Information Technology and Quantitative Management (ITQM2013)

Examining Disease Risk Factors by Mining Publicly Available Information

Jay Pedersen¹, Fangyao Liu¹, Fahad Alfarraj¹, Harry Ngondo¹

¹College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA
email: {jay.pedersen, fangyao.liu, fahad.alfarraj, hajingondo} @ unomaha.edu

Abstract

Human disease and associated risk factors are of great interest in the medical field. The skyrocketing cost of health care makes the understanding of disease risk factors of even greater importance. When risk factors are well understood for a disease it is possible to educate the public to reduce their risk by avoiding risk factors that they can control. There are publicly available data stores which document risk factors facing the general public in the United States. In particular, the Behavioral Risk Factor Surveillance System (BRFSS) has been maintained by the Center for Disease Control (CDC). This study examines the disease of diabetes using a set of risk factors which the BRFSS maintains. A logistic regression model is created which models diabetes as a function of risk factors. Nine years of existing data from BRFSS was used in building this model (for the years 2002 to 2010). The generated model shows promise in modeling diabetes as a function of risk factors and correctly identified obesity as an important risk factor for diabetes. Studies of this nature can be very cost effective and may generate insights into the risks of disease.

© 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-ND license.
Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management

Keywords: diabetes; data mining; risk factors;

1. Introduction

Diabetes is defined by the U.S. National Library of Medicine as a “lifelong disease in which there are high levels of sugar in the blood” [1]. The human body has the ability to produce insulin to control the amount of sugar in the blood stream. When the body is unable to produce insulin or effectively use insulin, the sugar concentration in the blood stream begins to build up. This condition is known as diabetes. In Type 1 diabetes, the body makes no insulin. In Type 2 diabetes, glands resist insulin and sugar in the blood is not effectively moved to the glands and broken down.

The Center for Disease Control (CDC) maintains a publically-accessible data store of various disease risk factors known as the BRFSS -- Behavior Risk Factor Surveillance System [2]. BRFSS provides a publicly accessible data store with the results of the annual risk factor survey conducted by the CDC. The annual survey has been conducted since 1984. The surveys are conducted by phone interview which queries the individual about their current medical condition and their behavior-related risk factors. Each surveyed individual is asked the same set of questions and the results are stored in a coded format. The exact questions asked each year and the coding of the answers is available online from the BRFSS. Personally identifiable information is not collected or stored. What is stored is the state the person resides in, their gender and ethnicity and the coded answers to various questions about health risk factors.
In this study we chose to study the data from 2002 to 2010 and to look at the following behavioral risk factors: obesity, overweight, age, smoking, alcohol use, blood pressure, gender, and ethnicity which were available for. We built a logistic regression model using all of those factors to predict the outcome of diabetes. At least one other study looked at diabetes and risk factors from the BRFSS [3], but did not try to build a factor model. The study examined data from 2001 and the current trend was of interest to our study.

A logistic regression model is a statistical model which requires a Boolean variable of interest as its dependent variable [4]. In this case, a variable which is YES when a person has diabetes and NO when they do not is the variable of interest. The model can predict a YES or NO result for the variable of interest based on other

2. Data Collection

The BRFSS data store consists of very large flat text files containing fixed-length records, one file for each year since 1984[5,6]. There is one record for each survey which was conducted. The fixed-length record contains coded values which describe the result of the survey. The format of the records changes from year to year. The CDC provides a “codebook” for each year which describes the encoding of each answer into the fixed-length records [7,8]. For example, in 2003, the variable “_BMI4CAT” is stored in position 858 in the fixed-length-record and describes the survey result for BMI (Body Mass Index). The value of 1 indicates a normal BMI, 2 indicates “Overweight” BMI, 3 indicates “Obese” BMI, and 9 indicates “Unknown/did not answer/Missing”.

It is therefore necessary to understand positions where the variables are stored, and what their coded values mean. The positions change from year to year and must be understood to properly process the data from different years. The necessary information is given in the form of the codebooks.

Each record in every year has a variable called “DISPCODE” which describes whether the survey was fully completed. A value of 110 indicates a fully completed interview. The data collection procedure discarded any record which was not fully completed.

Many of the variables were of a two-valued nature, where the result was YES or NO, with a third option allowed where the person did not provide an answer. These are considered three-valued for the purpose of this study.

