A User Manual for the Software Tools in Book "Practical approaches to causal relationship exploration"

Jiuyong Li, Lin Liu, and Thuc Duy Le

May 4, 2015

In this document we provide some examples of using software tools for running the four algorithms presented in the book entitled "Practical approaches to causal relationship exploration" [1], http://www.springer.com/computer/ai/book/978-3-319-14432-0.

1 An example data set and the ground truth

To demonstrate how to apply the four methods presented in the book (PC-simple, HITON-PC, CR-PA, and CR-CS), we generated a synthetic data set for the Bayesian network in Figure 1. We selected Z as the target variable, and therefore the parent and children set of Z is B, C, F. The data set was generated using the TETRAD software downloaded from *http://www.phil.cmu.edu/tetrad/*. The data set is in csv format (Example21.csv) in the cases of PC-simple and HITON-PC, and in C4.5 format (Example21.names and Example21.data) in the cases of CR-PA and CR-CS. The files can be downloaded from the home page of the book at: http://nugget.unisa.edu.au/Causalbook.

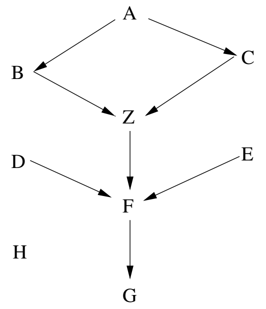


Figure 1: An example Bayesian network

2 Using PC-simple

Assume that the data set *Example21.csv* has been downloaded and stored in the *R* working directory. We now use the *pcSelect* function to find the parents and children set of the target Z as shown below.

```
> library(pcalg)
> data=read.csv("Example21.csv", header=TRUE, sep=",")
> pcSimple.Z=pcSelect(data[,9], data[,-9], alpha=0.01)
> pcSimple.Z
$G
    A     B     C     D     E     F     G     H
FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE
$zMin
[1]   1.85790685 144.57474808  16.65675448   0.27538423
[5]   0.76391512  15.94735566   0.07287629   0.23537966
```

The parents and children set of Z is $\{B, C, F\}$, which is consistent with the graph in Figure 1. We can also verify this result against the global causal structure (Figure 2) learned by the PC algorithm [3, 2]. The DAGs in Figures 1 and 2 have the same skeleton and Z has the same **PC** set in both Figures. The following codes are used to learn the causal structure from the *Example21.csv* data set using the PC algorithm.

```
> pc.example=pc(suffStat=list(dm=data, adaptDF=FALSE),
  indepTest=binCItest,alpha=0.01,labels=colnames(data))
> library(Rgraphviz)
> plot(pc.example)
```

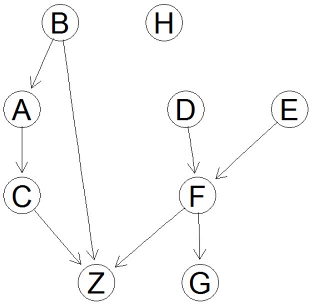


Figure 2: The causal structure learned by the PC algorithm using the data set generated based on the Bayesian network in Figure 1

3 Using HITON-PC

We now use HITON-PC to learn the local causal structure around node Z. The following code and screen output illustrate how to use HITON-PC to learn the local causal structure around the target node Z

```
> library(bnlearn)
> data=read.csv("Example21.csv", header=TRUE, sep=",")
> data[1:5,]
  A B C D E F G H Z
1 1 1 1 0 0 1 1 1 1
2 1 1 1 0 0 1 0 1 1
3 1 0 1 1 0 1 1 1 1
4 1 0 0 0 0 0 0 1 0
5 1 0 1 0 0 1 1 1 1
> for(i in 1:9){data[,i] = as.factor(data[,i]) }
#bnlearn requires numeric or factor data types.
#the above command converts data of the nine variables(nine columns)
#in the data set to factor data types.

> hiton.pc.Z=learn.nbr(data, ’Z’, method=’si.hiton.pc’,
                        alpha=0.01)
> hiton.pc.Z
[1] "B" "F" "C"
```

The parents and children set of Z is $\{B, F, C\}$, which is consistent with the ground truth shown in Figure 1.

