Predicting Suicidal Ideation using Symbolic and Statistical Machine Learning: National Youth Risk Behavior Survey (YRBS) - United States, 2011 & 2013

Abstract

Globally, child and adolescent suicide is a significant public health issue. In many countries it is one of the leading causes of death in youths. It is said that the majority of suicide completers and attempters never reach the attention of specialty mental health care. Reflecting that more effective school health initiates, policy and programmatic interventions are needed. And also highlighting the necessity of more dynamic scalable screening tools for suicidality. Rules, observed to be one of the most transparent knowledge representations is widely used in machine learning for rendering clinical decision support. However, in terms of prediction performance, feature set compactness and data understanding other algorithms, such as random forest, are increasingly useful for dealing with high dimensional datasets.

Introduction

The Youth Risk Behavior Surveillance System (YRBSS) is an epidemiologic surveillance system established by the Centers for Disease Control and Prevention (CDC). It monitors the prevalence of priority health-risk behaviors among youth and young adults. Interrelated and preventable, these priority health-risk behaviors (alcohol and drug use etc.) are associated with the leading causes of morbidity and mortality in this segment of the population [1]. YRBSS includes a national school-based Youth Risk Behavior Survey (YRBS) conducted by the CDC in conjunction with state and local education and health agencies. The national survey targets high school student, grades 9-12[1].

Since 1991, the earliest year of data collection, long term temporal changes in the prevalence of many priority health-risk behaviors have occurred, such as a decrease in current cigarette use, and current sexual activity. However, the prevalence of other health-risk behaviors has not changed, suicide attempts being one of them. Results from the 2013 national YRBS indicate that 8.0% of respondents had attempted suicide during the last 12 months. Variations by sex, race/ethnicity, and grade were observed, as in the case of the other assessed health-risk behaviors [2]. Globally, child and adolescent suicide is a significant public health issue. In many countries it is one of the leading causes of death in youths [3].

"Suicidal ideation" refers to thoughts of harming or killing oneself. It is a frequent precursor to suicide, defined as a fatal self-inflicted destructive act with explicit or inferred intent to die. The point prevalence of suicidal ideation in adolescence is approximately 15–25%. Suicidality refers to all suicide-related behaviors and thoughts: completing or attempting suicide, suicidal ideation or communications. Estimates of the risk of repetition of suicidal behavior range from 10 upon a 6-month follow-up to 42% upon 21-month follow-up [3]. It is said that the majority of suicide

completers and attempters never reach the attention of specialty mental health care [3]. Reflecting that more effective school health initiates, policy and programmatic interventions are needed [1]. And also highlighting the necessity of more dynamic scalable screening tools for suicidality.

Machine learning entails the discovery of models, patterns, and other regularities in data. It can be roughly categorized into two different typologies: symbolic approaches (involving inductive learning of symbolic descriptions, such as rules and decision trees) and statistical approaches (involving statistical or pattern-recognition methods, such as Bayesian classifiers and support vector machines). In supervised learning, models or theories (such as decision trees or rule sets) are induced from class-labeled data [4]. Most feature selection algorithms, used in machine learning don't strive to model mechanisms or uncover cause effect relationships between feature(s) and target, simply because it is not a requisite for making good predictions in a purely observational setting [5]. Yet, feature selection has been advanced as being the de facto standard for inducing causal hypotheses from massive datasets. It is recognized that feature selection and causal discovery share two essential goals, which are making good predictions and data understanding(identifying factors relevant to a target) [5]. There exists a wide variety of feature selection algorithms and a seeming lack of consensus as to which one works the best. This paper aims to uncover the strengths and weakness of variety of symbolic and statistical machine learning algorithms, in term of prediction performance, feature set compactness and data understanding [5]. The system studied is suicidal ideation as featured in the National Youth Risk Behavior Survey (YRBS) - United States, 2011 & 2013.

Materials and Methods

Data source

In this study, the 2011 and 2013 National Youth Risk Behavior Survey (YRBS) were used to generate a sampler of supervised leaners. The National Youth Risk Behavior Survey (YRBS) uses a three-stage cluster sample design to produce a representative sample of 9th through 12th grade students in the United States [2]. More information on the YRBSS methodology can be obtained at www.cdc.gov/yrbss.

