UM–SJTU Joint Institute

Fall 2018 VE593 Problem Solving with AI Techniques

Project 2 Report

Xinhao Liao

516370910037

Contents

1 Part 1. Analyzing the dataset of wine

1.1 Parameter determination

For the topological order for structural learning, it's not realistic to check for all the possible orders for this data set. The reason is that there are 12 variables in all, which lead to 12! = 479001600 kinds of orders. Assume testing each order cost only 1s, checking 12! orders need 5544 days. So random orders are generated and checked until a time limit is reached, and the produced graph with the largest total score is selected. The time limit is set to be 30 min by default.

For the data for training, the first 4746 cases are used for training. It's found that even all the cases are used for training, the time cost for one order is no more than 1 minute. So time is not a problem here.

However, after analyzing the data, we find that the maximum value of one variable, "freesulfurdioxide" appears in the 4746-th case for the first time. After that case appears, all the maximum and minimum values for the variables have been detected. Since the possible values are determined by the minimum and maximum values detected in training, we need to use at least 4746 cases for training if the list for training is continuous and from the beginning of the data set.

For the number K, it's set to be 11. The reason is that any one the first 11 variables describing some features of a wine, which can be experimentally measured, might affect the 12th variable, quality, of wine.

The variable of interest is quality for wine, which has the index 11 based on the original given order.

1.2 Result

The obtained graph when evaluating with K2 score is as follows,

$$\begin{array}{c}
[[0,0,1,0,0,0,0,0,1,0,0,0], \\
[0,0,0,0,0,0,0,0,0,0,1,1], \\
[0,0,0,0,0,0,0,0,0,0,0,0], \\
[0,0,0,0,0,0,0,0,0,0,0,0], \\
[0,0,1,0,0,0,0,0,0,0,1,0], \\
[0,0,0,0,0,0,0,0,0,0,0,0], \\
[1,0,0,0,0,1,0,1,0,0,1,0], \\
[0,0,1,1,0,0,0,0,0,1,1,0], \\
[0,0,0,0,0,0,0,0,0,0,0,0], \\
[0,0,0,0,0,0,0,0,0,0,0,0], \\
[0,0,0,0,0,0,0,0,1,1,0,1], \\
[0,0,0,0,0,0,0,0,0,0,0,0]]
\end{array}$$

with the total score of -40676.3858915201 (ln score). And the accuracy rate is 0.6052631578947368.

For the variables that are most useful to predict the value of the variable quality, when adjusting the size k of the observed variables, it's found that the subsets of variables of that size that appear first in order is shown in Table 1.

size k	First Subset with the greatest Accracy	Accuracy
1	[0]	0.5526315789473685
2	[1, 10]	0.6052631578947368
3	[0, 1, 10]	0.6052631578947368
4	[0, 1, 2, 10]	0.6052631578947368
5	[0, 1, 2, 3, 10]	0.6052631578947368
6	[0, 1, 2, 3, 4, 10]	0.6052631578947368
7	[0, 1, 2, 3, 4, 5, 10]	0.6052631578947368
8	[0, 1, 2, 3, 4, 5, 6, 10]	0.6052631578947368
9	[0, 1, 2, 3, 4, 5, 6, 7, 10]	0.6052631578947368
10	[0, 1, 2, 3, 4, 5, 6, 7, 8, 10]	0.6052631578947368
11	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]	0.6052631578947368

Table 1

As we can see, the variable with index 1 and 10 are most related to the quality, they are the volatile acidity and the alcohol.

The obtained graph when evaluating with BIC score is as follows,

$$[[0,0,0,0,0,0,0,0,0,0,0,0],$$

$$[0,0,0,0,0,0,0,0,0,0,0,0],$$

$$[0,0,0,0,0,0,0,0,0,0,0,0],$$

$$[1,1,1,0,0,1,0,1,1,1,1,1],$$

$$[0,0,0,1,0,0,1,0,0,0,0,0],$$

$$[0,0,0,0,0,0,0,0,0,0,0,0],$$

$$[0,0,0,0,0,0,0,0,0,0,0,0],$$

$$[0,0,0,0,0,0,0,0,0,0,0,0],$$

$$[0,0,0,0,0,0,0,0,0,0,0,0],$$

$$[0,0,0,0,0,0,0,0,0,0,0,0],$$

$$[0,0,0,0,0,0,0,0,0,0,0,0],$$

$$[0,0,0,0,0,0,0,0,0,0,0,0]]$$

with the total score of -3648.597266491105 (ln is applied for log). And the accuracy rate is 0.5526315789473685.

For the variables that are most useful to predict the value of the variable quality, when adjusting the size k of the observed variables, it's found that the subsets of variables of that size that appear first in order is shown in Table 2.

Since all the variables appear in order, and the accuracy remains the same, it can't be determined which variables are most related to the variable of interest from the result above.

