Risk-O-Meter: An Intelligent Clinical Risk Calculator

Kiyana Zolfaghar
Inst. of Technology, Ctr. for Web and Data Science (CWDS), Univ. of Washington Tacoma
kiyana@uw.edu

Jayshree Agarwal
Inst. of Technology, CWDS, Univ. of Washington Tacoma
jagarwal@uw.edu

Deepthi Sistla
Inst. of Technology, CWDS, Univ. of Washington Tacoma
dsistla@uw.edu

Si-Chi Chin
Inst. of Technology, CWDS, Univ. of Washington Tacoma
scchin@uw.edu

Senjuti Basu Roy
Inst. of Technology, CWDS, Univ. of Washington Tacoma
senjutib@uw.edu

Nele Verbiest
Ghent University
nele.verbiest@ugent.be

ABSTRACT

We present a system called *Risk-O-Meter* to predict and analyze clinical risk via data imputation, visualization, predictive modeling, and association rule exploration. Clinical risk calculators provide information about a person's chance of having a disease or encountering a clinical event. Such tools could be highly useful to educate patients to understand and monitor their health conditions. Unlike existing risk calculators that are primarily designed for domain experts, Risk-O-Meter is useful to patients who are unfamiliar with medical terminologies, or providers who have limited information about a patient. Risk-O-Meter is designed in a way such that it is flexible enough to accept limited or incomplete data inputs, and still manages to predict the clinical risk efficiently and effectively. Current version of Risk-O-Meter evaluates 30-day risk of hospital readmission. However, the proposed system framework is applicable to general clinical risk predictions. In this demonstration paper, we describe different components of Risk-O-Meter and the intelligent algorithms associated with each of these components to evaluate risk of readmission using incomplete patient data inputs.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based Services*

Keywords

Clinical risk calculator, risk of hospital readmission predication, clustering, association rule mining

1. INTRODUCTION

Clinical risk calculators and risk assessment tools are useful services to educate patients as well as healthcare providers to monitor the development of health conditions. Risk calculators are commonly used for diseases like cancer, diabetes, heart disease, and stroke etc. However, most clinical risk calculators require adequate knowledge of medical terminologies which are unfamiliar to regular patients. In addition, a sophisticated and effective predictive model often requires a large set of attribute values that may not all be available (or known) at the time when a patient or a healthcare provider uses the risk calculator. One recent popular risk calculator that predicts the risk-of-readmission related to heart attack, was proposed by the researchers at Yale University[1]. Unfortunately, the usefulness of this tool is largely limited only to domain experts such as physicians, who posses adequate expertise on the domain and are aware of the values of all the attributes[2].

In light of this opportunity, we develop *Risk-O-Meter*, an intelligent and effective system that allows predicting clinical risks with only limited user inputs. Novelty of Risk-O-Meter is manifold - a) Risk-O-Meter is worthwhile even to patients and healthcare providers, who are not necessarily domain experts. While the actual prediction algorithm is developed with the analyses of complex factors, unlike existing risk calculators, it allows users to provide very simple inputs. b) Unlike existing calculators, it accepts incomplete patient input, thus offers higher flexibility. Moreover, it offers intelligent visualization of other pertinent factors of the *similar* patients, based on the input values. c) Finally, along with the risk calculation, Risk-O-Meter also suggests meaningful explanation behind such prediction. We note that the proposed framework inside Risk-O-Meter is generic and could be applicable to a wide variety of clinical risk prediction tasks. However, for the demonstration purpose, we instantiate the system to predict the risk-of-readmission for patients within 30-days of discharge.

Hospital readmission refers to patient admission to a hospital after being discharged from an earlier hospital stay. 30-day readmission rate is considered as an indicator for evaluating the quality of care. In the U.S., starting from 2012, the Centers for Medicare & Medicaid Services (CMS) began to use preventable readmission rates as a quality metric to determine the reimbursement to hospitals [4]. Predict-

[1] http://www.readmissionscore.org/

[2] In order to predict the risk-of-readmission, this calculator requires as many as 22 different attribute values.

