Nursing Informatics 2014
K. Saranto et al. (Eds.)

doi:10.3233/978-1-61499-415-2-71

Using a Data Mining Approach to Discover Behavior Correlates of Chronic Disease: A Case Study of Depression

Sunmoo YOON[a,b,1], Basirah TAHA[a,c] and Suzanne BAKKEN[a,b]
[a]*School of Nursing,* [b]*Department of Biomedical Informatics, Columbia University, New York, NY, USA*
[c]*Rutgers University Behavioral Healthcare, Newark, NJ, USA*

Abstract. The purposes of this methodological paper are: 1) to describe data mining methods for building a classification model for a chronic disease using a U.S. behavior risk factor data set, and 2) to illustrate application of the methods using a case study of depressive disorder. Methods described include: 1) six steps of data mining to build a disease model using classification techniques, 2) an innovative approach to analyzing high-dimensionality data, and 3) a visualization strategy to communicate with clinicians who are unfamiliar with advanced statistics. Our application of data mining strategies identified childhood experience living with mentally ill and sexual abuse, and limited usual activity as the strongest correlates of depression among hundreds variables. The methods that we applied may be useful to others wishing to build a classification model from complex, large volume datasets for other health conditions.

Keywords. Data mining, knowledge discovery, depression, modeling, classification, high-dimensional data

Introduction

Today, many health-related organizations have large and complex datasets, which are difficult to process using traditional applications. The enhanced data mining/machine learning approaches offer advanced techniques for handling of complex data, and allow researchers to mine large volumes of the complex data [1; 8]. With the help of machine learning, artificial intelligence and computing power, researchers can investigate correlates from a broad range of datasets. For example, whereas it takes more than several minutes to handle a dataset consisting 400 attributes and 500,000 records using a conventional statistics software package such as SPSS or SAS, it takes less than a minute to handle a large dataset with data mining software such as R or Weka.

Data mining is not a single computing step for discovering new knowledge through applying algorithms [4; 8]. Instead, it is an interdisciplinary discovery process with multiple steps requiring input from domain expertise throughout the whole discovery process (e.g., selecting the correct database, deciding clinically meaningful

[1] Corresponding Author: Sunmoo Yoon, RN, PhD, 617 W 168 Street, New York, NY, U.S.; E-mail: sy2102@columbia.edu.

variables). The interdisciplinary component is vital to discover more comprehensive knowledge. For this reason, this study specifically describes how, when and where the domain experts are involved in the process of data mining. Further, one of the barriers that data mining has faced is efficiently and effectively communicating with domain experts such as clinicians, who are often unfamiliar with output of advanced statistics. Visualization of the output with images and diagrams may help the communication process between researchers and practitioners [6]. The purpose of study was to build a model for depressive disorder by applying data mining methods to a U.S. national behavioral risk factor data set.

1. Data Mining Methods With a Case of Depressive Disorder

1.1. Conceptual Framework

The data mining and knowledge discovery process model by Fayyad guides the data mining process [4]. This robust framework (Figure 1) depicts the steps of data mining including: 1) understanding and developing domain application, 2) selecting and understanding data, 3) cleaning and preprocessing data 4) reducing and projecting data, 5) choosing data mining task and algorithms and 6) interpreting and evaluating results. Domain expertise is required in every step.

1.2. Data and Tools

We applied classification methods of data mining to the U.S. Behavioral Risk Factor Surveillance System (BRFSS) in order to build a classification model for depression. BRFSS is a random annual phone-based (land-line and cell-phone) health survey tracking health behaviors and condition in the U.S. (http://www.cdc.gov/brfss/). The most up-to-date BRFSS data (2011-2012 cycles as of June 2013) were used for this study. PSAW SPSS version 20 was used for importing data, part of reducing the dimensionality of the dataset and converting the file format to 'csv', which is readable in Weka software. This study used Weka and R, a publically available data mining software, to assist in selection of variables and building a classification model (Weka available at http://www.cs.waikato.ac.nz/ml/Weka/downloading.html, R is available at http://www.r-project.org/).

1.3. Application of Data Mining for Depressive Disorder

In order to build a classification model for depression, this study followed the six analytic steps iteratively (Figure 1).

