Proceedings of the 3rd INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2008)
J. Li, D. Aleman, R. Sikora, eds.

AN INFORMATION VISUALIZATION APPROACH TO CLASSIFICATION AND ASSESSMENT OF DIABETES RISK IN PRIMARY CARE

Christopher A. Harle
Daniel B. Neill
Rema Padman

The H. John Heinz III School of Public Policy and Management
Carnegie Mellon University
Pittsburgh, PA
charle@cmu.edu

Abstract

Chronic disease risk assessment is a common information processing task performed by primary care physicians with many at-risk patients. However, effectively integrating information about many risk factors across many patients is cognitively difficult. Methods for visualizing multidimensional data may augment clinical disease risk assessment by providing reduced-dimensional displays which stratify patient data points according to risk level while providing additional insight into clinically important individual risk factor variables. This study combines medical evidence, dimensionality reduction techniques and information visualization to develop a new framework for visually classifying and interpreting patient data. This framework is then explored and analytically validated using a unique health information database from the American Diabetes Association that contains risk predictions made by the Archimedes model. Results show that the framework may generate models which visually classify a large patient population with accuracy comparable to common statistical methods. Further, the visualizations provide rich displays that give insight into (i) the relative importance of individual risk factors, (ii) confidence in individual patient risk predictions and (iii) overall distributions of risk in a population. The proposed approach may produce models that can be embedded in health information systems to provide interactive visual analysis tools that support physician decision making.

Keywords: data visualization, dimensionality reduction, risk assessment, diabetes

Motivation

Diabetes mellitus is a costly chronic disease. An estimated 13.0 million Americans were diagnosed as of 2002 and an additional 5.2 million were believed to be unaware of their condition [1]. Much of the burden of preventing, diagnosing and managing diabetes falls on the primary care physician who often has insufficient resources to effectively prevent and manage this complex disease [2]. At the individual patient level, clinicians must be able to quickly assess multiple laboratory tests, history and other risk factors to judge risk of disease. At the patient population level, monitoring and responding to changes in risk are important due to the rise of pay-for-performance initiatives [3]. Given the complexity of chronic disease prevention, diabetes risk assessment may be a critical area for improving clinical decision support.

Information visualization utilizes the high bandwidth processing capabilities of the human visual system to reveal patterns in data that are not evident in non-visual data analysis [4, 5]. Some visualization methods rely on graphical techniques for interacting with data by rotating, zooming and subsetting. Machine learning methods reduce multidimensional data to low dimensions while minimizing some measure of information loss [6]. Shneiderman proposes combining methods from each of these areas to enable more effective and responsible analysis. This study reports preliminary results from a case study that applies Shneiderman's proposal by combining automated algorithms and transparent data visualization to support diabetes risk assessment.

Research Objectives

The primary objective was to find low-dimensional mappings of multidimensional patient data that classified patients according to their risk of type II diabetes onset. The second goal was for these classifiers to be practically interpretable by a physician interested in visual analysis of a patient population. To preserve interpretability, models were limited to two-dimensional scatter plots with axes formed by linear combinations of diabetes risk factor variables.

The following describes the preliminary steps in developing and evaluating the model:

(1) *Variable selection* based on medical evidence and existing models of diabetes risk [7, 8]
(2) *Dimensionality reduction* of the chosen variables using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [6]
(3) *Enhanced visual interpretability* through an attraction metaphor similar to VIBE [9]
(4) *Training of classifiers and testing of classification accuracy* using data submitted to and processed by the American Diabetes Association (ADA) Diabetes PHD website [7]

Model Development and Evaluation

Variable Selection

To select factors that are most relevant to diabetes risk assessment, we relied on the medical literature on Type II diabetes risk, prior models of diabetes risk and discussion with an internal medicine physician. Table 1 lists the ten factors that were chosen. While each of these are not necessarily causally related to diabetes, there is evidence that they are associated with incidence of the disease, and they have been used in past, non-visual risk models. Variable selection was also limited to data that is commonly available in primary care physician practice records.

Table 1: Diabetes risk factors.

