Metadata Editor

Updated: September 22, 2015

Edits metadata associated with columns in a dataset

Category: Data Transformation / Manipulation (https://msdn.microsoft.com/en-us/library/azure/dn905863.aspx)

Module Overview

You can use the **Metadata Editor** module to change metadata that is associated with columns in a dataset. The data itself (including the values and the data types) are not actually altered—only the metadata inside Azure Machine Learning that tells downstream components how to use the column. Typical metadata changes might include:

- Treating Boolean or numeric columns as categorical values
- Indicating which column contains the *class* label, or the values you want to categorize or predict
- Marking columns as features
- Changing date/time values to a numeric value
- Adding or changing column names

Use **Metadata Editor** any time you need to tweak the definition of a column. For example, some learners require specific data types. Other modules require specific flags on the columns, such as **IsFeature** or **IsCategorical**. After performing the required operation, you can reset the metadata to its original state.

If you want to prevent changes to a particular metadata attribute, you can also select **Unchanged** from the drop-down menu or leave the new column name blank.

How to Use Metadata Editor

1. Use the Column Selector to choose columns by name, index, or type.

   Note that you must apply the metadata change to all columns in the selection, so you might need to use multiple instances of **Metadata Editor** when mixing different types of transformations.

   For example, if you are changing the data type of several numeric columns, you can select all the applicable columns, and then apply the change to all the columns.

However, suppose you want to convert a numeric column to a categorical column, and at the same time, convert a different set of columns to features. In that case, you would need to add a separate instance of **Metadata Editor**.

2. Specify the change that you want to apply to all the columns in the selection. You can:

    a. Change a column's datatype

    b. Rename columns

    c. Specify how the column is to be used by machine learning tasks

    d. Clear previously set metadata

3. Run the experiment.

Options

Use these parameters to modify the metadata of columns:

Column

Use the **Column Selector** to select the columns to which your changes should apply.

Data type

You can change the data type of the column if needed for specific operations. The data types supported are String, Integer, Float, Boolean, DateTime, and TimeSpan. If you do not specify a new data type, the column metadata is unchanged.

> **Tip**
>
> The change of data type occurs only in the metadata that is associated with the dataset. The actual column values are not altered unless you perform a different operation (such as rounding) on the column. Also, you can recover the original data type at any time by using **Metadata Editor** to reset the column data type.

Categorical

Use this option to specify that the values in the column be treated as categorical variables. The actual data values are not changed, but machine learning algorithms will handle the data differently. For example, numeric values will be treated as discrete, and they will not be used in calculations (such as calculating a mean value).

You can switch the metadata of the column back to the original type at any time by using another instance of **Metadata Editor**.

Fields

The metadata schema that is associated with each dataset consists of, at a minimum, the name and data type for each column. Additional metadata includes:

- Whether the column contains a feature, a label, or a weight.

- Whether the values in the column are numeric and can be used for computation or must be treated as discrete values (categorical variables).

By default, all columns are initially treated as features. This means that when the dataset is used for training a model, the learning algorithm uses the values as inputs (independent variables). If the column contains the values that you want to predict, you should use **Metadata Editor** to change the field type to **Label**.

If the dataset has previously been used for other models, you might want to clear all previous metadata changes, by using the **ClearFeatures** or **ClearLabels** options.

⚠ **Warning**

The ability to mark a column as a score is not available now, but will be added in a future release.

New column names

If you want to rename a column, type the new name. You can rename multiple columns, by typing the names as a comma-separated list in order of the column indices. No columns can be omitted.

Note that column names can use only characters that are supported by the UTF-8 encoding.

Examples

For examples of how **Metadata Editor** is used in preparing data and building models, see these sample experiments in the Model Gallery (http://azure.microsoft.com/documentation/services/machine-learning/models/):

- In the Breast cancer detection (http://go.microsoft.com/fwlink/?LinkId=525726) sample, metadata is changed to add column names to newly joined data and to ensure that the patient ID is handled as a categorical string value instead of a number.

- The Twitter sentiment analysis (http://go.microsoft.com/fwlink/?LinkId=525274) sample demonstrates how to use **Metadata Editor** to ensure that columns are treated as features and to clear the feature metadata.

- In the Data Processing and Analysis (http://go.microsoft.com/fwlink/?LinkId=525733) sample, **Metadata Editor** is used to define new columns names for data read from a webpage.

Technical Notes

- The following numeric data types are not supported: Double (decimal) and TimeStamp.
- Currently there is no option in **Metadata Editor** to flag a column as containing *scores*. However, you can use the Execute R Script (https://msdn.microsoft.com/en-us/library/azure/dn905952.aspx) module with a script similar to the following to indicate that a column contains scores:

```
dataset <- maml.mapInputPort(1)
attr(dataset$x, "label.type")= "True Labels"
attr(dataset$y, "feature.channel")= "Multiclass Classification
Scores"
attr(dataset$y, "score.type")= "Assigned Labels"
maml.mapOutputPort("dataset");
```

Expected Input

Name	Type	Description
Dataset	Data Table (https://msdn.microsoft.com/en-us/library/azure/dn905851.aspx)	Input dataset

Module Parameters

Name	Range	Type	Default	Description
Column	Any	ColumnSelection		Choose the columns to which your changes should apply.
Data type	List	Metadata editor datatype	Unchanged	Specify the new data type for the column.
Categorical	List	Metadata editor categorical	Unchanged	Indicate if the column should be flagged as categorical.
Fields	List	Metadata editor flag	Unchanged	Specify if the column should be considered a feature or label by learning algorithms.

New column names	any	String		Type the new names for the columns.

Output

Name	Type	Description
Results dataset	Data Table (https://msdn.microsoft.com/en-us/library/azure/dn905851.aspx)	Dataset with changed metadata

Exceptions

For a list of all exceptions, see Machine Learning Module Error Codes (https://msdn.microsoft.com/en-us/library/azure/dn905910.aspx).

Exception	Description
Error 0003 (https://msdn.microsoft.com/en-us/library/azure/dn906003.aspx)	An exception occurs if one or more of input datasets are null or empty.
Error 0017 (https://msdn.microsoft.com/en-us/library/azure/dn906039.aspx)	An exception occurs if one or more specified columns have a type that is unsupported by the current module.
Error 0020 (https://msdn.microsoft.com/en-us/library/azure/dn906040.aspx)	An exception occurs if the number of columns in some of the datasets that are passed to the module is too small.
Error 0031 (https://msdn.microsoft.com/en-us/library/azure/dn905832.aspx)	An exception occurs if the number of columns in the column set is less than needed.
Error 0027 (https://msdn.microsoft.com/en-us/library/azure/dn905865.aspx)	An exception occurs when two objects have to be of the same size, but they are not.
Error 0028 (https://msdn.microsoft.com/en-us/library/azure/dn905848.aspx)	An exception occurs when the column set contains duplicate column names and it is not allowed.

Error 0037 (https://msdn.microsoft.com/en-us/library/azure/dn905800.aspx)	An exception occurs if multiple label columns are specified and only one is allowed.

See Also

Data Transformation / Manipulation (https://msdn.microsoft.com/en-us/library/azure/dn905863.aspx)
Data Transformation (https://msdn.microsoft.com/en-us/library/azure/dn905834.aspx)
A-Z List of Machine Learning Studio Modules (https://msdn.microsoft.com/en-us/library/azure/dn906033.aspx)