Y\in S}DSA_Y$,将 S 插入 PRQ;//保持得分高低
4. RETURN PRQ;

FSMB 将分别推导目标结点的父/子结点与配偶结点,而配偶结点的推导也是依赖对优先推导所得的候选父/子结点进一步搜索其父/子结点而得。不过针对这两类结点的推导应用了截然不同的搜索策略:

① 后向选择(Backward Selection)。在推导相邻结点时,首先假设 T 与 $\forall X\in U\backslash\{T\}$ 相连,而后通过检查 T 与 X 之间是否存在 d-分割集来决定是否删除相应的边。算法 2 的 RecognizePCD 即基于后向搜索,该算法是整个学习过程的重要步骤(出现在算法 1 的第 2 行和第 6 行),绝大部分的统计测试发生在该步骤。

② 前向选择(Forward Selection)。配偶结点的推导(算法 1 的第 4—11 行)基于前向搜索,根据定理 2 直接从候选配偶结点中识别真配偶。

4 算法正确性证明

所有的证明基于两个(假设)条件:(1)忠实性;(2)正确的 CI 测试。根据 MB 的定义,将分别从 3 个方面证明输出的 $MB(T)$(1)包含所有父、子结点(引理 3);(2)包含所有配偶结点(引理 4);(3)不包含其他结点(引理 5)。

引理 2 给定 $X\in U$,RecognizePCD 除了输出 X 的父、子结点,还可能输出部分 X 的后代结点。

证明:RecognizePCD 首先将 T 和所有 $U\backslash\{T\}$ 相连。(1)首先证明 T 和任意的 $X\in PC(T)$ 之间的边 T-X 不会被误删。假设存在这样的边 T-X 被误删除,即存在 T 和 X 的 d-分割集(第 7—11 行)。这与定理 1 相悖,故 T 和真正相邻结点的边不会被误删,这样的相邻结点将始终保留在 $CADJ$ 集合中。(2)其次证明所有的非后代结点将从 $CADJ$ 集合中被删除。根据(1)可知所有的 $PC(T)$ 将始终包含在 $CADJ$ 中,故其自然包含所有的父结点 $Pa(T)$,根据定理 2,当 $dss==|Pa(T)|$ 时,将有机会从 $CADJ$ 中删除所有的非后代结点。(3)最后证明存在后代结点有可能无法被删除的场景。如图 3 中针对 T 的调用 RecognizePCD,当 $dss==0$ 时,结点 P 和 R 被识别为 $NotPC$,并在本次迭代结束时从 $CADJ$ 中删除,即 $CADJ=\{Q,S\}$;当 $dss==1$ 时,S 本应该被识别为 $NotPC$,但 d-分割 S 和 T 的最小结点集合应为 $\{Q,P\}$ 或 $\{Q,R\}$,而目前的 $CADJ$ 中不包含 P 或 R。所以本迭代中没有结点从 $CADJ$ 中被删除。由于 $dss==2==|CADJ|$,即没有更多的 CI 测试可执行,RecognizePCD 结束执行,将 $\{Q,S\}$ 返回给 FSMB。证毕。

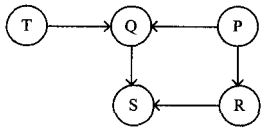

图 3 T 的后代结点可能无法被 RecognizePCD 识别的示例

引理 3 FSMB 能正确推导 T 的父、子结点。

证明:(1)由引理 2 可知,FSMB 第 2 行调用 RecognizePCD 后输出的 $PCD(T)$ 将包含所有 T 的父、子结点,但也可能包含 T 的后代结点。(2)FSMB 算法第 4—11 行将对 $\forall X\in PCD(T)$ 调用 RecognizePCD。若 $X\in PC(T)$,$T\in PCD(X)$,这样的 X 将被添加进 $MB(T)$(算法 1 第 8 行);若 $X\in Des(T)$,由于 $T\notin PCD(X)$,这样的 X 将没有机会进入 $MB(T)$。故 FSMB 能正确推导 $PC(T)$。

