HHS Public Access

Author manuscript

IFMBE Proc. Author manuscript; available in PMC 2016 August 30.

Published in final edited form as:

IFMBE Proc. 2014 ; 42: 114–117. doi:10.1007/978-3-319-03005-0_29.

PHARM – Association Rule Mining for Predictive Health

Chih-Wen Cheng[1], **Greg S. Martin**[2], **Po-Yen Wu**[1], and **May D. Wang**[3]

[1] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

[2] Center for Health Discovery and Well Being, Emory-Georgia Tech Predictive Health Institute, Atlanta, GA, USA

[3] Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA

Abstract

Predictive health is a new and innovative healthcare model that focuses on maintaining health rather than treating diseases. Such a model may benefit from computer-based decision support systems, which provide more quantitative health assessment, enabling more objective advice and action plans from predictive health providers. However, data mining for predictive health is more challenging compared to that for diseases. This is a reason why there are relatively fewer predictive health decision support systems embedded with data mining. The purpose of this study is to research and develop an interactive decision support system, called PHARM, in conjunction with Emory Center for Health Discovery and Well Being (CHDWB®). PHARM adopts association rule mining to generate quantitative and objective rules for health assessment and prediction. A case study results in 12 rules that predict mental illness based on five psychological factors. This study shows the value and usability of the decision support system to prevent the development of potential illness and to prioritize advice and action plans for reducing disease risks.

Keywords

Decision support system; predictive health; health modeling and prediction; association rule mining

I. INTRODUCTION

Health expenditures in the United States reached $2.7 trillion in 2011, over ten times the amount spent in 1980 [1], and the rate is still expected to grow faster than national income over the foreseeable future [2]. As over 90% of the medical spending is for patients with chronic diseases [3], “predictive health” (PH) is a transformation towards maintaining health (rather then treating diseases) by proactively predicting health-related events and disease development, and providing early and persistent interventions before being clinically overt.

Corresponding Author: May D. Wang, Ph.D., maywang@bme.gatech.edu, (404) 385-1790, 313 Ferst Dr. Atlanta, GA 30332.

Recognized as a pioneer in PH, the Center for Health Discovery and Well Being (CHDWB*) was established in 2008 as the major research component of a combined Emory University/ Georgia Institute of Technology strategic initiative: the Predictive Health Institute. [4] The specific goal of the center is to redefine health in a holistic fashion by broadly integrating health-related disciplines (e.g., ethics and sociology) with traditional disciplines of medicine, public health, and nursing through basic and clinical biomedical research. The CHDWB serves as an engine to drive the new healthcare definition through the conduct of a prospective cohort study of predominantly healthy individuals. It establishes a horizontal (as opposed to the traditionally vertical) relationship between participants and health partners who are trained to provide information and support for the participant. A health partner assists participants in completing surveys and other assessments, reviews and explains the health assessment report, helps in setting and achieving their health-related goals, and provides practical advice and moral support.

Currently, health partners are trained using didactic and practical experiences in the knowledge base and skill set. To pursue systematic and objective health advising and planning, the CHDWB has started investigating computer-based decision support systems embedded with advanced data mining methods. However, three major factors make data mining in PH more challenging compared to typical disease-based data mining. First, because the definition of health depends on interacting factors that are not limited to biology, PH data must contain measurements from multiple disciplines to provide a comprehensive picture of human health. However, multi-disciplinary data increases information heterogeneity among measurements, which poses a common challenge in data mining. Second, for PH data, measurements of some variables are highly homogeneous across the healthy cohort. It increases the difficulty of identifying accurate rules or effective separating boundaries for data given a measurement. Third, conventional data mining methods (e.g., support vector machine [5]) heavily depend on data distributions. A distribution-free data mining method may be preferable while dealing with heterogeneous data.