4 Using CR-PA

In this section, we use the data set in C4.5 format, which are stored in the two files *Example21.data* and *Example21.names*. With this data set, we can follow the steps below to run CR-PA.

1. Start CRE by double-clicking on the *CRE.jar* file. Note that Java SE runtime environment (jre7 or later) needs to be installed on the computer before running CRE. The main user interface of CRE is shown in Figure 3.

2. Open the input file. Select on the menu bar, File → Open File. Select the input *.names* file stored on local computer, i.e. *Example21.names* in this case.

3. Select the algorithm CR-PA from the left panel.

4. Set parameters for CR-PA. On the menu bar, select Configuration → Parameters. A pop-up window will appear for setting CR-PA parameters as shown in Figure 4. There are four parameters as follows:

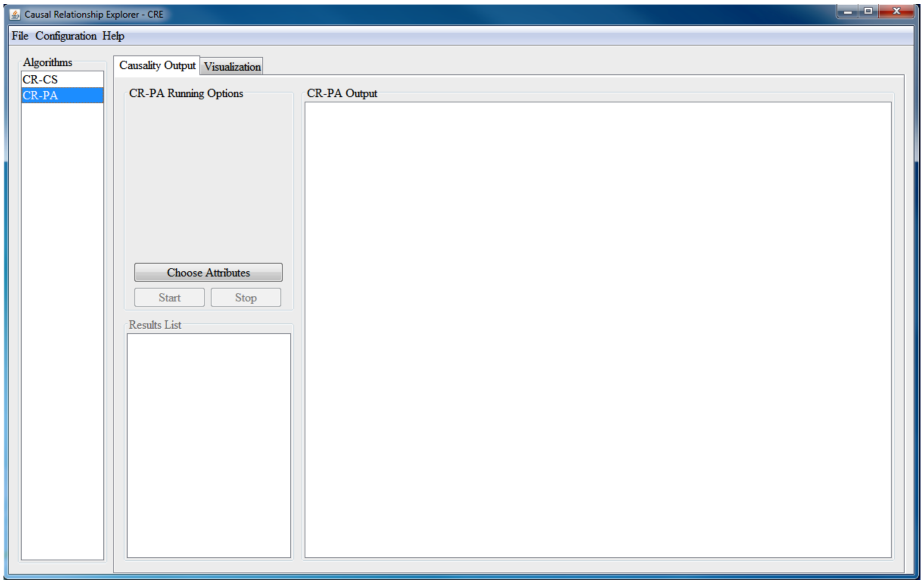


Figure 3: The main user interface of CRE

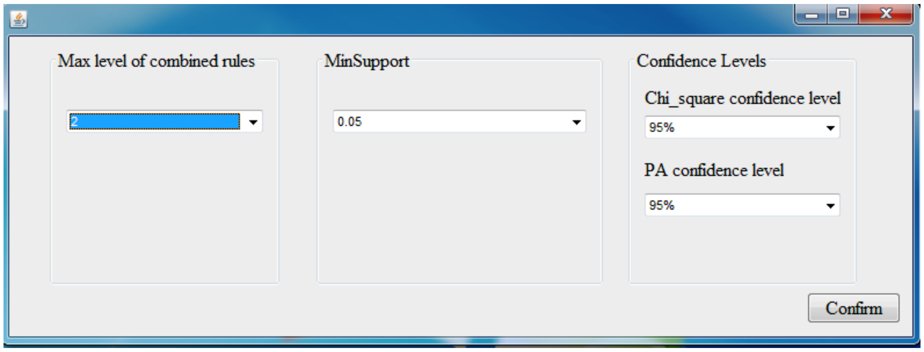
- *Max level of combined rules*: l_{max} in the CR-PA algorithm, the maximum level of combined causes, e.g. select 2 for mining single and combined (level 2) causal rules.
- *MinSupport*: the minimum support, i.e. m_{supp} in the CR-PA algorithm.
- *Chi-square confidence level*: the confidence level of the Chi-square tests in the CR-PA algorithm.
- *PA confidence level*: the confidence level for the partial association tests. By default, it is set to the same as the Chi-square confidence level.

