Analysis of Death in the United States

Bryan Brent
Nathan Lile
Alex Ray
Austin Pearman

ABSTRACT

This project serves to provide a better understanding of mortality, particularly with respect to the interrelationships between day-of-death and historical features such as race, sex, date, age, education, and circumstance of death in the United States from 2005 to 2015. The US Centers for Disease Control and Prevention releases extensive mortality data every year to allow for a variety of census analyses for life expectancy and death statistics, among other uses.

Our project will serve as a study of classification techniques on mortality data. Prior work with similar datasets has achieved mixed results with a number of algorithms including random forests and naive bayes; we hope to perform a more comprehensive survey of classification techniques as well as better understand the affect of different attributes on classification performance.

Finally, this project will delve into utilizing unsupervised learning techniques such as K-means to potentially gain insight into interesting patterns and trends of mortality in the US.

ACM Reference Format:
Bryan Brent, Nathan Lile, Alex Ray, and Austin Pearman. 2018. Analysis of Death in the United States. In *Proceedings of Data Mining*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.475/123_4

1 PROBLEM STATEMENT/MOTIVATION

Given CDC data on mortality in the United States from 2005 to 2015, which provides day-of-death data such as cause, age, education level, race, and marital status, we will initially attempt to predict the manner of death. Manner of death includes categories such as suicide, homicide, accidental, etc. We will be assessing classification success given only day-of-death information as well as performance with historical features and specific details of the death such as activity (at the time of death), location of injury as well as their family and descendant status.

Depending on the success of these initial manner of death prediction tasks, this project will also attempt to predict cause of death using the same attributes. Due to the specificity of cause of death in the data, this classification task also requires a preprocessing step to "roll up" the cause of death into meaningful supergroups. In theory, this work would allow for a more in-depth analysis of characteristics of individuals "at risk" for different manners and causes of death; given this information, various forms of government programs, social work, and other support systems may be able to adapt their methodologies.

Finally, we will be attempting to cluster mortality features to recognize interesting patterns or trends. This exploratory analysis affords an opportunity to potentially uncover novel interrelationships between attributes in the dataset as well as assess the "completeness" of the existing feature-set. If notable "holes" in attributes are found, it is possible to integrate past temporal data into the existing dataset–market trends and weather data are examples of readily available data ready to be integrated.

Interesting patterns would–like the supervised learning analysis–provide further context regarding what constitutes an individual with notable risk of some manner of death. Even a relatively unsuccessful analysis as determined by quantitative evaluation metrics provides insight into what sorts of features or attributes are necessary to meaningfully predict or cluster mortality events.

2 LITERATURE SURVEY (PREVIOUS WORK)

2.1 Death in the United States Kaggle Page

The Kaggle page for the CDC mortality dataset being used in this paper is a good resource for research questions, discussion on relevant topics, and a repository of existing non-academic work. For example, the main overview page contains interesting ideas on expanding previous work through increasing granularity in age data.

The discussion page contains useful threads on dealing with cause of death recodes (a significant part of our project goals). Discussions also include interesting, specific prior work using this dataset including clustering and predictive analyses.

Finally, the kernel page contains a curated, ranked list of prior work with this dataset. These provide examples of previous work, examples of handling a variety of cleaning and preprocessing steps in Python, and examples of analyses with varying levels of success and interestingness. [1]

2.2 Lessons learned from data mining WHO mortality database (Paoin W)

Previous studies of the CDC's mortality database have provided a template for the most effective data mining methods. Past researchers have concluded that, due to the issues mentioned in our problem statement, classification was generally ineffective for predicting cause of death.

The study cites a lack of correlation between variables and death cause as the root of the issue. On the contrary, it was found that clustering as well as association produced the most interesting patterns. The exact quote from the study can be found below:

Data Mining, Spring 2018, Boulder,Colorado USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

"Classification tools produced the poorest results in predicting cause of death. Given the inadequacy of variables in the WHO database, creation of a classification model to predict specific cause of death was impossible. Clustering and association tools yielded interesting results that could be used to identify new areas of interest in mortality data analysis. This can be used in data mining analysis to help solve some quality problems in mortality data." [4]. Our analysis hopes to elaborate on these classification findings by introducing algorithms beyond decision trees and naive bayes. Furthermore, we hope to emulate the prior success of unsupervised learning techniques on this sort of data.

