An Exposome Data Analysis Pipeline in R

Jeff Sorbo
Department of Computer Science
Texas Tech University
Lubbock, Texas 79409-3104
Email: jeffrey.s.sorbo@ttu.edu

***Abstract*—Brief description of the work described in this paper.**

1. Introduction

The Exposome data are described; the goals of the analytic pipeline are defined; data loading, cleaning, and merging are detailed; feature selection and data modeling are detailed; model evaluation and results are discussed; and areas of future work are listed.

2. Background

The exposome is the concept of the complete set of one's lifetime exposures. [1] In recent years public health investigators have gained unprecedented access to high volume, high dimension data sets combined from multiple, diverse sources, enabling the investigators to measure elements of the exposome driving health disparities. The exposome data, when analyzed in large data sets, allows investigators to uncover previously unknown relationships between factors affecting health outcomes such as preterm birth rates [2], obesity [3], and cardiovascular disease (CVD).

With the advent of such large data sets and the increased need to conduct analyses against them comes the increased pain of conducting such analysis work manually. Running data analysis processes manually can lead to slower throughput and increased rates of human error. The need to automate analytic processes increases further as the sophistication of the analysis rises and investigators need to reproduce the results.

Numerous data mining algorithms can be applied to the exposome data to help investigators gain insights into the relationships among the data or to confirm known relationships. Clustering methods can uncover groupings among data points, decision trees can find rule sets to predict outcomes, and association rules mining can show co-occurrences among the data. When combined together as ensemble learning methods, it is hoped that a greater predictive accuracy may be achieved than if the methods were used individually. [4], [5]

3. Data

The exposome data were provided by the Exposome research group at Texas Tech University and consisted of 3,125 data points representing county and parish units across the United States. The data attributes can be grouped into several categories, such as social factors, health factors, and environmental factors. The data were provided in two files: the independent variables file and the dependent variables file.

The independent variables file contained data in 63 attributes: 3 unique identifiers including a string attribute consisting of the county and state name; and 60 numeric attributes consisting of various data aggregated at the county level, such as population, bank offices, housing unit values, per capita income, and average daily precipitation.

The dependent variables file contained data in 9 attributes: the unique identifier consisting of county and state name, 7 numeric attributes related to cardiovascular disease (CVD) death; and an attribute containing the quintile of the age-adjusted CVD death rate.

4. Methodology

A pipeline was developed to load, clean, merge, and preprocess the data and to train an ensemble learning model to predict the CVD rate. Based on [6], the ensemble learning model combined clustering and decision trees. The pipeline was written in the R language to provide for potential reuse and adaptation by members of the Exposome research group.

4.1. Data Loading and Conversions

The exposome data were loaded from the independent and dependent attributes files in comma-separated values (CSV) format.

Many of the attribute names in the files were based on codes in the original data sources, *e.g.*, "AGE030200D," "HEA010200D," "HSG680200D;" such attributes were given friendly names based on a data dictionary provided by the Exposome group.

Data points with missing values were removed, and the independent data were merged with the dependent data based on the county and state names.

All numeric attributes were grouped in quintiles.

The CVD attribute in the merged file was converted to a binary type: the highest quintile (*i.e.*, the highest rate of CVD) was set to 1, and the lower 4 quintiles were set to 0.

4.2. Feature Selection

The exposome data were grouped into subsets for model training and evaluation. The first subset, hereafter referred to as "data set 1," consisted of 10 attributes identified as a paraclique by members of the Exposome research group. The second subset, hereafter referred to as "data set 2," included all 23 statistical attributes from the independent attributes file.

Some feature selection techniques were applied to data set 2: the χ^2 test, symmetrical uncertainty, and gain ratio, all described in [7]. The data sets resulting from these methods are hereafter referred to "data set 2a," "data set 2b," and "data set 2c" respectively.

4.3. Data Modeling

K-Means clustering [8] was applied to the paraclique data with k = 3, and each data point was labeled with the cluster id 1-3.

Decision trees using recursive partitioning [9] were trained against each cluster.

Each tree was simplified using cost-complexity pruning as described in [10].

5. Results

The first tree was trained against the full paraclique data set without clustering. This tree is shown in Figure 1. The tree was found to have a predictive accuracy of 0.8203; a confusion matrix is shown in Table 1.