To address the aforementioned three challenges, in this study we present a PH decision support system called Predictive Health Association Rule Mining (PHARM). PHARM is powered by a CHDWB dataset containing reports with 906 measurement variables from a large predominantly healthy cohort. The system features an interactive user interface to perform flexible association rule mining (ARM) to achieve personalized decision support for PH and CHDWB participants.

ii. METHODS AND PROCEDURES

A. CHD WB Dataset

Because most human diseases result from perturbations in common pathways involving oxidative stress, inflammation, and regenerative potential, CHDWB incorporates cutting edge biomarkers in these areas with established and novel assessments of health and healthy behavior. Initiated in May 2008, our current dataset contains 2,637 de-identified health reports from 696 healthy participants with 906 measurement variables. As tabulated in Table 1, each report consists of measurement outcomes, including questionnaires, assessments, physical measurements, laboratory tests, and research laboratory values. Together, these

measurements provide a comprehensive picture of human health, and the ability to discern and detect potential diseases.

B. Overview of Association Rule Mining

After importing the CHDWB dataset, association rule mining (ARM) is applied to unearth meaningful interactions among variables. Association rules are in the form of $X=>Y$, which represents that X (the antecedent) implies Y (the consequent) [6]. In PH data mining, since the most basic data tuple is a health report in one site visit, the rule A=>7 implies that if conditions in X occur in one visit, another set of conditions in Y are also likely to occur in the same visit. For example, a rule $\{SerumCholesterol > TH_{SC} \& SerumGlucose > TH_{SC}\} => \{CardioRisk = High\}$ implies that "in one visit, if a participant's serum cholesterol and serum glucose have been elevated above thresholds TH_{SC} and $THSG$, respectively, his/her cardiovascular risk may also be increased."

ARM has been used before in healthcare settings, such as heart disease prediction [7], healthcare auditing [8], and neurological diagnosis [9] with the following advantages: (1) unlike conventional statistical analysis that only indicates whether the relationship is significant or not (e.g., using p-value), ARM gives each rule a confidence value that determines its strength more quantitatively; (2) a rule composed of an antecedent and a consequent that provides a direction of the relationship; (3) the antecedent and consequent can consist of one or more factors, providing advanced knowledge of complex factor interactions instead of a monotonic relationship (e.g., logistic regression) [10]; and, (4) ARM accepts user-specified inputs, which ensure the strength of each rule to optimize the mining results. However, using ARM in decision support for PH has never been investigated.

C. Association Metrics

Two important metrics—*support* and *confidence*— quantify the frequency and strength of an association rule. The support of an association rule is defined as the fraction of the tuples in the dataset that contain all conditions in X and Y. A high support of an association rule means that a high portion of the dataset is applicable to the rule. The other metric of an association rule is its confidence. It indicates how frequently Y appears in those tuples that contain X. For example, if the confidence of an association rule is 95%, it implies that for data tuples that contain X, 95% of these tuples also contain Y. In other words, confidence reveals the level of the association between X and Y. In order to discover frequent and confident association rules, the mining process requires users to specify a minimum support ($Supp_{\min}$) and a minimum confidence ($Conf_{min}$) to eliminate infrequent and unconfident rules, respectively. Refer to [11] for more detail regarding the generation of frequent itemsets and confident rules.

D. System Use Cases and Interface

The PHARM system features an interactive user interface allowing mining association rules in the PH setting. The interface enables users to define health conditions, mine confident association rules from the CHDWB dataset, and display the rules to the user. The PHARM interface consists of two main windows, including a Rule Mining window and a New Item

window. The interface was implemented in MATLAB. The detailed characteristics of the two windows are described below.

The first main window is the Rule Mining window (Fig. 1) that enables health partners to extract association rules by matching participants' health conditions in predefined sets of antecedents and consequents. Based on these conditions, the system generates all frequent and confident rules from the CFIDWB dataset. However, health partners may not always find conditions of interest in the Rule Mining window. Therefore, the New Item window (Fig. 2) was designed to allow health partners to generate new conditions.

