Data Mining, CSCI 347, Fall 2019
Transformations, Nov. 4

Data Engineering – engineering the input data into a form suitable for the learning scheme chosen and engineering the output to make it more effective.

6 ways the data can be massaged to make it more amenable for learning methods:

1. Attribute selection
2. Attribute discretization
3. Data projections
4. Sampling the input
5. Data cleansing
6. Converting multiclass problems to two-class ones

Attribute selection

Fragmentation problem – small amount of data being available for attribute selection

Attribute selection can be done in two ways:

1. Filter method (scheme-independent selection) – select a subset of the attributes based on general characteristics of the data
2. Wrapper method (scheme-specific selection) - evaluate the subset using the machine learning algorithm that will ultimately be employed for learning, that is, the learning algorithm is "wrapped" into the selection procedure.

    - Methods using learning algorithms:
        - Apply a decision tree algorithm to the full dataset and select only those attributes that are used in the tree. A different learning method, such as instance-based learning, needs to be used once the attribute set is selected, since otherwise, the tree algorithm would output the same result
        - Use 1R to select attributes and then use a decision tree learning on selected attributes
        - Use an algorithm that builds a linear model (example linear support vector machine) and rank the attributes based on the size of their coefficients. (Be sure to normalize the values first so that the coefficients will be on the same scale.) Could use the attribute set for another learning algorithm. Alternatively, can remove low ranking attributes and use some algorithm again (called recursive feature elimination).

    The above methods help rank the attributes. Another issue is how many attributes to use. Can try a variety of numbers.

Attribute discretization

Some methods only deal with nominal attributes, so need to discretize. May also want to discretize numeric data.

Discretization can be supervised or unsupervised

Unsupervised discretization has 2 strategies:

- Equal interval binning
- Equal frequency binning

Supervised discretization (generally considered better)

- Entropy-based
  - Build decision tree but use the idea of minimum description length (MDL) as stopping criterion
  - Considered "state of the art"

Converse of discretization - make nominal values into numeric

- Assign numbers (if ordering is reasonable)
- Transform to binary (one hot encoding or dummy variables)

Data projections

Projection – function or mapping that transforms data in some way

Data projection – add new attributes that presents existing information in a new way

Difference of two attributes
Ratio of two attributes
Concatenation of the values of nominal attributes
Encoding cluster membership
Adding noise to the data
Removing data randomly or selectively
Obfuscating the data

Sampling the input

Sampling the input – only using part of the data. Includes a way of incrementally producing a random sample of given size.

Reservoir sampling

- Fill the reservoir, of size r, with the first r instances to arrive
- Subsequently, when the i-th item arrives (i>r)
  - With probability r/i keep the new item, discarding an old one, selecting which to replace at random (each with a 1/r change)
  - With probability 10r/i keep the old instance, ignoring the new instance