An attempt was made to prune the tree using cost-complexity pruning based on the lowest cross-validation error. This pruning attempt did not result in any changes to the tree. The process of training both unpruned and pruned trees was repeated 50 times. For the pruned trees, the mean predictive accuracy was 0.8382 with a standard deviation of 0.0130, whereas the pruned trees gave a mean predictive accuracy of 0.8368 with a standard deviation of 0.0139.

The second tree was trained against the statistical data set without clustering and is shown in Figure 2. This tree had a predictive accuracy of 0.8835; a confusion matrix is shown in Table 2.

Again a pruning attempt was made, and the tree was unchanged. The repetition process was applied; for the pruned trees, the mean accuracy was 0.8887 with a standard deviation of 0.0105, whereas the pruned trees gave a mean accuracy of 0.8891 with a standard deviation of 0.0101.

The two trees were unwieldy for practical use, so feature selection methods (χ^2, symmetrical uncertainty, and gain ratio) were applied to the data sets and the trees were created again. The accuracy resulting from each method for data set 1 is shown in Table 3, and the accuracy for data set 2 is in Table 4.

TABLE 1. CONFUSION MATRIX FOR TREE 1

	True	
Predicted	0	1
0	459	97
1	11	34

In either case, the highest levels of accuracy achieved following the application of the feature selection methods was only negligably different from the accuracy obtained by training a tree against the full feature set.

K-means was applied to Data Set 1 for k = 3, and a decision tree was trained on each of the three clusters to a maximum tree height of 3. The accuracy for each of the three resulting trees is listed in Table 5. These steps were repeated with Data Set 2; the accuracy is listed in Table 6.

The decision trees created from the 3 clusters in Data Set 2 were limited to a maximum depth of 3 levels. These trees are shown in Figure 3, Figure 4, and Figure 5.

Note any interesting outputs of the trees

6. Future Work

The ensemble method described in [6] included a step where the decision tree's depths were limited and then the a priori association mining algorithm was run against the data falling into the leaf nodes of the tree. Therefore a continuation of the work described in this paper should include an association mining step. In order to apply a priori against the data, one should complete a few major steps: first, the data should be converted into item sets for input into a priori; second, the decision tree should be "linearized" [10] to obtain production rules; and third, the data should be labeled with the corresponding production rules. Upon completion of these steps, one could run a priori against each subset, select the relevant association rules, and combine those association rules with the decision tree rules to measure their performance together.

Some of this work has been completed: the pipeline contains optional steps for converting the original exposome data into quintiles and from there into binary incidence matrices that can be passed to the a priori algorithm.

7. Conclusion

The shortcomings of this project's output underscore the need for a programmer to work closely with domain experts when conducting data modeling work: as in the case of this project, the programmer may focus much attention on the execution of the technical work, whereas only an expert in the problem domain can help the programmer determine whether the results will fulfill any practical needs.