HI. CASE STUDY AND RESULTS

Mental illness encompasses psychological patterns that disrupt an individual's feelings, mood, thinking, daily functioning, and social ability. Survey-based scales are the most common tools to measure the severity of mental disorders. Literature of mental health scales mainly focus on development [12], validation [13], modification for specific populations [14], and comparison between scales [15]. However, to our understanding, there is no research that tries to comprehensively find associations among these scales.

Because the development of mental illness can be related to a variety of psychological factors, we demonstrate the usability of PHARM by discovering association rules to predict mental illness based on scale scores of five psychological factors, including family functioning, social support, depressive symptoms, perceived empathic self-efficacy, and anxiety disorder. A summary score of each scale was used, and the mean and standard deviation (STD) of scores from 2,637 reports were calculated. Instead of using recommended cut-points provided by scales, we statistically defined disorder ranges. Cut-points for disorder ranges were set to be the mean + STD if high scores imply disorders, otherwise, mean - STD. Table 2 provides a complete list of targeted psychological factors, scales, and disorder ranges. We set $Supp_m$in= 1.5% since mental disorders are relatively rare in the healthy population, especially for those who have compounded disorders. Such threshold for *Supp* implies that each rule was mined from at least 40 records (out of 2,637 records). We set $Conf_{\min}$ = 80% as suggested by domain experts, to ensure the confidence of each rule. In Table 3, we list the final 12 rules that can predict potential mental illness that was measured by general mental health group in SF-36.

Our results provide important knowledge to (1) *prevent* the development of mental illness and (2) *prioritize* advice and action plans to reduce the risk of existing mental illness. According to Table 3, in general, rules with more antecedent items have lower support values because they represent more specific cases. On the other hand, specific cases tend to have higher confidence values. That is the reason why there is no confident rule with 1-item antecedent because people with only one psychological disorder are typically at low risk of mental illness compared to those with compounded disorders. However, we should pay more attention to *prevent* the development of compounded disorders even when a person is currently having only one psychological disorder. For example, if a person is currently having depressive symptoms (BDI), we may want to provide proactive advice to prevent the development of disorders especially in perceived empathic self-efficacy (PSSE, rule #10)

and family functioning (FAD, rule #12) because they are associated with mental illness risk if comorbid with BDI. On the other hand, we can also use these rules to *prioritize* the action plans to reduce the risk of existing mental illness.

For instance, according to the rule #1, individuals who have compounded disorders in social support (ESSI), depression (BDI), perceived empathic self-efficiency (PSSE), and anxiety (GAD7), are associated with the highest possibility (99.8%) of mental illness. Among these four factors, we should first focus and set action plans for disorders of social support (ESSI) because it can significantly (p=0.02 using $\%^2$-test) drop the risk from 99.8% to 82.1% by comparing rule #1 to rule #11.

iv. CONCLUSION

In this study, we researched and developed an interactive decision support system, called PHARM, to generate quantitative and objective rules in predictive health settings. By leveraging a predictive health dataset from 696 subjects, we adopted association rule mining to discover personalized health prediction rules. We utilized the system to investigate association rules to predict mental illness based on five psychological factors. Our results provide important knowledge to prevent the development of mental illness and prioritize advice and action plans to reduce the risk of worsening existing mental illness. Our future work will first include the temporal component of the CHDWB dataset and use association rule mining to provide temporal prediction rules. By doing so, we can more precisely predict and prevent chronic health conditions (or even diseases) within a specific timespan. Second, we will perform extensive validation (e.g., significance and actionable) on mined rules by combining ARM with other data mining techniques (e.g., classification).

ACKNOWLEDGMENT

The authors are gratefi.il to Dr. John Phan, Lynn Cunningham, Janani Venugopalan, and Chanchala Kaddi for their valuable assistance.