Click "Confirm" after all of the parameters are set.

5. Select attributes from the data set. As shown in Figure 5, click the "Choose Attributes" button in the middle panel of the CRE user interface to select the attributes (predictor variables) of interest, then click the "All" button to select all attributes in the data set, or tick the boxes to choose a subset of attributes. Note that the target variable, which is the last column in the data set, is not included in the list of attributes. Click "Confirm" after selecting the desired attributes.

6. Run CR-PA. Click the "Start" button to run the CR-PA algorithm.

7. Obtain the results. The results are shown in the CR-PA output panel as in Figure 6. Moreover, the results are also stored in the file *current_optforUI.csv*,



Max level of combined rules: 2

MinSupport: 0.05

Confidence Levels

Chi_square confidence level: 95%

PA confidence level: 95%

Figure 4: The interface for parameter setting in CR-PA

located in the current folder of *CRE*. Each row of the report is a rule. The “Causal/Noncausal” column tells us if a rule is causal rule or just an association rule. The “level” column shows the length of the LHS of a rule, with level 1 indicating a single variable in the LHS of a rule and level 2 indicating combined (level 2) rules. The last four columns show the statistics of the rule which are the values of the four cells in the contingency table. We can see that the result is consistent with the causal structure in Figure 1. Note that in this example, we are only interested in rules with $Z = 1$. To generate rules with $Z = 0$, we need to swap the positions of “0” and “1” at the first line of the *Example21.names* file before running CR-PA.

5 Using CR-CS

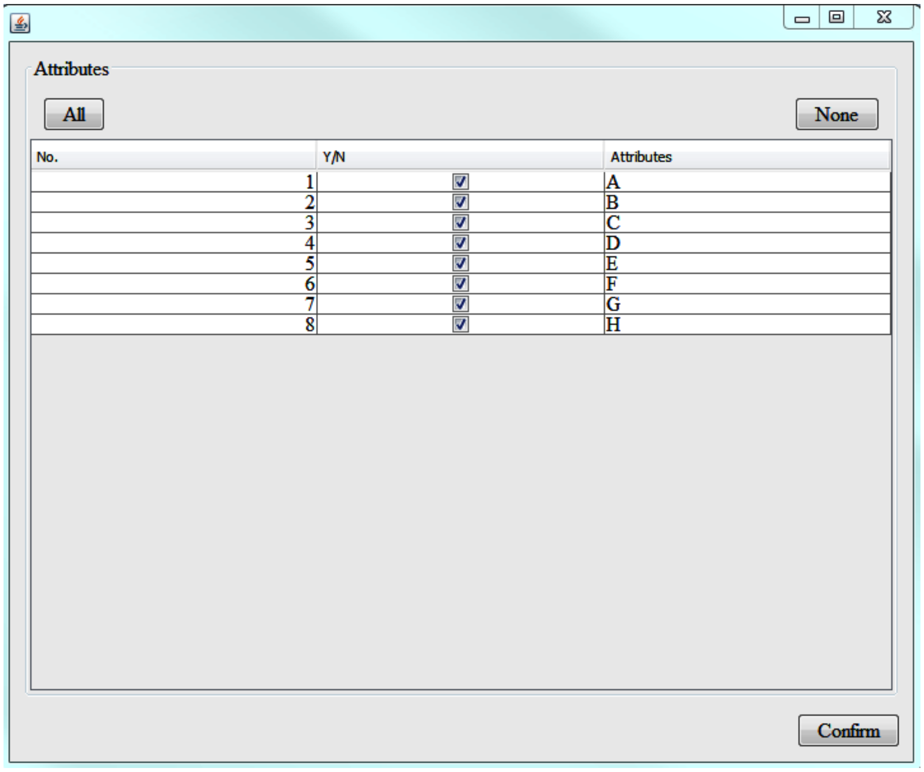
We assume that readers have installed the CRE software tool in their computers by following the instructions in the above section. In the following, we demonstrate how to use CRE to discover causal rules by CR-CS.