"Youth Risk Behavior Survey (YRBS) data are available in two file formats: Access® and ASCII. The Access and ASCII data can be downloaded and used as is... Additionally, SAS® and SPSS® programs are provided to convert the ASCII data into SAS® and SPSS® datasets for use in those packages." http://www.cdc.gov/healthyyouth/data/yrbs/data.htm

Variable selection

Dichotomous variables (QN#) that corresponded to the standard question variables, along with standard question variables pertaining to dietary behavior were omitted from analysis, in an effort to reduce noise in the data.

Total sample size of the 2011 and 2013 dataset were 15425 obs. (207 variables) and 13583 obs. (213 variables), respectively.

Target Binary Variable ("Suicidal ideation")

During the past 12 months, did you ever seriously consider attempting suicide?

A. Yes B. No

Other Suicidality Variables:

During the past 12 months, did you make a plan about how you would attempt suicide?

A. Yes B. No

During the past 12 months, how many times did you actually attempt suicide?

A. 0 times B. 1 time C. 2 or 3 times D. 4 or 5 times E. 6 or more times

If you attempted suicide during the past 12 months, did any attempt result in an injury, poisoning, or overdose that had to be treated by a doctor or nurse?

A. I did not attempt suicide during the past 12 months B. Yes C. No

Model Generation

For statistical machine learners, each dataset was divided into two non-overlapped samples: a training set (2/3 of cases, to fit model) and a testing set (1/3 of cases, to check prediction accuracy).

CRAN Packages Use for Machine Learning were:

e1071: Support vector machines, naive Bayes classifier etc (SVM algorithm)

rpart: Recursive Partitioning and Regression Trees (CART algorithm)

randomForest: Random forests for classification and regression

arules: Mining Association Rules and Frequent Itemsets (Apriori algorithm)

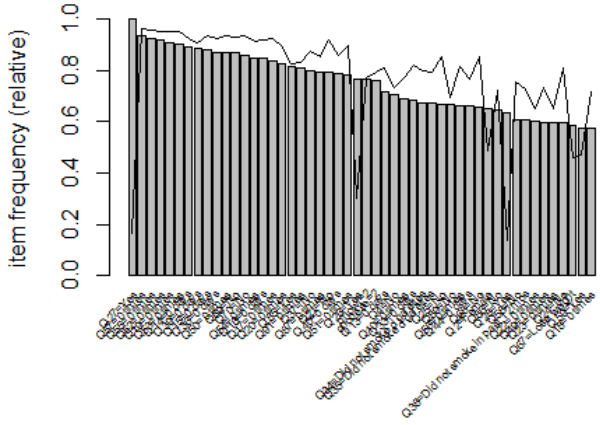
Version Notes

The calculations in this paper were performed with

R version 3.2.1.

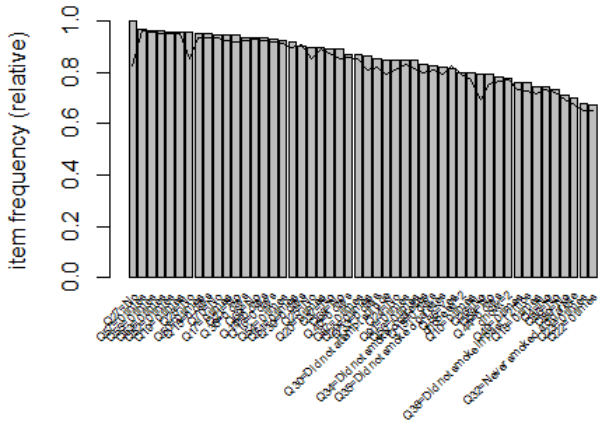
Results and Discussion

Figure 1: Item Frequency Plot, Top 50



Graph shows the top 50 most frequent items in respondents claiming to experience suicidal thoughts within the last 12 months (2259 students, Q27="Yes")