References

1. WHO Department of Health Statistics and Informatics. Aug 08. 2013 Available from: http://www.who.int/gho/publications/world_health_statistics/2012/en/index.html
2. Ginsburg, PB. High and rising health care costs: Demystifying US health care spending. Princeton, NJ: 2008.
3. Thorpe KE, Howard DH. The rise in spending among Medicare beneficiaries: the role of chronic disease prevalence and changes in treatment intensity. Health Affairs. 2006; 25(5):w378–w388. [PubMed: 16926180]
4. Johns MM, Brigham KL. Transforming health care through prospective medicine: The first step. Academic Medicine. 2008; 83(8):706. [PubMed: 18667878]
5. Yu J, et al. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. Bioinformatics. 2005; 21(10):2200–2209. [PubMed: 15784749]
6. Hipp J, Giintzer U, Nakhaeizadeh G. Algorithms for association rule mining—a general survey and comparison. ACM SIGKDD Explorations Newsletter. 2000; 2(1):58–64.
7. Konias, S., et al. Computers in Cardiology, 2003. IEEE; 2003. Uncertainty rule generation on a home care database of heart failure patients.
8. Shan, Y., et al. Mining medical specialist billing patterns for health service management. Conferences in Research and Practice in Information Technology; 2008.

9. Chaves R, et al. Efficient mining of association rules for the early diagnosis of Alzheimer's disease. Physics in medicine and biology. 2011; 56(18):6047. [PubMed: 21873769]
10. Laxminarayan P, et al. Mining statistically significant associations for exploratory analysis of human sleep data. Information Technology in Biomedicine, IEEE Transactions on. 2006; 10(3): 440–450.
11. Agrawal, R.; Srikant, R. Fast algorithms for mining association rules. Proc. 20th Int. Conf Very Large Data Bases, VLDB. 1994.
12. Streiner, DL.; Norman, GR. Health measurement scales: a practical guide to their development and use. Oxford university press; 2008.
13. Sidebottom AC, et al. Validation of the Patient Health Questionnaire (PHQ)-9 for prenatal depression screening. Archives of Women's Mental Health. 2012; 15:367–374.
14. Fazel M, et al. Mental health of displaced and refugee children resettled in high-income countries: risk and protective factors. The Lancet. 2012; 379(9812):266–282.
15. Bech P, et al. Measuring well-being rather than the absence of distress symptoms: a comparison of SF-36 Mental Health subscales and the WHO-Five well-being scale. International journal of methods in psychiatric research. 2003; 12(2):85–91. [PubMed: 12830302]

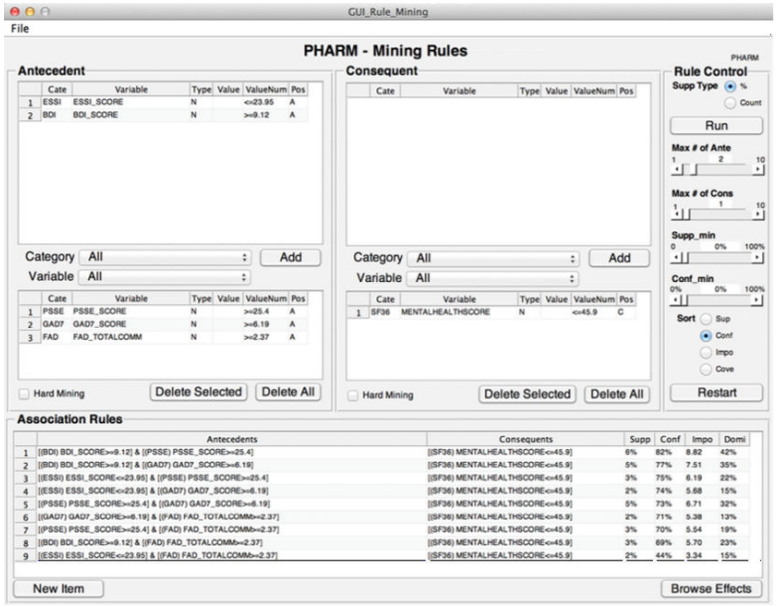

Fig. 1.
The Rule Mining Window.