1. Open the input file. Select on the menu bar, File → Open File. Select the input *.names* file from local computer, in this case, select *Example41.names*.

2. Select the algorithm CR-CS from the left panel.

3. Set parameters for CR-CS. On the menu bar, select Configuration → Parameters. A pop-up window will appear for setting CR-CS parameters as shown in Figure 7. There are three parameters as follows:

   - *Max level of combined rules*: l_{max} in the CR-CS algorithm, which is the maximum level of combined causes that users desire, e.g. select 2 for mining single and combined (level 2) causal rules.
   - *MinSupport*: the minimum support, i.e. m_{supp} in the CR-CS algorithm.



No.	Y/N	Attributes
1	☑	A
2	☑	B
3	☑	C
4	☑	D
5	☑	E
6	☑	F
7	☑	G
8	☑	H

Figure 5: The interface for choosing attributes in CR-PA

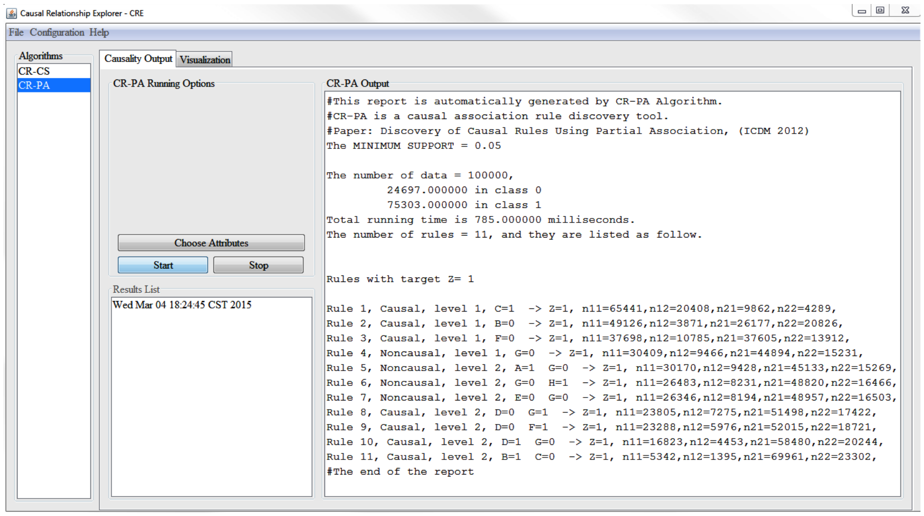
- *Odds ratio*: the odds ratio threshold for mining association rules and causal rules. The default value is “lower bound”, which uses the lower bound of the confidence interval of the odds ratio as the criterion (see Appendix A for more details)

Click “Confirm” after all of the parameters are set.

4. Select attributes from the data set. Click the “Choose Attributes” button in the middle panel of the CRE user interface to select the attributes (predictor variables) of interest. Click the “All” button to select all attributes in the data set, or alternatively tick the boxes to choose a subset of attributes. Note that the target variable, which is the last column in the data set, is not included in the list of attributes. Click “Confirm” after selecting the desired attributes.