引理 4 FSMB 能正确推导 T 的配偶结点。

证明：(1)首先证明满足算法 1 第 10 行条件的只可能是 T 的配偶结点。根据引理 3 的证明可知，满足算法第 7 行 $T\in PCD(X)$的 X 只可能是 T 的父、子结点。(1a)若 X 是 T 的父结点。对于任意的 $Y\in PCD(X)$且 $Y\notin MB(T)$，$DS_{(T,S)}$ 是 T 和 Y 的 d-分割集。易证 $X\in DS_{(T,Y)}$，否则 Y 和 T 之间的有向路径 $Y\to X\to T$、$Y\leftarrow X\to T$ 或 $Y\leftarrow\cdots\leftarrow X\to T$ 可能将无法被 d-分割而与事实相反。由于 $X\in DS_{(T,Y)}$，算法第 10 行 CI 测试无法通过，因此排除 X 是 T 的父结点的可能。(1b)若 X 是 T 的子结点，且 $Y\in PCD(X)$是 X 的子结点或后代结点。同(1a)可证 $X\in DS_{(T,Y)}$，故也无法满足第 10 行 CI 测试，排除 T 的子结点的子结点或后代结点的可能。(1c)若 X 是 T 的子结点，且 $Y\in PCD(X)$是 X 的父结点。当将 X 添加入 $DS_{(T,Y)}$后，有向路径 $Y\to X\leftarrow T$ 将不再隔断，从而 $DS_{(T,Y)}\cup\{X\}$不再 d-分割 Y 和 T。因此，满足算法第 10 行 CI 测试的只可能是 T 的子结点的父结点，即 T 的配偶结点。

(2)FSMB 将对所有 T 的父、子结点执行第 7－11 行，即不会漏掉任一个 T 的配偶结点。故 FSMB 能正确推导 $Sp(T)$。证毕。

引理 5 FSMB 不会输出非 $MB(T)$的结点。

证明：(1)引理 3 确保 FSMB 不会输出假的父、子结点。(2)引理 4 确保只有真正的配偶结点才可能出现在 FSMB 的输出中。证毕。

定理 3 FSMB 能正确推导 $MB(T)$。

证明：根据引理 3－引理 5 易证。

5 实验结果与分析

5.1 实验设置

从以下 3 个不同的维度分别检验新算法的性能。

①面向知识发现任务的学习效果。知识发现任务关注每个结点，无论是强关联还是弱关联，而 F-measure 由于综合了准确率和召回率而适合衡量这样的学习效果。该实验将基于两个真实网络(Asia 和 Alarm)以及两个基于 Weka 的 BN 模块随机生成的两个网络(Test100 和 Test200)，网络拓扑的更多信息见表 1。给定每个网络，将随机生成 1000～5000 大小不等的样本集，每种尺寸分别生成 10 份供实验所需。

②面向预测任务的学习效果。MB 作为最优特征子集，当应用到预测任务时最直接的评估是基于学习的特征子集训练一个分类器，并与基于全部特征集合训练的分类器比较预测准确率。该实验将基于 16 个 UCI 用于分类实验的经典数据集，并选择多个公认高效的分类器。

③学习效率。学习效率是基于 CI 测试的加权统计[16]。针对一个 CI 测试 $I_D(X,Y|Z)$，它的权重为 $2+|Z|$，反映了统计测试复杂度与涉及的变量(数)成正比。而一个算法的复杂度则为整个学习过程所执行的所有这样的 CI 测试的权重之和。这是基于 CI 测试的统计学习算法的标准比较方法，其最大优点是机器无关性。该实验也将基于表 1 所列的网络。

表 1 已知拓扑的 4 个贝叶斯网络

数据集	节点数	边数	变量取值
Asia	8	8	2
Alarm	37	46	2～4
Test100	100	120	2
Test200	200	250	2

新算法将通过与 PC 算法的对比来验证局部搜索的有效性；将与同属局部搜索的 IAMB、PCMB 和 IPC-MB 进行对比来验证新的启发式策略的优势。

所有算法和实验基于 Weka 框架实现，而面向预测的学习任务所涉及的经典分类器和非 MB 特征选择方法则是直接利用 Weka 已实现的版本。

5.2 面向知识发现任务的学习效果

给定 4 个已知网络拓扑的网络，将分别以每个结点为目标结点，应用 4 个基于局部搜索的算法学习其相应的 MB。从表 2 的实验结果可以得出：(1)当有更多的样本可用于学习时，算法能够获得更优的结果，除了 IAMB。这归结于 IAMB 的固有缺陷即学习策略相对“粗糙”，需要大量的样本才能保证可信的 CI 测试[17]；(2)PCMB、IPC-MB 和 FSMB 的学习效果相当；(3)面对较大网络时，PCMB、IPC-MB 和 FSMB 较 IAMB 有显著的优势。可见，FSMB 所采用的新策略可以保证有效推导。