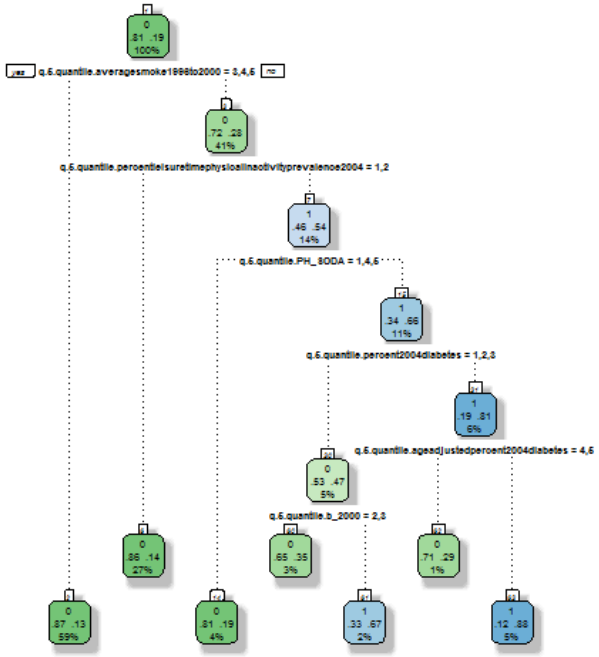

Figure 1. Decision tree based on paraclique features, no clustering

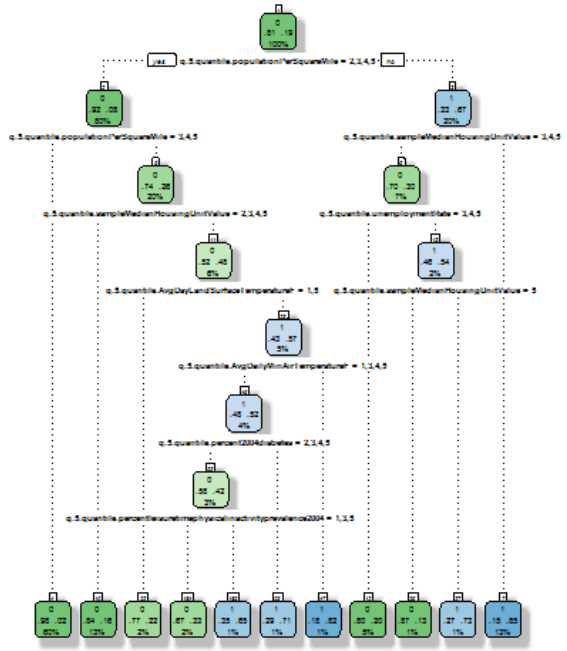

Figure 2. Decision tree based on statistical features, no clustering

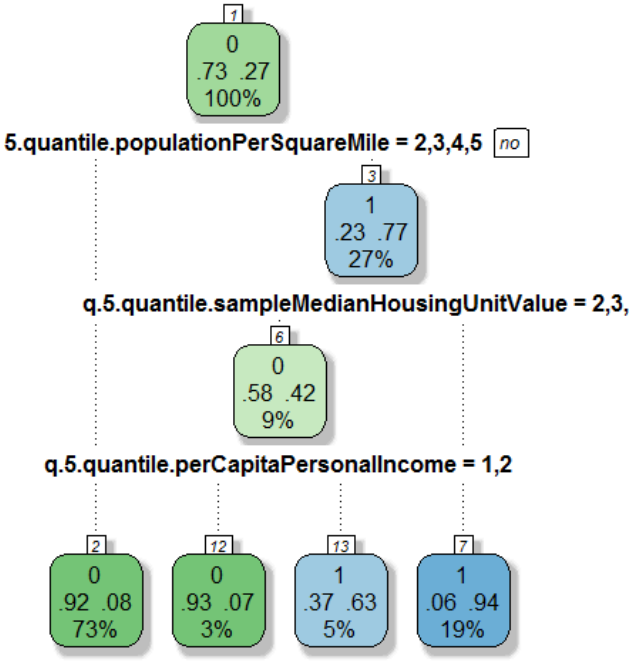

Figure 3. Decision tree based on statistical features, no clustering

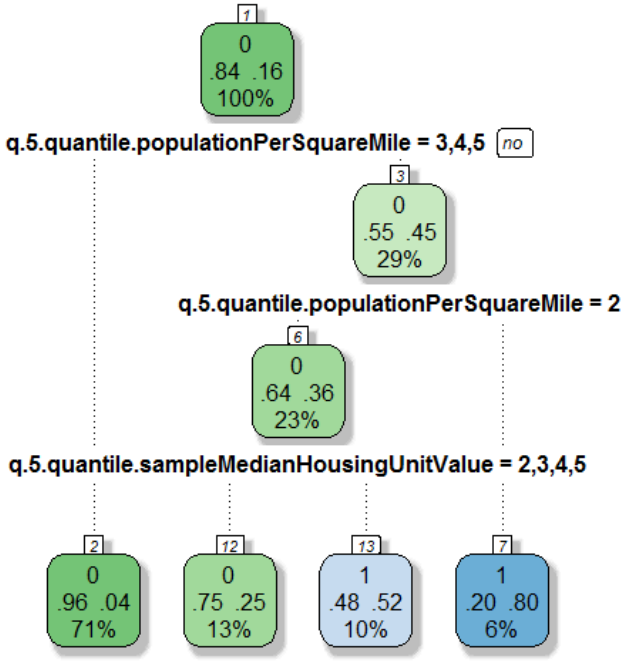

Figure 4. Decision tree based on statistical features, no clustering

References

[1] P. Juarez, P. Matthews-Juarez, D. Hood, W. Im, R. Levine, B. Kilbourne, M. Langston, M. Al-Hamdan, W. Crosson, M. Estes, S. Estes, V. Agboto, P. Robinson, S. Wilson, and M. Lichtveld, "The public health exposome: A population-based, exposure science approach to