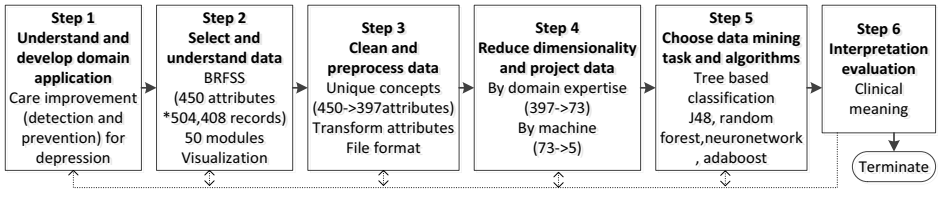


Figure1. Steps of data mining approach for building a classification model for depressive disorder

1.3.1. Step 1. Understanding and developing domain application

Diagnosis of depression has been a challenge to clinicians. Due to its variation of clinical manifestations, a depression diagnosis is often difficult to make. If classification of the disease is more accurate, providers can diagnose, treat and prevent the disease better [2]. The domain application for the model built in this study intends to assist clinicians to improve care for depressive disorder.

1.3.2. Step 2. Selecting and understanding data

Depression is a complex illness with many contributing factors including biological, personal behavior and social behaviors. Considering the complexity of the disease, BRFSS, which is one of the most comprehensive behavior data sets in the U.S., was chosen among other similar behavior data sets by domain experts in health surveys. Further, our outcome variable was operationalized from the question "Ever diagnosed with depressive disorder including depression, major depression, dysthymia or minor depression" question in the BRFSS. In order to understand the dataset, each variable was visualized to initially assess the distribution and/or odds ratio of each variable regarding the outcome using Weka (Figure 2). Some clinically meaningful factors showed a high association with depression during this initial assessment process. Figure 2 demonstrates some examples of the higher rate of depression among people who have low income, snoring, joint pain, and smoking.

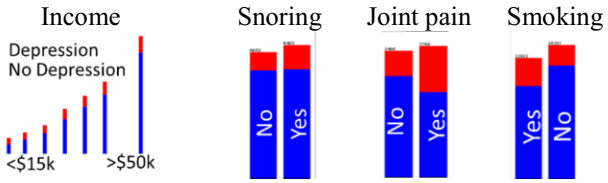


Figure 2. Initial visualization of depression distribution (red: depression, blue: no depression)

1.3.3. Step 3. Cleaning and preprocessing data

During the data cleaning step, conceptually duplicate variables (e.g., calculated variables) were removed by a BRFSS domain expert (SY). Using SPSS, file was converted from *xpt* (SAS transport format) format into *csv* (comma-separated value) which is a readable form in Weka. Then the file was saved as arfss (attribute-relation file format) within Weka. The data integrity of the arfss file (@attribute <attribute-name> <data type>) was checked using an editing software (e.g., notepad++; http://notepad-plus-plus.org/) for the further correction of the dataset. In the depressive disorder case, after 53 duplicate variables (e.g., calculated variable) were removed, 397 unique variables were left from the initial 450 variables.

1.3.4. Step 4. Reducing dimensionality and projecting data

According to the definition of high dimensional data as the one containing 100s-1000s attributes for each record [3], BRFSS is a high dimensional dataset containing 450 nearly unique attributes per record. The size of the dimensionality of the selected

dataset was noted as 500,000 (Euclidian distance, Framingham heart survey – 25,000). Unlike the traditional statistical analysis, variables were not predetermined in the data mining process. Instead, most variables were included without predetermination for the analysis. The study applied a hybrid (human + machine) approach to reduce dimensionality of the dataset (Figure 3).

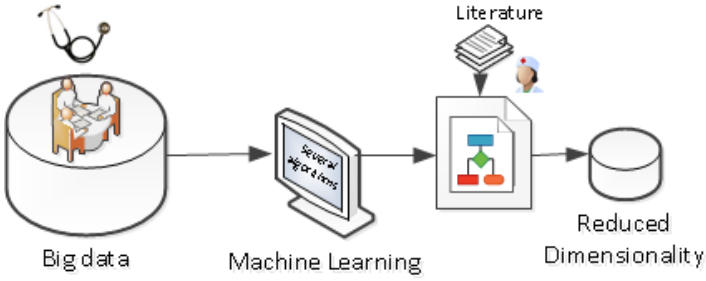


Figure 3. Reducing dimensionality of big data

By human filtering: First, the dimensionality of the data was reduced by manually filtering the variables which met the following criteria; a) completely irrelevant variables, which have no clinical or research implication (e.g., emergency preparedness), b) very highly correlated variables (e.g, "has a doctor or other health professional ever told you that you have depression, anxiety or post-traumatic stress disorder?" is similar to the outcome variable), c) redundant variables (e.g., age categories, race categories). Manual reduction process by BRFSS domain experts resulted in 73 variables from the initial 450.