表 2 基于 F-measure 的知识发现效果比较

数据集	样本数	IAMB	PCMB	IPC-MB	FSMB
Asia	1000	0.71	0.68	0.69	0.70
	2000	0.80	0.78	0.80	0.79
	3000	0.78	0.77	0.78	0.78
	4000	0.80	0.78	0.80	0.80
	5000	0.81	0.85	0.85	0.85
Alarm	1000	0.55	0.85	0.87	0.87
	200	0.55	0.93	0.94	0.94
	300	0.51	0.94	0.95	0.95
	400	0.51	0.95	0.96	0.96
	500	0.50	0.96	0.97	0.97
Test100	1000	0.60	0.69	0.69	0.70
	2000	0.61	0.78	0.78	0.78
	3000	0.64	0.79	0.81	0.82
	4000	0.60	0.83	0.83	0.84
	5000	0.61	0.84	0.85	0.85
Test200	1000	0.39	0.51	0.52	0.52
	2000	0.43	0.58	0.59	0.59
	3000	0.44	0.62	0.63	0.63
	4000	0.45	0.65	0.66	0.66
	5000	0.47	0.67	0.68	0.68

5.3 面向预测任务的学习效果

基于 UCI 的 16 个经典分类数据集(过滤掉特征数目少于 10 个的数据集)，并选择朴素贝叶斯(NB)[18]、隐朴素贝叶斯(HNB)[19]、决策树(DT)和 k 近邻(kNN) 4 个经典分类器，分别基于全部特征(表 3 中的 *(XX)行，其中 XX 对应的数字为特征数目)和(经 FSMB 和 CB-BF 算法所输出的)部分特征集训练分类器，继而通过 10-folder 实验方法获得它们的平均预测准确率。从表 3 中的实验结果可以得出：(1)基于 FSMB 的特征子集选择是有效的，反映在基于其所输出的特征集合所训练的分类器能够实现接近甚至超过基于所有集合训练的模型的分类准确率(表 3 中的阴影单元格表示基于 FSMB 输出所训练模型的性能不亚于基于全特征集所训练的模型)；(2)与非 MB 特征子集选择方法 CB-BF 比较，FSMB 在多个例子中显示出明显的优势。例如，给定 Anneal 数据集，FSMB＋HNB 可达到 97.26%的准确率，而 CB-BF＋HNB 的准确率却只有 74.11%；同样的情况还出现在 Ionosphere、Sonar、Vowel 等数据集上。