5. Run CR-CS. Click the “Start” button to run the CR-CS algorithm.

6. Obtain the results. The outputs are shown in the CR-CS output panel. Moreover, the results are also stored in the file *current_optforUI.csv*, which



```
#This report is automatically generated by CR-PA Algorithm.
#CR-PA is a causal association rule discovery tool.
#Paper: Discovery of Causal Rules Using Partial Association, (ICDM 2012)
The MINIMUM SUPPORT = 0.05

The number of data = 100000,
         24697.000000 in class 0
         75303.000000 in class 1
Total running time is 785.000000 milliseconds.
The number of rules = 11, and they are listed as follow.


Rules with target Z= 1

Rule 1, Causal, level 1, C=1  -> Z=1, n11=65441,n12=20408,n21=9862,n22=4289,
Rule 2, Causal, level 1, B=0  -> Z=1, n11=49126,n12=3871,n21=26177,n22=20826,
Rule 3, Causal, level 1, F=0  -> Z=1, n11=37698,n12=10785,n21=37605,n22=13912,
Rule 4, Noncausal, level 1, G=0  -> Z=1, n11=30409,n12=9466,n21=44894,n22=15231,
Rule 5, Noncausal, level 2, A=1  G=0  -> Z=1, n11=30170,n12=9428,n21=45133,n22=15269,
Rule 6, Noncausal, level 2, G=0  H=1  -> Z=1, n11=26483,n12=8231,n21=48820,n22=16466,
Rule 7, Noncausal, level 2, E=0  G=0  -> Z=1, n11=26346,n12=8194,n21=48957,n22=16503,
Rule 8, Causal, level 2, D=0  G=1  -> Z=1, n11=23805,n12=7275,n21=51498,n22=17422,
Rule 9, Causal, level 2, D=0  F=1  -> Z=1, n11=23288,n12=5976,n21=52015,n22=18721,
Rule 10, Causal, level 2, D=1  G=0  -> Z=1, n11=16823,n12=4453,n21=58480,n22=20244,
Rule 11, Causal, level 2, B=1  C=0  -> Z=1, n11=5342,n12=1395,n21=69961,n22=23302,
#The end of the report
```

Figure 6: The output of CR-PA algorithm for the data set

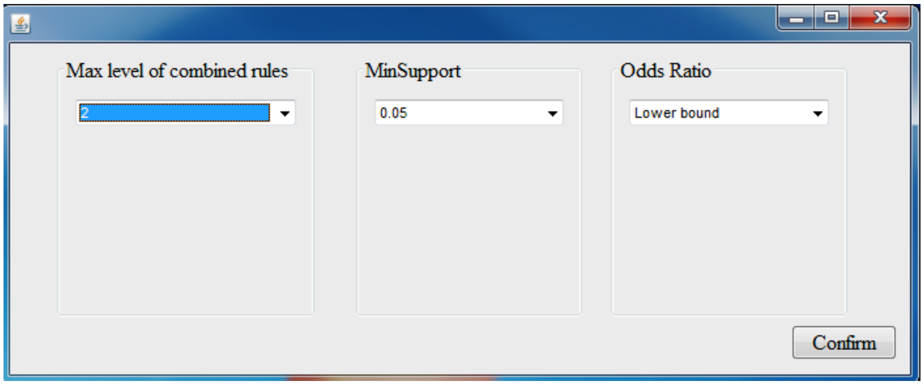
is located in the current folder of *CRE*. Each row of the report is a rule. The "Causal/Noncausal" column tells us if the rule is a causal rule or it is just an association rule. The "level" column shows the length of the LHS of a rule, with level 1 indicating single variable in the LHS of a rule and level 2 indicating 2 variables in LHS of a rule. The last four columns show the statistics of the rule which are the values of the four cells in the contingency table. We can see that the result is consistent with the causal structure in Figure 1.

References

[1] J. Li, L. Liu, and T. D. Le. *Practical Approaches to Causal Relationship Exploration*. Springer, 2015.

[2] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2003.

[3] P. Spirtes, C. C. Glymour, and R. Scheines. *Causation, Predication, and Search*. The MIT Press, 2nd. edition, 2000.



Max level of combined rules: 2

MinSupport: 0.05

Odds Ratio: Lower bound

Figure 7: The interface for parameter setting in CR-CS