The following risk factor variables were extracted from the BRFSS data files for the years from 2002 to 2010:

- DISPOSITION – indicates whether the survey was complete; only completed surveys were used.
- STATE – indicates the state in the US where the subject resides.
- YEAR – 2002 or 2003 or … or 2010
- DIABETES – indicates if the subject was ever told by a doctor that they had diabetes
- AGE – age of subject (values 18 to 99)
- GENDER – male or female
- ETHNICITY – ethnicity of subject
- ALCOHOL – indicates whether the subject consumed any alcoholic drinks in the last 30 days
- SMOKING – indicates whether the subject has smoked at least 100 cigarettes in their lifetime
- OVERWEIGHT – weight-related, indication that subject has BMI (body mass index) of at least 25
- OBESE – weight-related, indicates whether the subject has BMI of at least 30.
- BLOODPRESSURE – indicates whether the subject was told that they have high blood pressure

DIABETES is maintained as a two-valued variable, so that it can be used as the dependent variable in a logistic regression. Surveys which did not give a YES or NO answer for DIABETES were not collected due to this

2.1. Demographics of survey responders

Some demographic information about the persons who responded to the survey (from 2002 to 2010) was available within the BRFSS data. The following demographic information was obtained showing that more females were interviewed than males and that the median age of the subjects was about 54.
2.2. Trends found in Risk Factors from 2002 to 2010

The following figures shows the trends that were observed in individual risk factors as year increases from 2002 to 2010. In the graphs, year 1 is 2002 and year 9 is 2010. Note the general similarity in diabetes and obesity trends.
3. Creation of Logistic regression model

After data collection and cleaning, we have two data sets. One set includes data for all states except Nebraska and is used to build a logistic regression model. Another set includes data for the state of Nebraska which will be used to test the model.

The DIABETES variable is two-valued and is the dependent variable in the logistic regression model.

The following three-valued risk factor variables are used as independent variables (1=true, 2=false, 3=unknown): ALCOHOL, BLOODPRESSURE, GENDER, OBESE, OVERWEIGHT, SMOKING.

The following numeric variables are used as independent variables in the model: YEAR, AGE, ETHNICITY.

Logistic regression measures the relationship between a categorical dependent variable and usually a continuous independent variable (or several), by converting the dependent variable to probability scores:

\[
P_{ij} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}},
\]
The following is the resulting logistic regression model:

![Logistic Regression Model](image)

Fig 7. Logistic Regression model for Diabetes based on Risk Factors

### 4. Testing the model against known data

The generated logistic regression model was tested against the Nebraska data set, to see how well the model would predict cases of diabetes. There were 87,492 records in the Nebraska data set.

The following confusion matrix shows the result of the predictions of diabetes for those 87,492 records compared to the actual (known) diabetes value for those records.

<table>
<thead>
<tr>
<th>Table 1 – Confusion Matrix for Test of Model versus Nebraska data</th>
</tr>
</thead>
</table>
| Model-diabetes=
| Predicted-Value |
| Actual-Value |
| Yes (0) | No (1) | Total |
| Yes (0) | 12 | | 8805 | 8817 |
| No (1) | 23 | | 7439 | 7452 |
| Total | 35 | | 9545 | 9559 |

This result is preliminary and clearly has a lot of type 1 and type 2 errors.

### 5. Conclusion

The study presented a cost-effective way of examining risk factors for a disease which has publically available information about the disease and the behaviors of people who do and do not have the disease.

The model indicated that the most important risk factor, according to variable coefficient in the model – is the obesity variable. This matches known information that obesity is a risk factor of diabetes [9].

It may be possible to examine different risk factors within the BRFSS system related to Diabetes and create a better model than was created in this study. It may be possible to create the model with the same basic variables but using a different coding scheme and improve the result as well. The coding of 1=true, 2=false and 3=unknown may not be optimal.

Since behavioral risk factor data is publically available; it is cost-effective to apply multiple data mining methods, such as decision trees, neural networks and optimization to train the data for the hidden patterns and obtain a better classifier for of diabetes [10, 11, 12].
Diseases other than diabetes may be able to be similarly modeled wherever risk factor information is available in BRFSS or other publically available data stores.

References

[9]. Leong KS, Obesity and Diabetes, Best Practice & Research Clinical Endocrinology & Metabolism, Volume 13, Issue 2, July 1999, Pages 221-237
[12]. He, J., Y. Zhang, Y. Shi and G. Huang, Domain-Driven Classification Based on Multiple Criteria and Multiple Constraint-Level Programming for Intelligent Credit Scoring, IEEE Transactions on Knowledge & Data Engineering, Vol. 22 (2010) 826-838.