KDD '13 Chicago, IL USA

ing hospital readmission risk helps identify which patients would benefit most from care transition interventions, such as arranging a visiting nurse for the patient after the discharge. The topic has received a great deal of attention recently among healthcare professionals and researchers [3, 5, 6]. An effective risk calculator to assess the risk of 30-day hospital readmission risk can largely benefit patients and providers. This demo focuses on two tasks: predict and explain high risk of 30-day hospital readmission, and summarize and visualize the leading contributing factors. The proposed Risk-O-Meter has the following contributions:

- We address the challenge of incomplete input data to calculate clinical risks and devise novel solutions to that end (using clustering);
- We design effective supervised learning techniques (classification algorithms) to predict 30-day hospital readmission risk.
- To the best of our knowledge, Risk-O-Meter is the first ever tool that proposes explanation behind risk prediction and offers intelligent visualization of pertinent characteristics of the data, based on input values.
- To enable near real-time risk prediction, Risk-O-Meter leverages significant offline computation for both clustering and classification process. Additionally, given a set of input values, it uses effective indexing techniques to enable faster computation of risk prediction.

In the rest of this paper, we first describe the technical specifications of Risk-O-Meter and then demonstrate application scenario with the system.

2. TECHNICAL SPECIFICATION

2.1 System Overview

Risk-O-Meter is a web application to predict and analyze the risk of 30-day hospital readmission for a patient. The majority of the system components are precomputed and stored to increase the speed of the application. Figure 1 presents the system overview. Since the users can input any subset of the n input attributes presented in the interface, the first task is to map the input values to a group of patients who are most similar to the provided user profile. Therefore a set of clusters are precomputed as described in Section 2.2. The actual predictive models is built considering a larger attribute set, beyond the small set of attribute values specified by the user. Therefore, the inputted attribute set needs to be transformed to a larger set of attribute values on which the predictive model is trained. For that purpose, we maintain a cluster *centroid* which is used to *complete* the remaining missing attribute values. Additionally, we also visualize the characteristics of similar patients using top N pertinent attributes. Finally, we also offer an explanation behind the risk of readmission prediction using association rule mining. Given a set of input attribute values, we perform efficient matching to map the inputs to its closest cluster centroid. The centroid computation, visualization, predictive model training, and association rule mining are performed offline leveraging the properties of respective clusters, while the centroid matching is the only online process that takes place at runtime. Section 2.2 describes in details the offline system components and Section 2.3 presents the online cluster mapping algorithm.

Risk-O-Meter is primarily implemented using RStudio Shiny package[3] and Javascript. We use R project[4] to implement statistical analysis, clustering, predictive modeling, and association rules mining. The choice of the implementation allows us to combine extensible data mining algorithms, statistical analysis tools, and compelling data visualizations into a single web application.

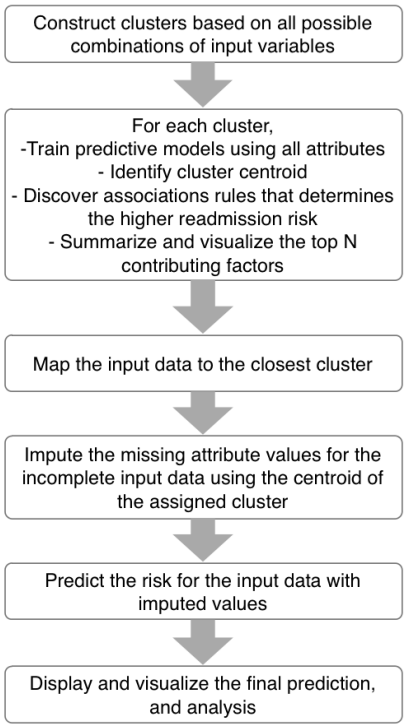

Figure 1: System overview

2.2 Offline Computation

In this section, we present the details of the five offline system components: clustering, building predictive models, visualizing cluster characteristics, coping with incomplete input, and explaining predictions.

