ESRB Classifier

Testing Design Document

## Dataset

### Data Sources

The classification relies on previously recorded official ESRB ratings to make a prediction about a hypothetical game’s content rating. Although this data isn’t made available by the ESRB itself, many unaffiliated and hobbyist projects have compiled data sets through publicly available sources for use in applications like ours. The data set used in the initial training for the classification contains what we believe is a good cross-section of gaming titles across the parental advisory categories, platform and genre. The data is provided and ingested by the software in a comma separated value (.csv) formatted file which contains:

* A record containing the headings for each column.
* 2396 records of past official ESRB game ratings.
* 34 data points per entry.
* Each game’s title.
* Whether the game released on a console platform.
* Whether the official list of the content descriptors for the title was not available.
* 31 content descriptors indicated by the ESRB.

#### Testing & Validation Data

The original data, once read from the original file, will be stored in memory as a simple java string object and then passed into a custom data set object that can split itself into randomly selected disjoint sets. In the initial read, we will split the entire data set into three parts:

* 70% for use in bootstrapping working data partitions to generate decision trees.
* 20% for sample data used in testing.
* 10% for validation of testing and tuning model parameters.

30% of the data is reserved for our tests and further refining the system to achieve high classification accuracy. To create the most representative sample data set for testing, we use a subset of our initial records as opposed to similar data from another source. We also use a programmatic approach to selection to avoid any bias or human error.