Description of and Grammar for Mark's File Format

COSC-575

August 26, 2018

Introduction

Since machine-learning algorithms operate on examples, we need a way to get examples into our programs. One way is to read the data directly from a database. Another is to read the information from a file. Even if the data were in a database, it is often convenient to put at least a portion of the information in a file for development and experimentation.

For information stored in a file, we need a simple format for specifying examples, attributes, domains, and values. Based on other commonly used, but more complicated formats, I devised a simple format that should suit our purposes. A grammar for the format follows, but you should be able to parse files by reading tokens, or by reading lines and processing tokens. An example of the the Bikes data set follows the grammar.

Grammar

$$
\begin{aligned}
\langle \textit{dataset} \rangle &\rightarrow \langle \textit{header} \rangle\ \langle \textit{attributes} \rangle\ \langle \textit{examples} \rangle \\
\langle \textit{header} \rangle &\rightarrow \text{`@dataset'}\ \langle \textit{identifier} \rangle \\
\langle \textit{attributes} \rangle &\rightarrow \langle \textit{attribute} \rangle\ \langle \textit{attribute} \rangle\ \langle \textit{attribute-list} \rangle \\
\langle \textit{attribute-list} \rangle &\rightarrow \langle \textit{attribute} \rangle\ \langle \textit{attribute-list} \rangle \mid \epsilon \\
\langle \textit{attribute} \rangle &\rightarrow \text{`@attribute'}\ \langle \textit{identifier} \rangle\ \langle \textit{declaration} \rangle \\
\langle \textit{declaration} \rangle &\rightarrow \text{`numeric'} \mid \langle \textit{nominal-values} \rangle \\
\langle \textit{examples} \rangle &\rightarrow \text{`@examples'}\ \langle \textit{example-list} \rangle \\
\langle \textit{example-list} \rangle &\rightarrow \langle \textit{example} \rangle\ \langle \textit{examples-list} \rangle \mid \epsilon \\
\langle \textit{example} \rangle &\rightarrow \langle \textit{attribute-value} \rangle\ \langle \textit{example} \rangle \mid \epsilon \\
\langle \textit{attribute-value} \rangle &\rightarrow \langle \textit{identifier} \rangle \mid \langle \textit{number} \rangle \\
\langle \textit{nominal-values} \rangle &\rightarrow \langle \textit{identifier} \rangle\ \langle \textit{identifier} \rangle\ \langle \textit{identifier-list} \rangle \\
\langle \textit{identifier-list} \rangle &\rightarrow \langle \textit{identifier} \rangle\ \langle \textit{identifier-list} \rangle \mid \epsilon \\
\langle \textit{identifier} \rangle &\rightarrow \langle \textit{non-whitespace-character} \rangle\ \langle \textit{identifier} \rangle \mid \epsilon \\
\langle \textit{number} \rangle &\rightarrow -\infty, \ldots, +\infty \\
\langle \textit{non-whitespace-character} \rangle &\rightarrow \text{`a'}, \ldots, \text{`z'}, \text{`A'}, \ldots, \text{`Z'}, \text{`0'}, \ldots, \text{`9'}, \text{`\_'}, \ldots, \text{`+'}
\end{aligned}
$$

Example

```
@dataset bikes

@attribute make trek bridgestone cannondale nishiki garyfisher
@attribute tires knobby treads
@attribute bars straight curved
@attribute bottle y n
@attribute weight numeric
@attribute type mountain hybrid

@examples

trek knobby straight y 250.3 mountain
bridgestone treads straight y 200 hybrid
cannondale knobby curved n 222.9 mountain
nishiki treads curved y 190 hybrid
trek treads straight y 196.8 hybrid
```