表 3　基于预测准确率间接评价特征选择效果

数据集	方法	NB	HNB	DT	kNN
Anneal	*(39)	93.24	97.48	97.90	97.94
	FSMB	93.30	97.26	97.63	97.83
	CB-BF	92.70	74.11	93.94	94.27
Cleveland-14-heart-disease	*(14)	83.74	83.57	77.99	78.00
	FSMB	82.28	83.86	81.07	79.51
	CB-BF	85.28	84.03	82.55	81.16
Hungarian-14-heart-disease	*(14)	83.81	84.32	81.81	82.16
	FSMB	84.74	85.55	81.41	82.52
	CB-BF	85.08	84.64	80.66	84.87
Heat-statlog	*(14)	84.74	84.00	77.96	77.78
	FSMB	85.11	84.07	79.96	80.56
	CB-BF	84.41	84.93	78.78	78.30
Hepatitis	*(20)	85.28	85.47	80.35	81.72
	FSMB	87.02	84.60	82.28	82.83
	CB-BF	85.67	84.45	82.53	85.58
Hypothyroid	*(30)	91.41	92.63	92.84	92.00
	FSMB	91.06	92.75	92.84	92.35
	CB-BF	89.78	92.77	92.81	92.83
Ionosphere	*(35)	88.61	92.45	89.63	90.94
	FSMB	90.72	93.34	90.00	91.79
	CB-BF	90.71	89.43	88.83	90.63
Labor	*(17)	93.87	93.73	84.97	89.57
	FSMB	94.60	95.77	82.10	91.07
	CB-BF	93.30	93.97	81.80	90.53
Lymphography	*(19)	84.42	84.22	78.35	80.55
	FSMB	84.36	82.53	78.82	83.68
	CB-BF	82.43	82.77	79.77	81.84
Anneal-ORIG	*(39)	84.38	87.85	89.83	88.46
	FSMB	84.38	87.85	88.46	89.83
	CB-BF	80.06	82.46	78.24	82.18
Primary-tumor	*(18)	47.20	47.85	41.01	38.59
	FSMB	46.79	46.81	40.27	38.61
	CB-BF	46.70	47.23	41.04	39.86
Sick	*(30)	91.99	93.25	93.88	93.88
	FSMB	93.90	94.01	93.88	93.80
	CB-BF	93.85	93.81	93.88	93.57
Sonar	*(61)	72.87	79.51	74.17	83.05
	FSMB	76.42	80.88	76.25	82.66
	CB-BF	76.81	79.15	72.76	74.74
Vote	*(17)	90.21	94.36	96.27	93.49
	FSMB	94.41	96.18	95.29	95.91
	CB-BF	94.76	96.00	95.29	94.67
Vowel	*(14)	59.16	90.65	77.08	92.30
	FSMB	63.18	83.88	76.61	90.25
	CB-BF	56.31	70.23	70.11	74.89
Zoo	*(18)	93.30	97.81	92.61	96.15
	FSMB	93.49	97.61	92.61	96.15
	CB-BF	94.09	94.77	93.08	96.04

注：表中为基于全部特征集合训练的分类器(*)与基于 FSMB 和 CB-BF 输出的特征子集训练的分类器的比较

5.4　基于统计 CI 测试数的学习效率

实验结果如表 4 所列，可以得出：(1)PCMB 算法复杂度最高，IAMB 最低；(2)FSMB 较 PCMB 有显著的优势，比如基于 Alarm 网络的实验中，FSMB 较 FSMB 能节省近 98%的计算量；(3)FSMB 和 IPC-MB 的总体架构相似，但新的学习策略同样让 FSMB 较 IPC-MB 有明显的优势。例如基于 Test200 网络的实验中，FSMB 所需的计算量仅是 IPC-MB 的 50%；(4)IAMB 算法公认是目前最快速的算法(虽然学习效果很差)，但在处理大网络时(Test100 和 Test200)，FSMB 的学习效率甚至超越了 IAMB(表 4 中阴影单元格)。考虑到 FSMB 的学习效果远远优于 IAMB，FSMB 的综合优势则显得十分突出。

表 4　基于加权 CI 测试的计算效率比较

数据集	样本数	IAMB	PCMB	IPC-MB	FSMB
Asia	1000	81	1940	359	218
	2000	75	1764	357	222
	3000	75	1820	388	260
	4000	81	1870	411	269
	5000	79	2155	439	292
Alarm	1000	496	31449	3036	1798
	200	633	47787	3612	2158
	300	754	67036	3945	2360
	400	801	82510	4343	2540
	500	868	102298	4631	2721
Test100	1000	894	7452	1776	979
	2000	1382	9463	2148	1152
	3000	1381	9979	2281	1207
	4000	1964	11049	2437	1303
	5000	1959	12112	2545	1365
Test200	1000	5335	17303	4508	2395
	2000	6894	20048	5181	2703
	3000	8636	21880	5578	2890
	4000	8636	22938	5828	2979
	5000	8636	24500	6143	3137

结束语　马尔科夫毯属于数据挖掘和机器学习的重要概念，也有诸多关于有效推导马尔科夫毯的工作被提出。目前为止，基于拓扑信息的约束学习类算法占明显主导，这得益于其更佳的实用效果。提出一种新的推导算法 FSMB，挖掘并利用了更多的拓扑信息，在保证理论正确的前提下显著提高了学习效率。未来的工作将继续探索并利用拓扑信息以进一步提高学习效率。另外，相关的学习策略也被应用到贝叶斯网络的结构学习任务上，且被证明同样有效。