`Clustering.` To allow simple and incomplete patient input in Risk-O-Meter, we precompute the clusters based on different permutations of input attributes using K-mode algorithm. To accommodate all possible scenarios, we constructed $k * 2^n$ clusters, where n is the number of attributes used in the predictive model and k is the predetermined number of clusters for each combination of user input (there is a total of 2^n permutations of attributes). For example, if a user inputs the age, gender, and blood pressure, he or she will be mapped to one of the k clusters built based on the three attributes. In this demo, we used 49 ($n = 49$) attributes to train predictive models and set $k = 5$ based on the empirical experiments. Future work may explore methods to automatically identify the optimal n and k parameters. To allow efficient and speedy cluster mapping, we constructed an inverted index on the clusters as described in Section 2.3. Clustering is a necessary component for both Step 2 and 3 of the system as shown in Figure 2.

[3]http://rstudio.github.com/shiny/
[4]http://www.r-project.org/

`Building Predictive Models.` We implemented Naive Bayes classifiers using the complete set of attributes ($n = 49$) for each precomputed clusters. Naive Bayes classifier utilizes a probabilistic method for classification by multiplying the individual probabilities of every attribute-class pair. The method has been shown as one of the effective classifiers in [5]. The outcome of predictive modeling is shown in Step 2 (Figure 2).

`Visualizing Cluster Characteristics.` We summarized the characteristics of a cluster via the analysis and visualization of significant contributors to high risk of readmission. Chi-Square test was used to discover the top N correlated attributes ranked by the P-value (observed significance level). This component implements Step 3 of the system (Figure 2).

`Coping with Incomplete Input.` Coping with incomplete input data is similar to the problem of imputing missing values, a common problem in all types of medical research. In this demo, we use the precomputed cluster centroid to replace the missing values of the input data for the prediction. This component is part of the implementation for Step 2 (Figure 2).

`Explaining Predictions.` We used Apriori association rules mining to identify patterns that may associated with high risk of readmission. For each cluster, we extracted rules of high support and confidence scores. We then summarized the discovered patterns for each cluster in a paragraph as part of the cluster feature. This component is also part of the implementation for Step 2 (Figure 2).

2.3 Online Cluster Mapping and Risk Prediction

One novelty of Risk-O-Meter is its flexibility in accepting inputs from the user. In the user interface, the user enters all or only a subset of the attribute values that she is aware of. After that, the task is to map the inputted parameters of the patient to a cluster of patients in our database that she *resembles* the most. More specifically, the process thus becomes computing the *similarity (or distance)* [2] between the cluster-centroid and the inputted values and selecting that centroid that has the highest similarity (smallest distance) with the inputs. While we are aware of several measures to compute similarity or distance, our current implementation considers Cosine Similarity [2] for that purpose. However, since the total number of materialized cluster centroids is exponential (actually, $k \times 2^m$, where k-clusters are materialized for each possible subset of m attributes in the predictive modeling) to the number of attributes (m) in the input user interface, the challenge is to be able to perform this similarity computation efficiently at run-time (after the user has specified the inputs). We propose an effective indexing scheme to that end that bears resemblance with inverted-index [1] in IR literature. The advantage of using this indexing scheme is it avoids the exponential number of similarity computation (given the input parameters) at run-time, and only performs k searches to find out the centroid with the highest Cosine Similarity.

Intuitively, the indexing scheme is designed as follows: For every possible subset of attributes, a list of k centroids are pre-computed and materialized. The inverted index file contains 2^m rows (each corresponds to a subset of attributes), and each row contains a list of k cluster centroids (considering distribution of values of that attribute subset). Given input values on a subset of attributes, the task is to first perform an efficient look-up into the index file to retrieve the row that corresponds to that attribute subset, and then perform k similarity computations (one for each centroid) to find the centroid with highest similarity. Assuming that one can perform constant time look-up to retrieve the row in the index-file given the inputted attributed set, the complexity of finding the best clustered centroid becomes at $O(k)$ in our case.

By leveraging the cluster centroid along with the inputted attribute values, the trained predictive model outputs the clinical risk of the input patient profile.

3. SYSTEM DEMONSTRATION

In this section, we use a cohort provided by MultiCare Health System[5] to demonstrate Risk-O-Meter. The cohort consists approximately 8,600 patients diagnosed with heart failure.Figure 2 shows a screenshot of Risk-O-Meter. Our demonstration contains three main steps: 1) flexible user input; 2) prediction and explanation; and 3) real-time data exploration and visualization.

