The max-min-hill-climbing algorithm

Michael Bauer

M.Sc. Comp. Science

June 3, 2014

Algorithm 1 $\overline{MMPC}$ Algorithm

procedure $\overline{MMPC}$ $(T,\mathcal{D})$
Input: target variable T; data $\mathcal{D}$
Output: the parents and children of T in any Bayesian network faithfully representing the data distribution
%Phase I: Forward
$\mathbf{CPC} = \emptyset$
repeat
$\langle F, assocF \rangle = MaxMinHeuristic(T; \mathbf{CPC})$
if $assocF \neq 0$ **then**
$\mathbf{CPC} = \mathbf{CPC} \cup F$
end if
until CPC has not changed

%Phase II: Backward
for all $X \in \mathbf{CPC}$ **do**
if $\exists \mathbf{S} \subseteq \mathbf{CPC}$, s.t. $Ind(X; T|\mathbf{S})$ **then**
$\mathbf{CPC} = \mathbf{CPC} \setminus \{X\}$
end if
end for

return CPC
end procedure

procedure MaxMinHeuristic(T,**CPC**)
Input: target variable T; subset of variables **CPC**
Output: the maximum over all variables of the minimum association with T relative to **CPC**, and the variable that achieves the maximum
$assocF = \max_{X \in \mathcal{V}} MinAssoc(X; T|\mathbf{CPC})$
$F = \arg\max_{X \in \mathcal{V}} MinAssoc(X; T|\mathbf{CPC})$
return $\langle F, assocF \rangle$
end procedure

Definition

We define the minimum association of X and T relative to a feature subset $\mathbf{Z}$, as

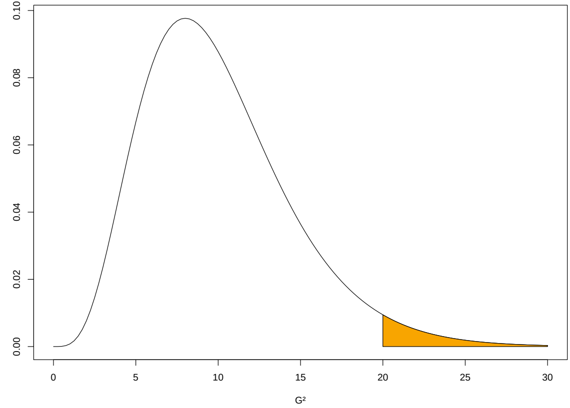
$$MinAssoc(X; T|\mathbf{Z}) = \min_{\mathbf{S} \subseteq \mathbf{Z}} Assoc(X; T|\mathbf{S}). \tag{1}$$

Definition

We calculate the G^2 value under the nullhypothesis of the conditional independence of $Ind_P(X_i, X_j | \mathbf{X}_k)$ holding. Then, the G^2 value is defined as:

$$G^2 := 2 * \sum_{a,b,c} S_{ijk}^{ab\mathbf{c}} * ln \left(\frac{S_{ijk}^{ab\mathbf{c}} * S_k^{\mathbf{c}}}{S_{ik}^{a\mathbf{c}} * S_{jk}^{b\mathbf{c}}} \right), \tag{2}$$

where S_{ijk}^{abc} is the number of times in the data where $X_i = a$, $X_j = b$ and $\mathbf{X}_k = \mathbf{c}$. We define in a similar fashion S_{ik}^{ac}, S_{jk}^{bc} and S_k^c, respectively.



$\alpha = 0.05$

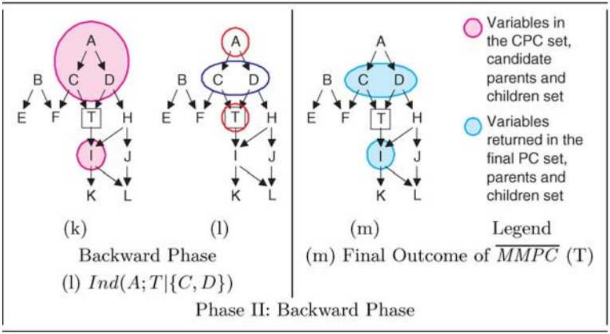
Algorithm 1 $\overline{MMPC}$ Algorithm

procedure $\overline{MMPC}$ $(T,\mathcal{D})$
    Input: target variable T; data $\mathcal{D}$
    Output: the parents and children of T in any Bayesian network faithfully representing the data distribution
    %Phase I: Forward
    $\mathbf{CPC} = \emptyset$
    repeat
        $\langle F, assocF \rangle = MaxMinHeuristic(T; \mathbf{CPC})$
        if $assocF \neq 0$ **then**
            $\mathbf{CPC} = \mathbf{CPC} \cup F$
        end if
    until CPC has not changed