Figure 2: Item Frequency Plot, Top 50



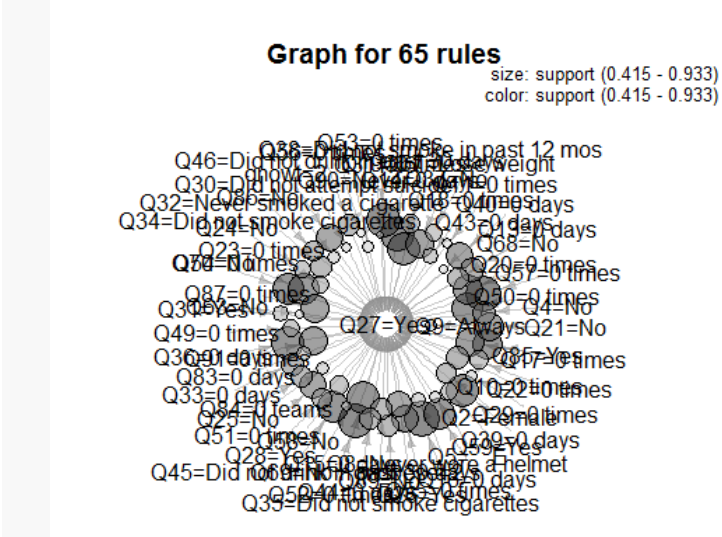
Graph shows the top 50 most frequent items in a matched sample population claiming not having any suicidal thoughts within the last 12 months (2259 random students, Q27="No")

Table 1: Sample list of rule finding the antecedents for suicidal ideation (rhs="Q27=Yes")

```
##   lhs                          rhs          support    confidence    lift
## 53 {Q85=Yes}                => {Q27=Yes} 0.8162904  1          1
## 46 {Q26=Yes}                => {Q27=Yes} 0.7671536  1          1
## 45 {qnowt=2}                => {Q27=Yes} 0.7653829  1          1
## 44 {qnobese=2}              => {Q27=Yes} 0.7605135  1          1
## 32 {Q2=Female}              => {Q27=Yes} 0.6502877  1          1
## 30 {Q28=Yes}                => {Q27=Yes} 0.6347942  1          1
## 23 {Q67=Lose weight}        => {Q27=Yes} 0.5869854  1          1
## 22 {Q59=Yes}                => {Q27=Yes} 0.5776892  1          1
## 20 {Q31=Yes}                => {Q27=Yes} 0.5528995  1          1
## 12 {Q9=Always}              => {Q27=Yes} 0.4710049  1          1
```

A set of 15222 rules were generated for rhs="Q27=Yes", where the minimum support was set to 0.4, confidence was set to 0.9, and maximum length to 4 items. The maximum support value achieved was 0.933. And the maximum value for confidence and lift was 1.

Figure 3: Antecedents for suicidal ideation (rhs="Q27=Yes")



The learning of rule-based models has been a main activity in the field of machine learning since its inception in the early 1960. A rule-based classification model consists of a set of if–then rules. Each rule has a conjunction of attribute values in the conditional part of the rule, and a class label in the rule consequent. [1]. The Apriori algorithm requires specification of two thresholds, the minimal support and the minimal confidence, to output rules that exceed the thresholds prompted by the user. Thus, it is contended that the user must possess a certain amount of expertise in order to find the right parameter settings to obtain the best rules [7]. The Apriori algorithms generated an extremely large number of

association rules. Traditional techniques, such as constraining item set length, were used to improve the comprehensibility of discovered rules. Nevertheless, the set of association rules remained unwieldly, making validation seemingly impossible. The usefulness of the data mining results were undermined [7, 8]. Prediction Speed was slow to intractable, memory usage large, however, interpretability was moderate.

Table 2: Classification: Naïve Bayes: Suicidal Ideation

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |-------------------------|
##
## ==================================
##              actual
## predicted      Yes      No   Total
## ----------------------------------
## Yes            676     279     955
##              0.708   0.292   0.214
##              0.903   0.075
## ----------------------------------
## No              73    3426    3499
##              0.021   0.979   0.786
##              0.097   0.925
## ----------------------------------
## Total          749    3705    4454
```

Prediction accuracy of classes between 90 and 92%

- Naïve Bayes, prediction speed was medium, memory usage small and interpretability was easy. The major drawback of this algorithm was that its simple representation does not support data understanding.

Figure 4: Decision Tree

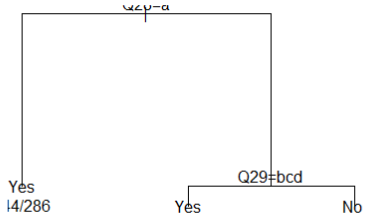


Table 3: Contingency Table Decision Tree

