Expert Data Mining:

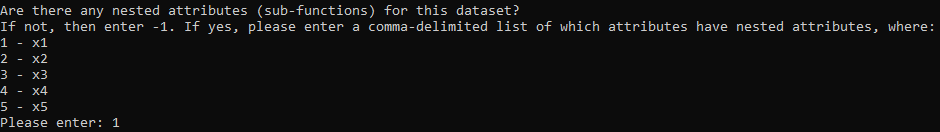
Nested Functions

# Introduction

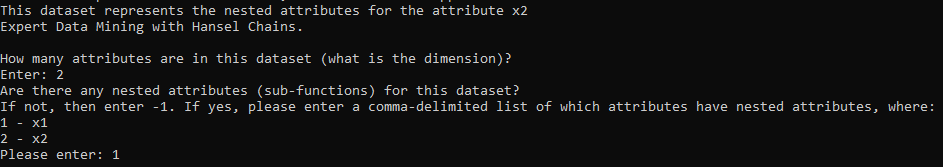
The Expert Data Mining software is now able to use nested functions. Essentially, any attributes can have sub-attributes, meaning that there can be multiple sequences of asking questions to the user. By default, the attributes of the first Boolean function will be given names that start with “x,” for example, there is attribute “x1.” Then, the next level Boolean function that is created will have attributes that start with “y.” Then, the next level will start with “z.” Finally, any subsequent levels will start with “a” and increment down back to “w.” A common function that we like to use for testing is “x1x3x5 v x2x3x5 v x4x5.” Any one of these attributes can have a couple sub-attributes. Let’s say that attribute x3 has two sub-attributes, y1 and y2. These two attributes will have their own Boolean function which makes the attribute x3 true or false. The program uses both a top-down approach (identifying sub-attributes) and a bottom-up approach (identifying groups of attributes). The top-down approach is easier to use due to the nature of the user input. Both of the implementations are recursive depth-wise. Moreover, the user now has the option to name the Boolean function as the first pilot question. Note: as always, improper user input can lead to some unintended behavior. This software will be polished up when it approaches the more sophisticated stage

# Top-down Approach

The implementation is quite simple: before we ask the user any questions, we ask if there are any sub-attributes.



In this example, we entered one so that the attribute x1 would have sub-attributes. Multiple attributes can be designated to have sub-attributes. Once all the pilot questions (such as the dimension and order of Hansel Chains) have been answered, we move onto asking questions about the sub-attributes, starting with how many of them there are.



We entered two so that that there are 2 attributes in the dataset (dimension of 2), and then specified that there are sub-sub-attributes. The process is iterative, so any depth of sub-attributes could be used, although it would likely be better to group the attributes in a more efficient way if the depth gets too deep. Once all the pilot questions for each level of Boolean function is asked and answered, now we start asking questions on the datapoints that belong to that Boolean function. We start with the uppermost level and proceed downwards breadthwise.

Furthermore, the output file of the results uses a slightly different nomenclature and format to account for the sub-functions and sub-attributes. The top-level function’s corresponding file is simply called “parent.csv.” If there are any children or grandchildren, then they will be labelled “child\_of\_[insert attribute].csv.” If there is a child of the first attribute of the parent function, then it will simply be called “child\_of\_x1.csv.” Moreover, there is another results file that is simply called “function\_hierarchy.txt.” It describes the mathematical hierarchy of the different level of functions. For example, “f(x1) = y1, y2” will mean that x1 is a function of child attributes y1 and y2. In the case that there are no child attributes, then simply “f(x1) = “ would be left blank. It is assumed to be the identity function, since no child attributes were given.

# Bottom-up Approach

The bottom-up approach is similar in that it can achieve the same result as the top-down approach, but it takes a bit more work and pilot questions to get there. Note: using a top-down approach and bottom-up approach at the same time isn’t allowed for groups. If we create a grouping of attributes, it doesn’t make sense to ask for sub-attributes of the attribute that correlates to that group (they are the attributes in that group). However, it is possible and allowed to create sub-attributes before creating any groups.

In the implementation, before we ask any questions, and only if there are more than 5 attributes, we ask the user if there are any groups of attributes. Then, if those groups of attributes still have more than 5 attributes, we ask the user to group them further, creating more than 2 levels of attributes. We ask the user to list the specific attributes that they would like to go in each group. Then, the current grouping of attributes is destroyed because all those attributes were moved into separate groups. The next sequence of questions will be about the groups that were just made. Let’s say that we split 8 attributes into 2 groups: one with 2 attributes and one with 6 attributes. These 2 groups form their own parent function of 2 attributes, so we will ask more pilot questions about that parent function first. Then we will go down to the next level to ask the pilot questions, and so on until all pilot questions have been answered. In this case, we ask the user to split the group of 6 into more groups (the user doesn’t have to). Just like before, the current group is destroyed, a new parent that corresponds to the groups is created, and children that are the groups themselves are also created. Finally, the expert data mining process will be initiated, starting at the uppermost level, and moving downwards.