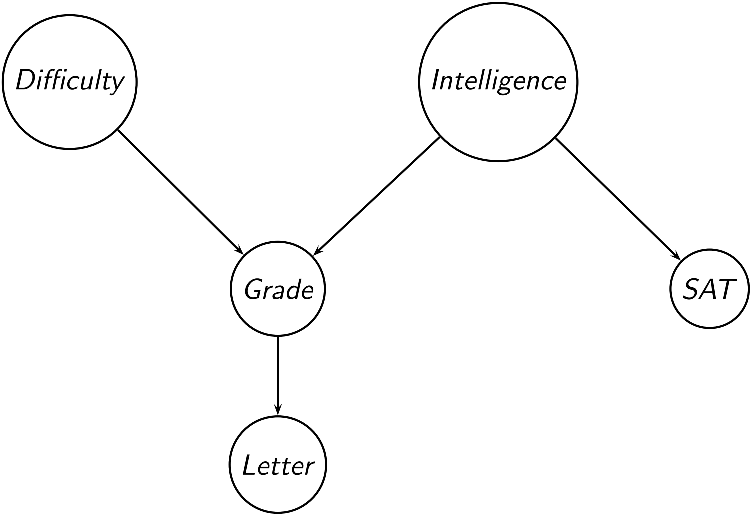
    %Phase II: Backward
    for all $X \in \mathbf{CPC}$ **do**
        if $\exists \mathbf{S} \subseteq \mathbf{CPC}$, s.t. $Ind(X; T|\mathbf{S})$ **then**
            $\mathbf{CPC} = \mathbf{CPC} \setminus \{X\}$
        end if
    end for

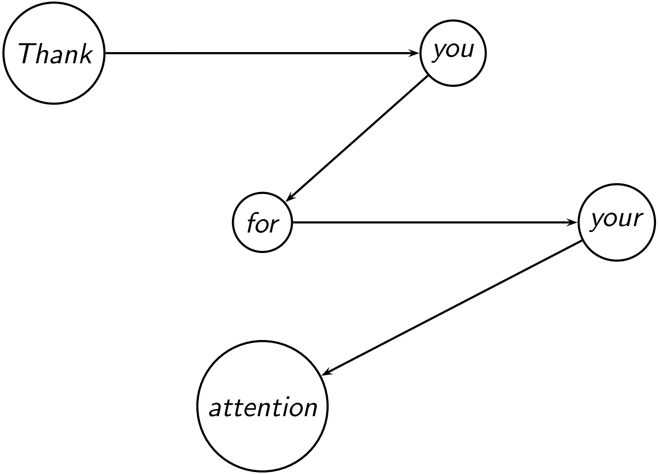
    return CPC
end procedure

procedure MAXMINHEURISTIC$(T,\mathbf{CPC})$
    Input: target variable T; subset of variables **CPC**
    Output: the maximum over all variables of the minimum association with T relative to **CPC**, and the variable that achieves the maximum
    $assocF = \max_{X \in \mathcal{V}} MinAssoc(X; T|\mathbf{CPC})$
    $F = \arg\max_{X \in \mathcal{V}} MinAssoc(X; T|\mathbf{CPC})$
    return $\langle F, assocF \rangle$
end procedure


A
B
C
D
E
F
T
H
I
J
K
L
Variables in the CPC set, candidate parents and children set
Variables returned in the final PC set, parents and children set
(k)
(l)
(m)
Legend
Backward Phase
(l) $Ind(A; T|\{C, D\})$
(m) Final Outcome of $\overline{MMPC}$ (T)
Phase II: Backward Phase

The last goal:




Thank
you
for
your
attention