参 考 文 献

[1] Pearl J. Probabilistic reasoning in expert systems[M]. San Matego: Morgan Kaufmann, 1988

[2] Koller D, Sahami M. Toward optimal feature selection[C]// the 13th International Conference on Machine Learning (ICML). Bari, Italy: Morgan Kaufmann, 1996

[3] Chickering D M, Geiger D, Heckerman D. Learning Bayesian Network is NP-Hard[R]. Microsoft Research, 1994

[4] Campos C P D, Zeng Z, Ji Q. Efficient structure learning of Bayesian networks using constraints[J]. Journal of Machine Learning Research(JMLR), 2011, 12(11): 663-689

[5] Tsamardinos I, Aliferis C, Statnikov A, et al. Algorithms for large scale Markov blanket discovery[C]// 16th International FLAIRS Conference, 2003. AAAI, 2003

[6] Zhang Y, Zhang Z, Liu K, et al. An improved IAMB algorithm for Markov blanket discovery[J]. Journal of Computers, 2010, 5(11): 1755-1761

[7] Zhang Y, Xu H, Huang Y, et al. S-IAMB algorithm for Markov blanket discovery[C]// Asia-Pacific Conference on Information Processing(APCIP'09). Washington: IEEE Computer Society

[8] Tsamardinos I, Aliferis C F, Statnikov A. Time and sample efficient discovery of Markov blankets and direct causal relations [C]// 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD). ACM, 2003

[9] Aliferis C, Tsamardinos I, Statnikov A. HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection[C]// Annual Symposium on American Medical Informatics Association (AMIA). 2003

[10] Pena J M, Nilsson R, Bjorkegren J, et al. Towards scalable and

data efficient learning of Markov boundaries[J]. International Journal of Approximate Reasoning,2007,45(2):211-232

[11] Fu S,Desmarais M C. Fast Markov blanket discovery algorithm via local learning within single pass[C]//21st Conference of the Canadian Society for Computational Studies of Intelligence(Canadian AI). Springer,2008

[12] Zeng Y X,Xiang H Y,Mao H. Dynamic ordering-based search algorithm for Markov blanket discovery[C]//15th Pacific-Asia Conference on Data Mining,2011. Shenzhen,China:Springer,2011

[13] Acid S,De Campos L M,Castellano J G. Learning Bayesian network classifiers:Searching in a space of partially directed acyclic graphs[J]. Machine Learning,2005,59(3):213-235

[14] Fu Shun-kai, Minn M C D S,Lv Tian-yi. A Survey of Advances in Feature Selection by Markov Blanket[C]// ICNC-FSKD, 2014. Xiamen,China,2014

[15] Koller D,Friedman N. Probabilistic graphical models:Principles and Techniques[M]. MIT Press,2009

[16] Bromberg F, Margaritis D, Honavar V. Efficient Markov network structure discovery using independence tests[J]. Journal of Artificial Intelligence Research,2009,35(1):449-484

[17] Fu S, Desmarais M C. Tradeoff analysis of different Markov blanket local learning approaches[C]//12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD). Osaka,Japan:Springer,2008

[18] Duda R O,Hart P E. Pattern Classification and Scene Analysis [M]. John Wiley & Sons,1973-2-9

[19] Zhang H,Jiang L,Su J. Hidden Naive Bayes[C]//AAAI. 2005

(上接第 21 页)

对 CDE 算法中的参数进行初始化:最大进化代数 $g_{max}=100$,种群规模 $Np=40$,其他参数依据前文设定。SVM 惩罚参数 C 和核参数 σ 均设置在$[2^{-10},2^{15}]$。最终的仿真结果如图 2—图 4 所示。