Below is an exact transcript of some user input that uses the bottom-up approach (grouping instead of creating sub attributes). This entry is exactly the same as the short example that was described above. It is a lengthy process, and creating the same hierarchy of functions might be easier to execute with the top-down approach. The results files that correspond to this example are given in this folder.

Expert Data Mining.

What would you like to name the Boolean function/dataset that the following questions will be about? Please Enter: fruits

Now asking pilot questions:

How many attributes are in this dataset (what is the dimension)?

Enter: 8

Are there any nested attributes (sub-functions) for this dataset?

If not, then enter -1. If yes, please enter a comma-delimited list of which attributes have nested attributes, where:

1 - x1

2 - x2

3 - x3

4 - x4

5 - x5

6 - x6

7 - x7

8 - x8

Please enter: -1

Are there any groupings of attributes?

If not, then enter -1. If yes, please enter how many groups there are (please keep this to 2-5 groups):

2

Please note that all attributes must be in a group since this option was chosen, even if that group is of attributes that are unrelated.

Please enter a comma-delimited list of what attributes are in group 1: 1,3

Please enter a comma-delimited list of what attributes are in group 2: 2,4,5,6,7,8

Since parent attributes have been specified (the groupings), the expert data mining process will start with those attributes first.

The following questions will be about the groups (parent attributes aka super attributes) that were just created. Since we just created groups from those attributes, the previous function will be ovewritten by the new parent function and the new child functions.

Expert Data Mining.

What would you like to name the Boolean function/dataset that the following questions will be about? Please Enter: fruits

Now asking pilot questions:

What order to use for the Hansel Chains?

0 - Shortest Hansel Chain First

1 - Longest Hansel Chain First

2 - Manual Hansel Chain Order

3 - Default Hansel Chain Algorithm Order

Enter: 0

Is there any attribute which must be true for the datapoint to be true?

Enter the number assigned to each attribute or -1 if there is no such attribute.

If there are multiple attributes, separate them with a comma.

a - 0

b - 1

Enter: 1

Do you want to use chain jumping (1/0)?

Enter: 0

Expert Data Mining.

The following questions are for the child function (aka sub-function or nested function) that belongs to the parent function named fruits

The associated parent attribute is x1

What would you like to name the Boolean function/dataset that the following questions will be about? Please Enter: apples

Now asking pilot questions:

What order to use for the Hansel Chains?

0 - Shortest Hansel Chain First

1 - Longest Hansel Chain First

2 - Manual Hansel Chain Order

3 - Default Hansel Chain Algorithm Order

Enter: 0

Is there any attribute which must be true for the datapoint to be true?

Enter the number assigned to each attribute or -1 if there is no such attribute.

If there are multiple attributes, separate them with a comma.

a - 0

b - 1

Enter: 1

Do you want to use chain jumping (1/0)?

Enter: 0

Expert Data Mining.

The following questions are for the child function (aka sub-function or nested function) that belongs to the parent function named fruits

The associated parent attribute is x2

What would you like to name the Boolean function/dataset that the following questions will be about? Please Enter: oranges

Now asking pilot questions:

Are there any groupings of attributes?

If not, then enter -1. If yes, please enter how many groups there are (please keep this to 2-5 groups):

2

Please note that all attributes must be in a group since this option was chosen, even if that group is of attributes that are unrelated.

Please enter a comma-delimited list of what attributes are in group 1: 1,2,3

Please enter a comma-delimited list of what attributes are in group 2: 4,5,6

Since parent attributes have been specified (the groupings), the expert data mining process will start with those attributes first.

The following questions will be about the groups (parent attributes aka super attributes) that were just created. Since we just created groups from those attributes, the previous function will be ovewritten by the new parent function and the new child functions.

Expert Data Mining.

The following questions are for the child function (aka sub-function or nested function) that belongs to the parent function named fruits

The associated parent attribute is x2

What would you like to name the Boolean function/dataset that the following questions will be about? Please Enter: oranges

Now asking pilot questions:

What order to use for the Hansel Chains?

0 - Shortest Hansel Chain First

1 - Longest Hansel Chain First

2 - Manual Hansel Chain Order

3 - Default Hansel Chain Algorithm Order

Enter: 0

Is there any attribute which must be true for the datapoint to be true?

Enter the number assigned to each attribute or -1 if there is no such attribute.

If there are multiple attributes, separate them with a comma.

a - 0

b - 1

Enter: 1

Do you want to use chain jumping (1/0)?

Enter: 0

Expert Data Mining.

The following questions are for the child function (aka sub-function or nested function) that belongs to the parent function named oranges

The associated parent attribute is z1

What would you like to name the Boolean function/dataset that the following questions will be about? Please Enter: tangerine

Now asking pilot questions:

What order to use for the Hansel Chains?

0 - Shortest Hansel Chain First

1 - Longest Hansel Chain First

2 - Manual Hansel Chain Order

3 - Default Hansel Chain Algorithm Order

Enter: 0

Is there any attribute which must be true for the datapoint to be true?