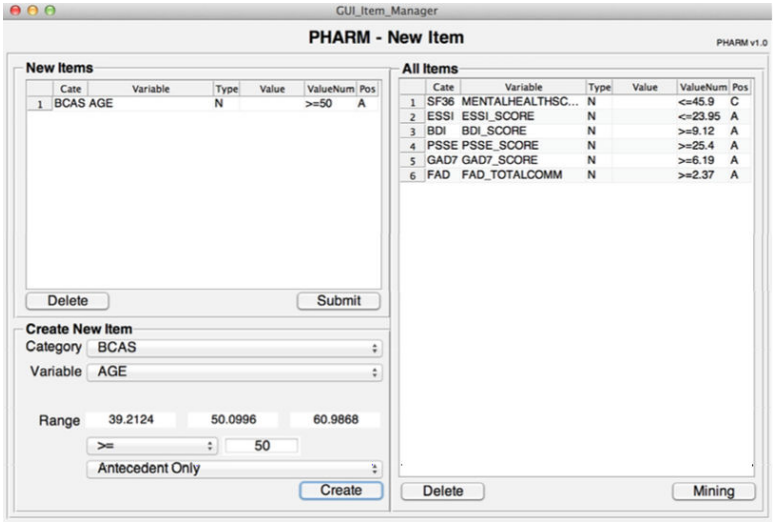

Fig. 2.
The New Item Window.

Table 1

Examples of CHDWB measurement.

Measurement Type	Examples
Questionnaires	Demographics, personal and family health history, occupational history and exposures, tobacco and alcohol usage.
Assessments	Perceived stress scale, block food frequency questionnaire, CAPS physical activity questionnaire, SF-36 (v2).
Physical measurements	Resting blood pressure, resting heart rate, bioelectrical impedance analysis, dual-energy X-ray absorptiometry scan.
Lab tests	Lipid panel, blood chemistries, urine creatinine and microalbumin, vitamin B12, iron and total iron binding capacity.
Research lab values	Oxidized and reduced glutathione, cysteine, cystine, CysGSH, CysRedox, serum protein nitrotyrosine.

Table 2

Assessment Scales and Disorder Range.

Scale	Disorder Range
SF36M: Short Form (36) Survey Mental Health	< 45.91
FAD: Family Assessment Device	> 2.37
ESSI: ENRICHD Social Support Inventory	< 23.95
BDI: Beck Depression Inventory	> 9.12
PSSE: Perceived Empathic Self-Efficacy Scale	> 25.41
GAD7: Generalized Anxiety Disorder 7-item	> 6.19

Table 3

Rules predicting general mental problem (SF36M < 45.91).

Rule #	Antecedent	*Supp*	*Conf*
1	ESSI + BDI + PSSE + GAD7	1.6%	99.8%
2	ESSI + BDI + PSSE	3.1%	93.4%
3	BDI + PSSE + GAD7 + FAD	1.6%	93.8%
4	ESSI + PSSE + GAD7	2.6%	92.6%
5	ESSI + BDI + PSSE +FAD	1.6%	91.3%
6	ESSI + BDI + GAD7	2.7%	89.5%
7	BDI + PSSE + FAD	2.5%	88.1%
8	ESSI + BDI + FAD	2.2%	86.6%
9	BDI + GAD7 + FAD	4.1%	84.4%
10	BDI + PSSE	6.8%	82.0%
11	BDI + PSSE + GAD7	2.8%	82.1%
12	BDI + FAD	5.5%	81.9%
	4-factor	1.6%	95.0%
Average	3-factor	2.9%	88.1%
	2-factor	6.2%	82.0%

All antecedent factors are within disorder range.

IFMBE Proc. Author manuscript; available in PMC 2016 August 30.