2.3 Graphs of trends of deaths in the United states from 1900 until 2015 NCHS data visualization

The below graph displaying a century long longitudinal study showing trends in life expectancy and age adjusted death rates is indicative of the previous work that has been done in the study of mortality. An example graph is provided below. [2]

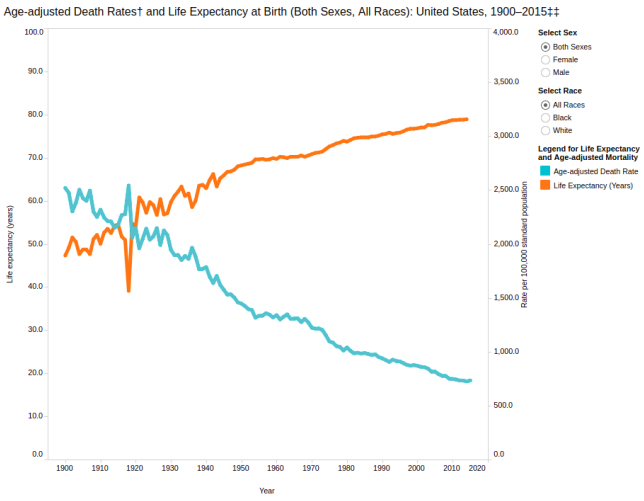

Figure 1: Age-adjusted Death Rates and Life Expectancy at Birth; United States, 1900-2015

2.4 CDC NVSS (National Vital Statistics System) Publication Page

This is the publication page for our database. It contains a list of studies that have been previously done with regards to the CDCs mortality data. We will be using this as a reference for selecting our clustering and association techniques. It has provided us with a roadmap for what does and doesn't work in terms of analyzing CDC mortality data. [3]

2.5 Applying data mining in medical data with focus on mortality related to accident in children

This study used data mining techniques and was seeking conclusions in line with our objectives. The researchers achieved results regarding classifying mortality rate with both decision trees as well as Bayes' theorem. We intend to use their methods as examples of possible techniques in our analysis. [5]

3 PROPOSED WORK

3.1 Data cleaning

(1) Several areas in the json files are of an incorrect syntax, essentially missing quotation marks. This will require a hands on reformatting of the database source.
(2) Converting cause-of-death and age "recodes" to a more usable format.
  (a) This step also includes potentially converting non-numeric data (in education "recodes" for example) into a matching numeric representation (e.g. numeric grade level).
(3) Ensuring continuity of feature representation and format across the entire database.

3.2 Data preprocessing

(1) "Rolling up" approximately 500 causes of death to create meaningful supergroups that are large enough to yield interesting results.
  (a) Note that this step is only needed for a subset of our final analysis tasks described below. It may require more advanced topic extraction methods if we attempt to autonomously perform this step.
(2) Identifying extraneous attributes that we can drop from the study before we begin analysis.
(3) Joining features to decrease dataset complexity while maintaining continuity and information integrity.
  (a) For example, there are multiple race-related features in this mortality dataset including multiple race recodes and Hispanic recodes that can be consolidated without loss of information.

3.3 Data integration

(1) We may use other databases to explore the possibility that the database is overlooking features which contribute to the cause of death. Particularly if the classification and/or clustering analyses do not achieve the intended success given our evaluation metrics, integration of external datasets may provide opportunities for more successful analyses.
  External datasets may include (but are not limited to) the following:
  (a) Weather
  (b) Sporting events
  (c) Political events
  (d) Financial markets

4 DATASETS

We will be using the CDC mortality database (URL below). The dataset contains data from deaths in the United States from 2005-2015. The data is compiled yearly by the CDC in the National Vital Statistics system. All deaths are accompanied by a wide array of attributes including age, race, cause of death, descendant status, education level, marital status, month, and day, among many others.

The database has been downloaded by Bryan Brent on his personal machine.

URL: `https://www.kaggle.com/cdc/mortality`

4.1 Temporality

4.1.1 Temporality in the CDC dataset opens the door for integration of other data to add additional features (weather, stock market, political events, etc).

5 TOOLS

5.1 Pandas

5.1.1 Will be utilized for all data analysis and manipulation unless need arises for more powerful tools.

5.1.2 Pandas provides intuitive and easy to use helper functions for data manipulation, analysis, and cleaning. Pandas dataframes afford powerful summarization and correlation tools as well as useful preprocessing functionality such as interpolation.

5.2 Scikit-learn

5.2.1 Scikit-learn provides a wide array of machine learning models and tools. These tools include but are not limited to classification and clustering algorithms, vectorization algorithms, and preprocessing algorithms like principal component analysis.

5.3 Matplotlib

5.3.1 For simple visualizations such as scatter plots, box and whisker plots, and histograms. Note that Pandas integrates Matplotlib to provide a simple plotting interface for dataframes.

