OSEMN methodology
Step 2: Scrubbing data
Excerpts from How to Prepare Data For Machine Learning
By Jason Brownlee[1]

Stepan Oskin

July 20, 2019

Abstract

Machine learning algorithms learn from data. It is critical that we feed them the right data for the problem we want to solve. Even if we have good data, we need to make sure that it is in a useful scale, format and even that meaningful features are included.

In this document, we will discuss how to prepare data for a machine learning algorithm. This is a big topic, only the essentials are covered here, references for further reading are included in section 3.

1 Data Preparation Process

The more disciplined we are in our handling of data, the more consistent and better results we are like likely to achieve. The process for getting data ready for a machine learning algorithm can be summarized in three steps:

- Step 1: Select Data
- Step 2: Preprocess Data
- Step 3: Transform Data

Data preparation process can be followed in a linear manner, but it is very likely to be iterative with many loops.

1.1 Select Data

This step is concerned with selecting the subset of all available data that we will be working with. There is always a strong desire for including all data that is available, that the maxim ”more is better” will hold. This may or may not be true.

We need to consider what data we actually need to address the question or problem we are working on. We have to make some assumptions about the data we require and be careful to record those assumptions so that we can test them later if needed.

Below are some questions to help us think through this process:

- What is the extent of the data that we have available? For example, through time, database tables, connected systems. We need to ensure that we have a clear picture of everything that we can use.
- What data is not available that we wish we had available? For example, it can be data that is not recorded or cannot be recorded. We may be able to derive or simulate this data.
- What data don’t we need to address the problem? **Excluding data is almost always easier than including data.** We have to note down which data we excluded and why.

It is only in small problems, like competition or toy datasets where the data has already been selected for us.

1.2 Preprocess Data

After we have selected the data, we need to consider how are we going to use the data. This preprocessing step is about getting the selected data into a form that we can work.

Three common data preprocessing steps are formatting, cleaning and sampling:

- **Formatting**: The data we have selected may not be in a format that is suitable for us to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.

- **Cleaning**: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data we believe we need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely.

- **Sampling**: There may be far more selected data available than we need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. We can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

It is very likely that the machine learning tools that we use on the data will influence the preprocessing we will be required to perform. We will likely revisit this step.

1.3 Transform Data

The final step is to transform the process data. The specific algorithm we are working with and the knowledge of the problem domain will influence this step and we will very likely have to revisit different transformations of our preprocessed data as we work on our problem.

Three common data transformations are scaling, attribute decompositions and attribute aggregations. This step is also referred to as **feature engineering**.

- **Scaling**: The preprocessed data may contain attributes with a mixtures of scales for various quantities such as dollars, kilograms and sales volume. Many machine learning methods like data attributes to have the same scale such as between 0 and 1 for the smallest and largest value for a given feature.

- **Decomposition**: There may be features that represent a complex concept that may be more useful to a machine learning method when split into the constituent parts. An example is a date that may have day and time components that in turn could be split out further. Perhaps only the hour of day is relevant to the problem being solved.

- **Aggregation**: There may be features that can be aggregated into a single feature that would be more meaningful to the problem that we are trying to solve. For example, there may be data instances for each time a customer logged into a system that could be aggregated into a count for the number of logins allowing the additional instances to be discarded.

A lot of time can be spent engineering features from our data and it can be very beneficial to the performance of an algorithm.

2 Summary

In this document we have discussed the essence of data preparation for machine learning. It was presented as a three step framework for data preparation with tactics for each step:

- **Step 1: Data Selection** Consider what data is available, what data is missing and what data can be removed.
- **Step 2: Data Preprocessing** Organize your selected data by formatting, cleaning and sampling from it.
- **Step 3: Data Transformation** Transform preprocessed data ready for machine learning by engineering features using scaling, attribute decomposition and attribute aggregation.

Data preparation is a large subject that can involve a lot of iterations, exploration and analysis. Getting good at data preparation will make you a master at machine learning.

3 Further reading

Resources listed below can be used to dive deeper into the subject:

- *From Data Mining to Knowledge Discovery in Databases, 1996*[2]
- *Data Analysis with Open Source Tools, Part 1*[3]
- *Machine Learning for Hackers, Chapter 2: Data Exploration*[4]
- *Data Mining: Practical Machine Learning Tools and Techniques, Chapter 7: Transformations: Engineering the input and output*[5]

References

[1] J. Brownlee, “How to Prepare Data For Machine Learning,” 2013.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” *AI Magazine*, vol. 17, no. 3, pp. 37–53, 1996.

[3] P. K. Janert, *Data Analysis with Open Source Tools: A Hands-On Guide for Programmers and Data Scientists.* Sebastopol, CA: O’Reilly Media, Inc., first ed., 2010.

[4] D. Conway and J. M. White, *Machine Learning for Hackers: Case Studies and Algorithms to Get You Started.* Sebastopol, CA: O’Reilly Media, Inc., first ed., 2012.

[5] I. H. Witten, F. Eibe, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco, CA: Elsevier, third ed., 2011.