As shown in Figure 2, Step 1 involves the web form on the left displaying first the intuitive input fields. Although, if possible, a user can provide complete information for all the required attributes (49 attributes in the predictive modeling), the system can accommodates limited partial inputs. For example, a user can provide only the age(=71), gender(=M), ethnic group(=African American), the history of diabetes(=yes), and the blood pressure(<130/80). When the "Calculate Risk" button is clicked, Risk-O-Meter maps the user profile to the closest cluster and display the prediction in Step 2.

Step 2 outputs the probability of prediction obtained from Naive Bayes classifier. In addition, it presents the explanation generated from association rule mining process. In the screenshot, the provided input values, combined with the additional attribute values imputed from the centroid of the assigned cluster, lead to 72% probability of 30-day readmission. Association rule mining offers an explanation behind risk calculation, for example, for the attached screenshot, it explains that males in the age group of 70s with diabetes have high risk of readmission. Users of Risk-O-Meter may decide whether they want to learn more about the cluster by clicking the "Explore Data" button.

Step 3 explores and visualizes the characteristics of the assigned cluster for the given input values. It shows that the top three contributor to readmission are history of depression, discharge severity, and history of pneumonia. The stacked bar charts shows the distributions of people readmitted to hospital within 30 days for each factor. For example, we observe that patients similar to the given profile tends to have level 3 discharge severity, and several patients in our dataset with discharge severity 3 tends to exhibit high risk or 30-day readmission.

4. CONCLUSIONS

In this paper, we propose Risk-O-Meter, a system aims to predict and analyze clinical risk via data imputation, visualization, predictive modeling, and association rule exploration. Risk-O-Meter is designed for patients and healthcare providers, who may only have a limited or incomplete infor-