By machine leaning: Several machine learning algorithms including correlation feature selection (CFS), correlation attributes evaluator (CAE), principal component analysis (PCA) and multiple-regression were applied to reduce the dimensionality using Weka. CFS attribute evaluator is used to find features which are highly correlated with the class, yet uncorrelated with each other based on three correlation measures (Minimum Description Length, Symmetrical Uncertainty and Relief) instead of common correlation coefficients (Pearson's r or Spearman's ρ) [5]. On the contrary, CAE simply evaluates variables by measuring the correlation (Pearson's) between it and the class. PCA performs a principal components analysis, and chooses enough eigenvectors to account for some percentage of the variance in the original data. The coefficients of each variable calculated by machine learning regression were visualized with a heat-map in order to simply represent the outcome of the analysis for the better communication with a clinical domain expert who is unfamiliar with advanced statistics and machine learning. The domain experts compared variables generated by each algorithm to the known variables from literature, and selected the final set of variables to build a classification model. The different machine learning algorithms resulted in various numbers of attributes to project the data (Table 1). A heap-map was created to represent the level of the influence of different variables on depressive disorder (Figure 4).

Table 1. Examples of number of variables selected by different algorithms for depressive disorder

Methods			
correlation feature selection	**correlation attributes evaluator**	**principal component analysis**	**literature**
5 variables	All variables by ranking	161 formula	23 variables
Childhood- sexually touched Own home Limited usual activity Mental health days	Mental health days (83) Employment (82) Poor health days (81) Childhood sexually touched (81)…	General health Arthritis burden-social Interaction Physical health Limited usual activity…	smoking, sleep deprivation, friendship , instability, pain …

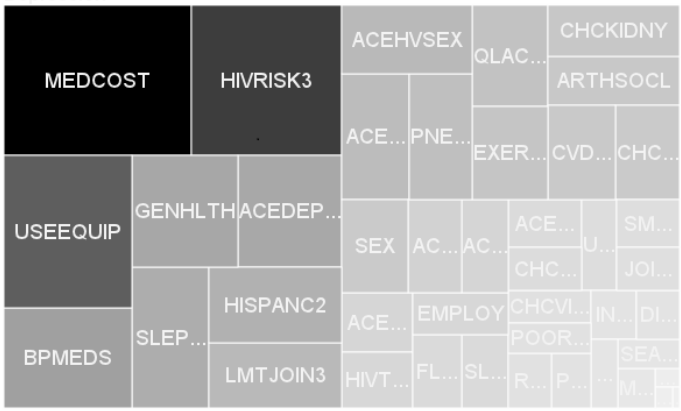


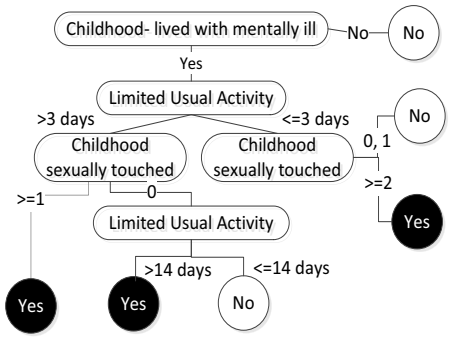
Figure 4. Heat map visualization of depressive disorder

After several iterations of comparing the results from the different sources, clinical and BRFSS domain experts selected the following three, which were repeatedly appeared from the different methods in Table 1, as the final set for the next modeling process;living with anyone who was depressed, mentally ill or suicidal, 2) childhood adverse experience – sexually touched and 3) days of limited usual activities, such as self-care, work, or recreation.

1.3.5. Step 5. Choosing data mining task and algorithms

In this step, classification models were built with the final set of the selected variables. Several classification algorithms including J48, Random Forest, Multilayer Perception, AdaboostM1, and Support Vector Machine were applied to iteratively generate classification models in order to avoid algorithm dependency. J48 generates a decision tree using a classification tree, C4.5 in Weka [7]. Although studies report that ensemble-based methods (e.g. Random Forest) outweigh the benefits of classification trees, the model built by classification trees (J48) was chosen because it is relatively easier to interpret the results with the tree visualization. In order to validate the model, 10-fold cross validation was applied to randomly partition a dataset for training and testing. We evaluated the model's performance each time using proportion correctly classified (model accuracy rates 80-82%) and the area of under the receiver operating characteristic curve (AUC: 0.70-0.72). A final prediction model was chosen based on

predictability. The complicated tree model built by J48 was re-illustrated according to clinicians' request in a simplified way to facilitate interpretation and translation of the knowledge into practice. J48 decision tree algorithm computed the final set of variables and depicted 'childhood experience living with mentally ill' as the most related variable to current depression diagnosis. The model by J48 shows that individuals with the adverse experience of living with mentally ill or being sexually touched in childhood are likely to be depressed if they complain that their usual activity was limited for greater than three days. Figure 5 shows the two different presentations of the model displayed using a diagram (left) and using a logistic regression formula (right).