5.4 D3.js

5.4.1 D3.js provides more advanced plotting functionality that allows for the creation of interesting infographics and designs to display results.

6 EVALUATION METHODS

Both the supervised and unsupervised analyses have traditional concrete quantitative evaluation metrics.

6.1 Classification

6.1.1 Evaluation of results can be performed using a number of metrics including accuracy, precision, recall, and F-score (a conglomeration of precision and recall). The specific metric to be used is highly dependent on the goal of the classification, as they each evaluate using a false positive-false negative tradeoff.

6.2 Clustering

6.2.1 Clustering allows a more distance-based quantitative evaluation metric. This can come in the form of using different distance formulas to assess inter-cluster distance (Manhattan distance, Euclidean distance, etc. as well as more refined "closeness" measures.

6.2.2 Evaluating clustering analysis results also includes a notable qualitative aspect that returns to the idea of "interesting" patterns. The success of a cluster analysis for this dataset certainly depends on the novelty of the findings as well as the relevance to real-world questions.

7 MILESTONES

(1) Combine each annual dataset into single complete dataset. ***March 12***
(2) Clean and preprocess data for first manner-of-death classification task. ***March 12***
   (a) Convert `education` related revision attributes into reasonable "rolled up" education summary.
   (b) Convert `age` related revision attributes into reasonable age summary with necessary granularity.
   (c) Convert `race` and `hispanic` related revision attributes into race summary with necessary granularity.
(3) Perform classification analysis to evaluate ability to determine manner of death from day-of-death attributes. ***March 19***
   (a) Perform and evaluate binary classification for a suicide vs non-suicide manner of death using support vector machine, random forest, and multi-layer perceptron, among others.
   (b) Given reasonable results with binary classification, extend analysis to multi-class classification.
(4) Clean and preprocess data for unsupervised learning analysis. ***March 23***
   (a) Summarize and process yet-unprocessed attributes such as place of death, injury at work, method of disposition, autopsy, and activity.
(5) Perform unsupervised learning analysis to evaluate ability to extract interesting information from mortality trends and patterns. ***March 30***
   (a) Perform and evaluate the results of clustering techniques such as K-means. This includes an evaluation using different hyperparameters such as number of clusters and distance formula used.
   (b) Assess need to extend analysis to include further techniques or algorithms.
(6) Roll-up cause of death attribute recodes into meaningful summary supersets. ***Loosely March 30***
(7) Given successfully preprocessing cause of death attributes, repeat previous classification analysis for cause of death prediction. ***April 13***
   (a) Include non-day of death attributes such as place of death, activity, and descendant status.

8 SUMMARY OF PEER REVIEW SESSION

The peer review session provided context for a successful data mining project as well as steps to further iterate on the project outline. Most notably, the peer review session illustrated the need to utilize prior work to narrow the scope of our analysis; as discussed previously, prior work in the study of mortality data has seen more success with clustering analyses yet also presents results of classification with a number of different algorithms.

Furthermore, the peer review session presented a variety of examples of integrating external datasets to further flesh out an analysis. Depending on the initial success of classification analysis in this paper, this context may be used to lend insight to integrating other temporal data into the CDC dataset.

Finally, many groups received feedback to further refine their evaluation metrics and specifics on procedure, particularly with data cleaning and data preprocessing (which has been covered in this course). We find that being more explicit in our procedure steps and quantitative evaluation metrics provide insight into next steps for this project.

REFERENCES

[1] kaggleMortality 2018. Death in the United States. (2018). Retrieved March 5, 2018 from https://www.kaggle.com/cdc/mortality/kernels

[2] nchsVisTool 2017. Mortality Trends in the United States, 1900-2015. (2017). Retrieved March 5, 2018 from https://www.cdc.gov/nchs/data-visualization/mortality-trends/

[3] nvss 2015. NVSS - Mortality Data. (2015). Retrieved March 2, 2018 from https://www.cdc.gov/nchs/nvss/deaths.htm

[4] W. Paoin. 2011. Lessons Learned from Data Mining of WHO Mortality Database. *Methods of Information in Medicine* 50, 4 (jun 2011), 380–385. https://doi.org/10.3414/me10-02-0019

[5] M. H. Saraee, Z. Ehghaghi, H. Meamarzadeh, and B. Zibanezhad. 2008. Applying data mining in medical data with focus on mortality related to accident in children. In *2008 IEEE International Multitopic Conference*. 160–164. https://doi.org/10.1109/INMIC.2008.4777728