Enter the number assigned to each attribute or -1 if there is no such attribute.

If there are multiple attributes, separate them with a comma.

a - 0

b - 1

c - 2

Enter: 1

Do you want to use chain jumping (1/0)?

Enter: 0

Expert Data Mining.

The following questions are for the child function (aka sub-function or nested function) that belongs to the parent function named oranges

The associated parent attribute is z2

What would you like to name the Boolean function/dataset that the following questions will be about? Please Enter: mandarin

Now asking pilot questions:

What order to use for the Hansel Chains?

0 - Shortest Hansel Chain First

1 - Longest Hansel Chain First

2 - Manual Hansel Chain Order

3 - Default Hansel Chain Algorithm Order

Enter: 0

Is there any attribute which must be true for the datapoint to be true?

Enter the number assigned to each attribute or -1 if there is no such attribute.

If there are multiple attributes, separate them with a comma.

a - 0

b - 1

c - 2

Enter: 2

Do you want to use chain jumping (1/0)?

Enter: 0

There are no more pilot questions.

Now starting to ask questions on the boolean function named fruits

Use majority flag(1/0)?

Enter: 0

Enter the class for this data point:

a = 0

b = 1

Enter Class: 1

Now starting to ask questions on the boolean function named apples

This dataset represents the nested attributes for the attribute x1 of the function that is one level higher.

Use majority flag(1/0)?

Enter: 0

Enter the class for this data point:

a = 0

b = 1

Enter Class: 0

Enter the class for this data point:

a = 1

b = 1

Enter Class: 1

Now starting to ask questions on the boolean function named oranges

This dataset represents the nested attributes for the attribute x2 of the function that is one level higher.

Use majority flag(1/0)?

Enter: 0

Enter the class for this data point:

a = 0

b = 1

Enter Class: 1

Now starting to ask questions on the boolean function named tangerine

This dataset represents the nested attributes for the attribute z1 of the function that is one level higher.

Use majority flag(1/0)?

Enter: 0

Enter the class for this data point:

a = 0

b = 1

c = 0

Enter Class: 0

Enter the class for this data point:

a = 1

b = 1

c = 0

Enter Class: 1

Enter the class for this data point:

a = 0

b = 1

c = 1

Enter Class: 1

Now starting to ask questions on the boolean function named mandarin

This dataset represents the nested attributes for the attribute z2 of the function that is one level higher.

Use majority flag(1/0)?

Enter: 0

Enter the class for this data point:

a = 1

b = 0

c = 1

Enter Class: 0

Enter the class for this data point:

a = 0

b = 1

c = 1

Enter Class: 0

Enter the class for this data point:

a = 1

b = 1

c = 1

Enter Class: 1

Now beginning to ask about f-changes for the function named fruits

Checking for f-changes:

Number Datapoint Class

1.1 1, 0 0

2.1 0, 0 0

2.2 0, 1 1

2.3 1, 1 1

Please enter the number of any vectors which need any changes in a comma-separalited list (e.g. 1.1, 3.2, ..., 7.4):

The vector specified does not exist, so it has been skipped.

Now beginning to ask about f-changes for the function named apples

Checking for f-changes:

Number Datapoint Class

1.1 1, 0 0

2.1 0, 0 0

2.2 0, 1 0

2.3 1, 1 1

Please enter the number of any vectors which need any changes in a comma-separalited list (e.g. 1.1, 3.2, ..., 7.4):

The vector specified does not exist, so it has been skipped.

Now beginning to ask about f-changes for the function named oranges

Checking for f-changes:

Number Datapoint Class

1.1 1, 0 0

2.1 0, 0 0

2.2 0, 1 1

2.3 1, 1 1

Please enter the number of any vectors which need any changes in a comma-separalited list (e.g. 1.1, 3.2, ..., 7.4):

The vector specified does not exist, so it has been skipped.

Now beginning to ask about f-changes for the function named tangerine

Checking for f-changes:

Number Datapoint Class

1.1 0, 1, 0 0

1.2 1, 1, 0 1

2.1 1, 0, 0 0

2.2 1, 0, 1 0

3.1 0, 0, 0 0

3.2 0, 0, 1 0

3.3 0, 1, 1 1

3.4 1, 1, 1 1

Please enter the number of any vectors which need any changes in a comma-separalited list (e.g. 1.1, 3.2, ..., 7.4):

The vector specified does not exist, so it has been skipped.

Now beginning to ask about f-changes for the function named mandarin

Checking for f-changes:

Number Datapoint Class

1.1 0, 1, 0 0

1.2 1, 1, 0 0

2.1 1, 0, 0 0

2.2 1, 0, 1 0

3.1 0, 0, 0 0

3.2 0, 0, 1 0

3.3 0, 1, 1 0

3.4 1, 1, 1 1

Please enter the number of any vectors which need any changes in a comma-separalited list (e.g. 1.1, 3.2, ..., 7.4):

The vector specified does not exist, so it has been skipped.