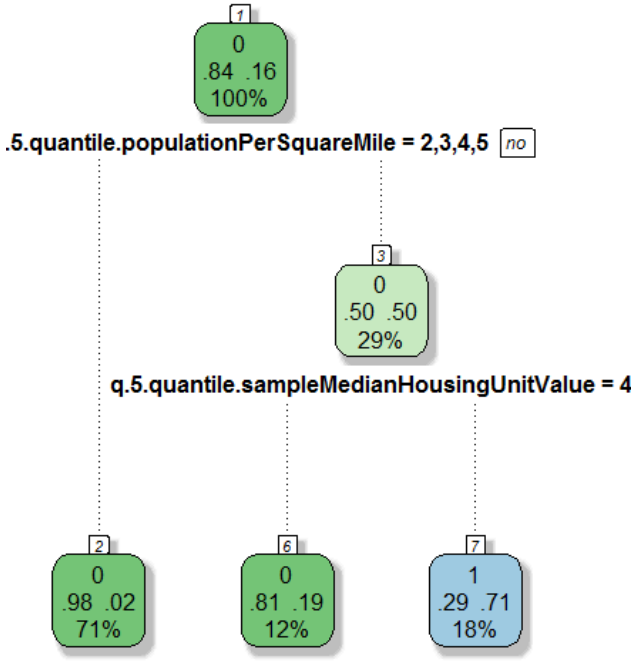

Figure 5. Decision tree based on statistical features, no clustering

TABLE 2. CONFUSION MATRIX FOR TREE 2

	True	
Predicted	0	1
0	445	45
1	25	86

TABLE 3. ACCURACY FROM FEATURE REDUCTION METHODS ON DATA SET 1

Method	Unpruned Accuracy	Pruned Accuracy
χ^2	0.8040	0.7995
Symmetrical Uncertainty	0.8317	0.8316
Gain Ratio	0.8317	0.8316

TABLE 4. ACCURACY FROM FEATURE REDUCTION METHODS ON DATA SET 2

Method	Unpruned Accuracy	Pruned Accuracy
χ^2	0.8892	0.8893
Symmetrical Uncertainty	0.8887	0.8888
Gain Ratio	0.8887	0.8888

TABLE 5. ACCURACY OF TREES FROM K-MEANS CLUSTERS ON DATA SET 1

Cluster #	Unpruned Accuracy	Pruned Accuracy
1	0.8190	0.8189
2	0.8438	0.8441
3	0.8456	0.8407

TABLE 6. ACCURACY OF TREES FROM K-MEANS CLUSTERS ON DATA SET 2

Cluster #	Unpruned Accuracy	Pruned Accuracy
1	0.8669	0.8690
2	0.7983	0.7919
3	0.9919	0.9919

health disparities research," *International Journal of Environmental Research and Public Health*, vol. 11, pp. 12866–12895, 12 2014.

[2] A. Kershenbaum, M. Langston, R. Levine, A. Saxton, T. Oyana, B. Kilbourne, G. Rogers, L. Gittner, S. Baktash, P. Matthews-Juarez, and P. Juarez, "Exploration of preterm birth rates using the public health exposome database and computational analysis methods," *International Journal of Environmental Research and Public Health*, vol. 11, pp. 12346–12366, 11 2014.

[3] R. Raman, "Epigenetics and the exposomes: Obesity and beyond," *International Journal of Nutrology*, vol. 6, no. 3, 2013.

[4] M. Bramer, *Principles of Data Mining*. 2013.

[5] C. C. Aggarwal, *Data Mining: The Textbook*. 2015.

[6] S. Datta, *A Multi-Stage Decision Algorithm for Rule Generation for Minority Class*. PhD thesis, Texas Tech University, Lubbock, TX, 8 2014.

[7] P. Romanski and L. Kotthoff, "Package 'FSelector'." Available at https://cran.r-project.org/web/packages/FSelector/FSelector.pdf (2017/04/23).

[8] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[9] T. Therneau and B. Atkinson, "Package 'rpart'." Available at https://cran.r-project.org/web/packages/rpart/rpart.pdf (2017/04/23).

[10] J. Quinlan, "Simplifying decision trees," *International Journal of Man-Machine Studies*, vol. 27, no. 3, pp. 221 – 234, 1987.