Clearly for this data set of wine, the implementation using K2 score function is more accurate.

size k	First Subset with the greatest Accracy	Accuracy
1	[0]	0.5526315789473685
2	[0, 1]	0.5526315789473685
3	[0, 1, 2]	0.5526315789473685
4	[0, 1, 2, 3]	0.5526315789473685
5	[0, 1, 2, 3, 4]	0.5526315789473685
6	[0, 1, 2, 3, 4, 5]	0.5526315789473685
7	[0, 1, 2, 3, 4, 5, 6]	0.5526315789473685
8	[0, 1, 2, 3, 4, 5, 6, 7]	0.5526315789473685
9	[0, 1, 2, 3, 4, 5, 6, 7, 8]	0.5526315789473685
10	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]	0.5526315789473685
11	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]	0.5526315789473685

Table 2

2 Part 2. Analyzing the dataset of protein

2.1 Parameter determination

For the topological order for structural learning, there are only 5! = 120 orders in all, and we can check for all the possible orders for this data set within 30 minutes.

For the data for training, the first 15000 cases are used for training. The time cost for one order is no more than 10 seconds. So time is not a problem here.

Assume the data are checked from the first line downward. Since all the maximum and minimum values have appeared in the first 18000 cases, and the possible values are determined by the minimum and maximum values detected in training, it's ok to check the first 18000 cases for training.

For the number K, it's set to be 5. The reason is that any one the last 5 variables, corresponding to amino acid types in a given protein DNA sequence, might affect The first variable nuc representing the risk for a certain disease.

The variable of interest is ”nuc” representing the risk for a certain disease, which has the index 0 based on the original given order.

2.2 Result

The obtained graph when evaluating with K2 score is as follows,

$$\begin{gathered}[[0,0,0,0,0,0],\\ [0,0,0,0,0,0],\\ [0,0,0,0,0,0],\\ [0,0,0,0,0,0],\\ [0,0,0,0,0,1],\\ [0,0,0,0,0,0]]\end{gathered}$$

with the total score of -93156.3648569196 (ln score). And the accuracy rate is 0.5344594594594595.

For the variables that are most useful to predict the value of the variable quality, when adjusting the size k of the observed variables, it's found that the subsets of variables of that size that appear first in order is shown in Table 1.

size k	First Subset with the greatest Accracy	Accuracy
1	[0]	0.5344594594594595
2	[0,1]	0.5344594594594595
3	[0, 1, 2]	0.5344594594594595
4	[0, 1, 2, 3]	0.5344594594594595
5	[0, 1, 2, 3, 4]	0.5344594594594595
6	[0, 1, 2, 3, 4, 5]	0.5344594594594595
7	[0, 1, 2, 3, 4, 5, 6]	0.5344594594594595
8	[0, 1, 2, 3, 4, 5, 6, 7]	0.5344594594594595
9	[0, 1, 2, 3, 4, 5, 6, 7, 8]	0.5344594594594595
10	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]	0.5344594594594595
11	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]	0.5344594594594595

Table 1

Since all the variables appear in order, and the accuracy remains the same, it can't be determined which variables are most related to the variable of interest from the result above.

The obtained graph when evaluating with BIC score is as follows,

$$
\begin{gathered}
[[0,0,0,0,0,0], \\
[0,0,0,0,0,0], \\
[0,0,0,0,0,0], \\
[0,1,1,0,0,0], \\
[1,0,0,1,0,1], \\
[1,0,0,1,0,0]]
\end{gathered}
$$

with the total score of -8419.287851245937 (*ln* is applied for *log*). And the accuracy rate is 0.5526315789473685.

For the variables that are most useful to predict the value of the variable quality, when adjusting the size k of the observed variables, it's found that the subsets of variables of that size that appear first in order is shown in Table 2.

size k	First Subset with the greatest Accracy	Accuracy
1	[0]	0.5344594594594595
2	[0,1]	0.5344594594594595
3	[0, 1, 2]	0.5344594594594595
4	[0, 1, 2, 3]	0.5344594594594595
5	[0, 1, 2, 3, 4]	0.5344594594594595
6	[0, 1, 2, 3, 4, 5]	0.5344594594594595
7	[0, 1, 2, 3, 4, 5, 6]	0.5344594594594595
8	[0, 1, 2, 3, 4, 5, 6, 7]	0.5344594594594595
9	[0, 1, 2, 3, 4, 5, 6, 7, 8]	0.5344594594594595
10	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]	0.5344594594594595
11	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]	0.5344594594594595

Table 2

Since all the variables appear in order, and the accuracy remains the same, it can't be determined which variables are most related to the variable of interest from the result above.

For this data set of protein, the result of the 2 score functions have the same accuracy. And they both fail to find out the variable most useful to predict the variable of interest.