```
table(prediction, testdt$Q27)

##
## prediction  Yes   No
##        Yes  548  158
##        No   201 3547
```

Decision Tree, prediction speed was fast, memory usage small and interpretability was easy. The major failing of this algorithm was that it was too greedy, potentially over-fitting data. It does not support rich data understanding and risk to being generalizable.

Table 4: Cross Validation Using Random Subsampling and Random Forest, with Rough Imputation of Missing

```
## prediction  Yes   No
##        Yes  494  115
##        No   251 3621
## [1] 0.9183218
##
## prediction  Yes   No
##        Yes  499  134
##        No   246 3602
## [1] 0.9151975
##
## prediction  Yes   No
##        Yes  502  125
##        No   243 3611
## [1] 0.9178755

accuracies

## [1] 0.9183218 0.9151975 0.9178755

mean(accuracies)

## [1] 0.9171316
```

Table 5: Tree Ensembles, with Rough Imputation of Missing Values

```
table(prediction, yrbs2013siTest$Q27)

##
## prediction  Yes   No
##        Yes  505  120
##        No   240 3616

importance (model)
```

Table 6: Mean Decrease Gini

```
##          MeanDecreaseGini
## Q7             31.044000
## Q24            31.038757
## Q26           151.615186
## Q28           395.153515
## Q29           196.169407
## Q30           207.415659
## Q78            31.205483
## Q79            37.384864
## Q80            39.286879
## Q81            35.079109
## Q82            36.004142
## Q88            33.441511
## Q92            34.264111
## bmipct         34.581384
```

Random Forest, prediction speed was fast, memory usage small and interpretability was moderate. In term of prediction performance, feature set compactness and data understanding this was the best performing algorithm.

Table 7: Support Vector Machine

```
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  1
##
## Number of Support Vectors:  9101
##
##  ( 7587 1514 )
##
##
## Number of Classes:  2
##
## Levels:
##  Yes No
```

```
##       true
## pred   Yes   No
##   Yes    0    0
##   No   745 3736
```

Support Vector Machine, prediction speed was slow, memory usage was large and interpretability was hard. This SVM failed to predict the class “Yes”, most likely due to the insufficient training data and the huge class imbalance.

Conclusion

Decision support systems and knowledge discovery are the primary applications of machine learning in healthcare. Wojtusiak, J. (2014), contends that “successful application of machine learning in healthcare requires accuracy, transparency, acceptability, ability to deal with complex data, ability to deal with background knowledge, efficiency, and exportability” to be attributes of any implemented learning models. As observed in this paper, acceptance of high-dimensional predictive data-analytic models continues to be hindered by their poor interpretation capability.

References

1. Eaton, D.K., et al., *Youth risk behavior surveillance-United States, 2011.* Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002), 2012. **61**(4): p. 1-162.
2. Kann, L., et al., *Youth risk behavior surveillance—United States, 2013.* MMWR Surveill Summ, 2014. **63**(Suppl 4): p. 1-168.

3. Bridge, J.A., T.R. Goldstein, and D.A. Brent, *Adolescent suicide and suicidal behavior.* Journal of Child Psychology and Psychiatry, 2006. **47**(3-4): p. 372-394.
4. Fürnkranz, J. and D. Gamberger, *Foundations of rule learning*. 2012: Springer Science & Business Media.
5. Guyon, I., C. Aliferis, and A. Elisseeff, *Causal feature selection.* Computational methods of feature selection, 2007: p. 63-86.
6. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* The Journal of Machine Learning Research, 2003. **3**: p. 1157-1182.
7. García, E., et al. *Drawbacks and solutions of applying association rule mining in learning management systems*. in *Proceedings of the International Workshop on Applying Data Mining in e-Learning (ADML 2007), Crete, Greece*. 2007.
8. Kotsiantis, S. and D. Kanellopoulos, *Association rules mining: A recent overview.* GESTS International Transactions on Computer Science and Engineering, 2006. **32**(1): p. 71-82.