[5]http://www.multicare.org/

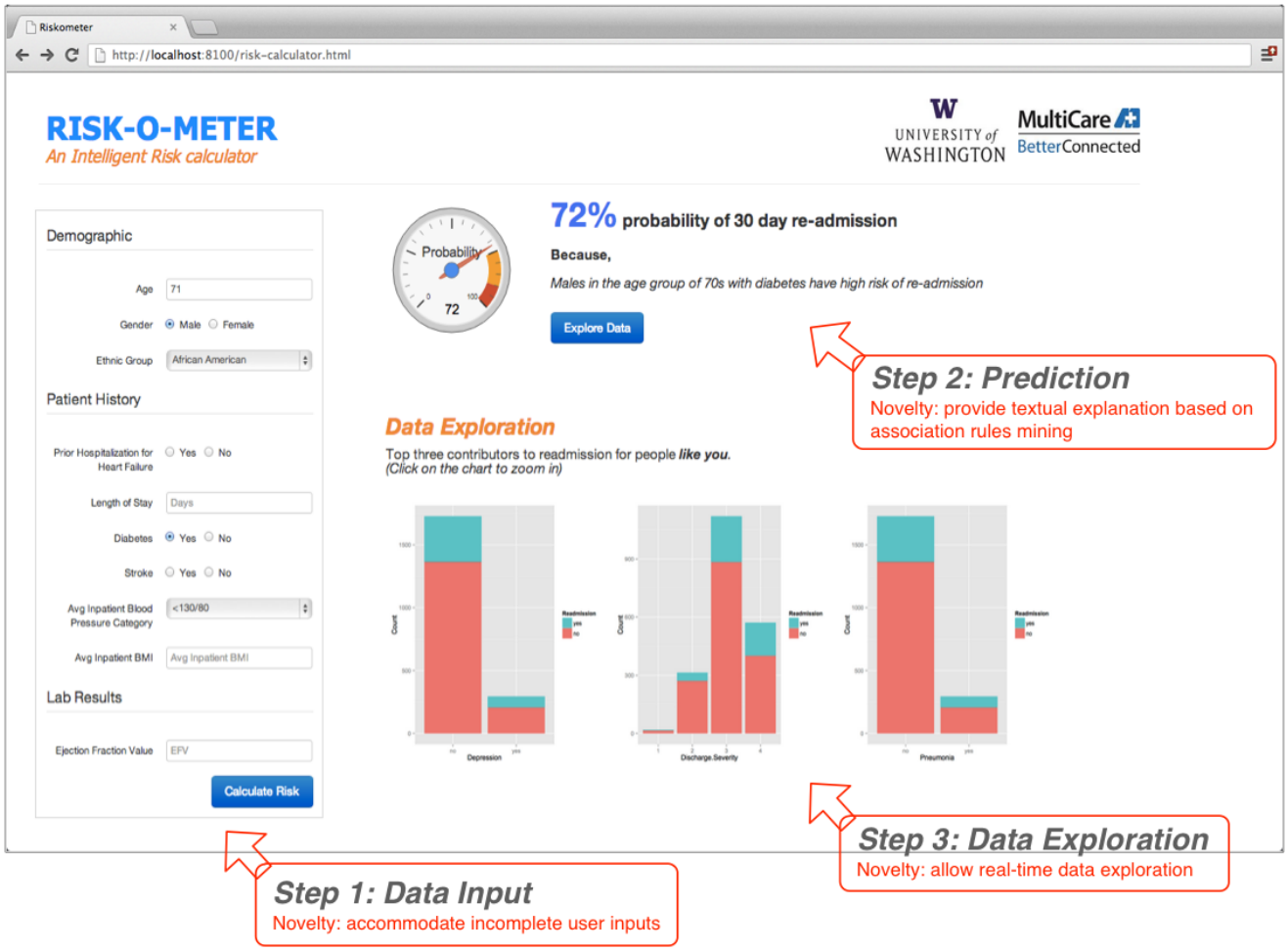

Figure 2: Risk-O-Meter to Calculate Risk-of-Readmission for Congestive Heart Failure Patients

mation at the time they use a risk calculator. One novelty of Risk-O-Meter, compared to currently available clinical risk calculators, is the capability to efficiently and effectively extrapolate the limited input values to a meaningful group of patient data to proceed the prediction and analyses. Moreover, Risk-O-Meter provides explanation of risk predictions and intelligent visualization of other patients similar to the given patient profile, which, to the best of our knowledge, is the first ever risk calculation tool with the improved functionalities.

5. ADDITIONAL AUTHORS

Additional authors: Ankur Teredesai (Inst. of Technology, CWDS, Univ. of Washington Tacoma, email: `ankurt@uw.edu`), David Hazel (Inst. of Technology, CWDS, Univ. of Washington Tacoma, email: `dhazel@uw.edu`), Paul Amoroso (Multicare Healthcare Systems, email: `Paul.Amoroso@multicare.org`), and Lester Reed (Multicare Healthcare Systems, email: `Lester.reed@multicare.org`).

6. REFERENCES

[1] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England, 2011.

[2] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[3] E. H. Kansagara D. Risk prediction models for hospital readmission: A systematic review. *JAMA*, 306(15):1688–1698, Oct. 2011.

[4] H. M. Krumholz, A. R. Merrill, E. M. Schone, G. C. Schreiner, J. Chen, E. H. Bradley, Y. Wang, Y. Wang, Z. Lin, B. M. Straube, M. T. Rapp, S.-L. T. Normand, and E. E. Drye. Patterns of hospital performance in acute myocardial infarction and heart failure 30-day mortality and readmission. *Circulation: Cardiovascular Quality and Outcomes*, 2(5):407–413, Sept. 2009.

[5] N. Meadem, K. Zolfaghar, J. Agarwal, S.-C. Chin, S. Basu Roy, A. Teredesai, D. Hazel, P. Amoroso, and L. Reed. Prediction risk of readmission for congestive heart failure patients. (under review), 2013.

[6] K. Zolfaghar, N. Verbiest, J. Agarwal, N. Meadem, S.-C. Chin, S. B. Roy, A. Teredesai, D. Hazel, P. Amoroso, and L. Reed. Predicting risk-of-readmission for congestive heart failure patients: A multi-layer approach, 2013. (under review).