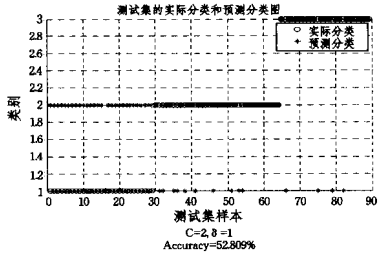

图 2 基于 SVM Wine data set 数据分类图

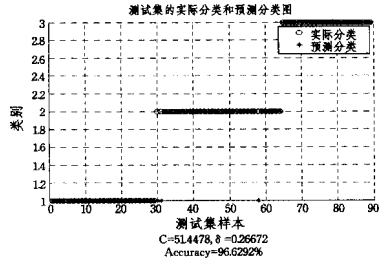

图 3 基于 DE-SVM Wine data set 数据分类图

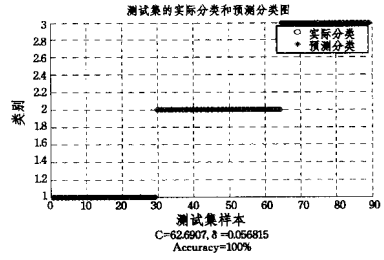

图 4 基于 CDE-SVM Wine data set 数据分类图

通过图 2 可看出,在不使用任何算法对 SVM 进行参数优化(默认参数)的前提下,SVM 对 Wine data set 的分类存在随机性而且效果比较差;图 3 中经过 DE 优化的 SVM 分类准确率达到了 96.63%,但仍有个别样本点未能正确分类;而图 4 中基于 CDE-SVM 模型的分类准确率达到了 100%,所有样本点全部得到了准确分类,由此可以看出 CDE 算法在数据分类中也同样体现出了有效性和优越性。

结束语 本文将克隆选择算法引入差分进化算法中,提出了基于克隆选择的差分进化算法,并将该算法应用于支持向量机的参数优化中。实验结果表明,基于克隆选择的差分进化算法不仅能有效避免早熟收敛现象,而且具有较强的寻优能力,同时经过 CDE 算法优化过的 SVM 分类精度得到了很大的增强,进一步提高了 SVM 的学习和泛化能力。

参 考 文 献

[1] 赵海洋,徐敏强,王金东. 改进二叉树支持向量机及其故障诊断方法研究[J]. 振动工程学报,2013(5):764-770

[2] 彭光金,司海涛,俞集辉,等. 改进的支持向量机算法及其应用[J]. 计算机工程与应用,2011,47(18):218-211

[3] 于明,艾月乔. 基于人工蜂群算法的支持向量机参数优化及应用[J]. 光电子·激光,2012,23(2):374-378

[4] 庄严,白振林,许云峰. 基于蚁群算法的支持向量机参数选择方法研究[J]. 计算机仿真,2011,28(5):216-219

[5] Das S, Suganthan P N. Differential evolution: a survey of the state-of-the-art[J]. IEEE Transactions on Evolutionary Computation,2011,15(1):4-31

[6] Angira R,Babu B V. Optimization of process synthesis and design problems:a modified differential evolution approach [J]. Chemical Engineering Science,2006,61(14):4707-4721

[7] Babu B V,Angira R. Modified differential evolution (MDE) for optimization of nonlinear chemical processes [J]. Computer and Chemical Engineering,2006,30(6):989-1002

[8] de Castro L N, Tmanis J. Artificial Immune Systems: A New Computational Intelligence Approach [M]. British: Springer Press,2002

[9] Leandro N de C,Fernando J Von Z. Learning and optimization using the clonal selection principle[J]. IEEE Transactions on Evolutionary Computation,2002(3):239-251

[10] 张向荣,焦李成. 基于免疫克隆选择算法的特征选择[J]. 复旦学报(自然科学版),2004,43(5):926-929

[11] 王俊,田玉玲. 用于入侵检测的动态克隆选择算法的研究[J]. 计算机与数字工程,2010(6):108-110

[12] 刘倩,仇宾. 基于克隆选择算法的花卉图像分割[J]. 计算机工程与应用,2012,48(14):185-189

[13] 徐佳,张卫. 人工免疫系统中的抗体生成与匹配算法[J]. 计算机工程,2010,36(9):181-183

[14] 胡超杰,章兢. 一种采用克隆选择的免疫差分进化算法[J]. 计算机应用研究,2013,30(6):1640-1642