Logistic Regression

Variable	Coefficients
Childhood-living with mentally ill	+ 1.30
Childhood-sexually touched>once	+ 0.57
Childhood-sexually touched =never	- 0.46
Childhood-sexually touched=once	+ 0.21
Limited usual activity	+ 0.05
Intercept	1.64

(Accuracy: 82%, Area under ROC=0. 72)

Figure 5. Visualization (left) versus conventional presentation (right) of disease modeling using a depression case

1.3.6. Step 6. Interpreting and evaluating the results

In this step, domain experts critically interpret the results of the data mining according to the clinical meaningfulness and iterate the process if it's necessary. In the depressive disorder case, clinical domain experts evaluated the implication and clinical significance of computed model. Among some documented predictors (e.g., age, marriage status, insomnia), clinical domain experts noted that modifiable factors (e.g., pain, insomnia or smoking) are more important for clinical practice than demographic factors. The data mining process was terminated after several iterations and confirmation by clinical domain experts.

Discussion

The data mining methods described in this paper support analytic strategies for building a disease classification model from a large and complex dataset using a case study of depressive disorder. The paper also highlights innovative approaches to reduce the dimensionality. Furthermore, this study used a heat-map and flow chart in order to encourage participation of clinical domain experts who are critical for the mining process yet unfamiliar with machine learning or advanced statistics. We further discuss how clinical domain expertise was integrated into the classification modeling process and the implications for the data mining researchers.

Clinical meaningfulness

Figure 5 shows that even if an individual has an adverse experience living with mentally ill during childhood, if their usual activity is limited less than half of the days of the month, then the individual is more likely not depressed when the person has not experienced childhood sexual abuse. During the step 6 interpretation process, clinical domain experts pointed out how this might be useful in their practice; clinicians should routinely assess individual daily functional status. In fact, the clinical meaningfulness of function variables (e.g., days of physical illness, activity limitation due to health problems, or use of equipment) was supported by the different machine learning algorithms during the data mining process. In particular, principal component analysis revealed several variables related to functional burdens (e.g., limited because of joint symptoms, social activities limited because of joint symptoms). Further, clinical domain experts indicated that individual function can be limited due to physical and mental reasons. With this guidance, further logistic regression analysis was conducted using a separate concept (functional limitation due to physical illness). Although the coefficient of physical function (0.05) was smaller than the childhood adverse events (0.21-1.30) in the built model, clinical domain experts emphasized that clinical meaningfulness of this variable; assessing functional days (e.g., days of limited usual activities due to physical illness such as pain) underlying the source of physical illness should not be overlooked because physical illness such as pain also influences the progress of depression treatment.

Implications for mining researchers

This study specifically emphasizes data mining as an interdisciplinary process. For this reason, we described how, when and where clinical domain experts were involved in the process of data mining. Some visualization techniques (e.g., heat map, initial exploratory visualization, illustrative decision tree) can be a useful strategy for the researchers to collaborate with clinical domain experts who can translate the knowledge into the practice. One of the problems around data mining/machine learning is the interpretation issue of correlation and causation. Classification methods (e.g., multiple regression) applied in this study are categorized as "predictive modeling" by some authors [8; 9]. Although the plausibility of most of variables (childhood adverse experience living with mentally ill and sexually touched) in our final model are clear, the direction of causation between a variable (the days of limited usual activity) and depression remains unclear; domain experts confirmed it as bidirectional relationship. For this reason, the data mining results are considered a correlational model instead of a prediction model.

Conclusion

The data mining methods that we applied may be useful to others wishing to mine a large and complex behavior dataset for other chronic conditions.

Acknowledgments

This study was funded by the Washington Heights Inwood Informatics Infrastructure for Comparative Effectiveness Research (R01 HS019853, Suzanne Bakken, Principal Investigator)

References

[1] R. Alizadehsani, J. Habibi, M.J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, and Z.A. Sani, A data mining approach for diagnosis of coronary artery disease, *Comput Methods Programs Biomed* **111** (2013), 52-61.
[2] P.C. Austin, J.V. Tu, J.E. Ho, D. Levy, and D.S. Lee, Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes, *J Clin Epidemiol* **66** (2013), 398-407.
[3] R. Clarke, H.W. Ressom, A. Wang, J. Xuan, M.C. Liu, E.A. Gehan, and Y. Wang, The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, *Nat Rev Cancer* **8** (2008), 37-49.
[4] U. Fayyad, G. Piatetsky-Shapiro, P. Smith, and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI, MIT Press, MA, 1996.
[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, The weka data mining software: An update, *SIGKDD Explorations* **11** (2009), 10-18.
[6] G.J. Myatt and W.P. Johnson, *Making sense of data III: a practical guide to designing interactive data visualizations*, A john Wiley & Sons, INC., 2012.
[7] R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
[8] P. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*, Addison Wesley, 2006.
[9] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*, Morgan Kaufmann, San Francisco, CA, 2005