Systolic Blood Pressure	continuous
Diastolic Blood Pressure	continuous
LDL	continuous
HDL	continuous
Diabetes Family History	binary (yes/no)
Smoker	binary (yes/no)
Regular Check Up	binary (yes/no)
Physical Activity Level	ordinal (vigorous, sedentary, moderate, light)
Age	continuous
BMI	continuous

Dimensionality Reduction

The machine learning literature contains many methods for projecting multidimensional data to a reduced space. Representing data in low dimensions is often useful for simplifying subsequent analysis, including the ability to visualize data. This study explored Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [6] for finding two-dimensional representations of clinical data. While PCA finds linear combinations of variables that maximize variance, LDA is explicitly concerned with classification and finds projections that maximize the ratio of between-class variance to within-class variance. Two two-dimensional projections were evaluated for classification accuracy and clinical interpretability. The first model scaled the health data to mean zero and unit standard deviation and then mapped the original ten variables (Table 1) to the first and second principal components found by PCA (Figure 1). This method was not necessarily expected to provide good classification. However, PCA is often useful for simplifying multidimensional data, and for a given dataset, it may provide good classification. The second model mapped the health data to the first principal component and the linear discriminant computed by LDA on the original ten-dimensional space (Figure 2).

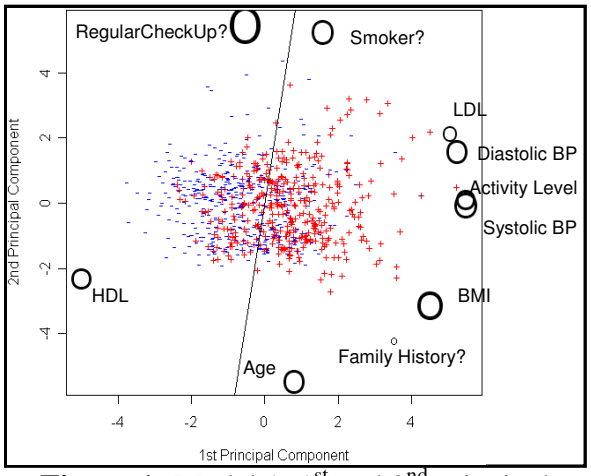

Figure 1: Model 1: 1st and 2nd principal components.

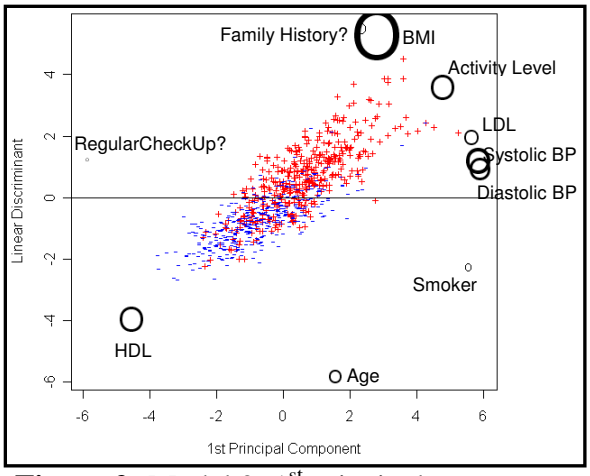

Figure 2: Model 2: 1st principal component and linear discriminant.

Enhancing Interpretability

To enhance interpretability of the reduced space and classification, an attraction metaphor concept used in prior visualization methods, including [9], was integrated with the mappings. In each two dimensional projection, the plot axes are linear combinations of ten factors related to diabetes risk. To convey the relative weights of the individual variables in these combinations, "**attracting anchors**", one for each of the risk factors in the high-dimensional data, were plotted on a circle that circumscribes the data. In terms of polar coordinates, each anchor's direction of attraction (θ_i) was defined using the ratio of each attribute's weights on the vertical and horizontal components of the two-dimensional space:

$$\theta_i = \begin{cases} \tan^{-1}(d_{i1} / d_{i2}) & d_{i2} \geq 0 \\ \pi + \tan^{-1}(d_{i1} / d_{i2}) & d_{i2} < 0 \end{cases} \quad (1)$$

d_{i1} is the weight of attribute i on the vertical axis plot component, and d_{i2} is the weight of attribute i on the horizontal axis component. To express the magnitude of the coefficients (i.e. the strength of attraction), the size of each anchor (S_i) was defined to be proportional to the length of the vector formed by the component coefficients d_{i1} and d_{i2}:

$$S_i = C\left(\sqrt{d_{i1}^2 + d_{i2}^2}\right) \quad (2)$$

C is a constant set such that the smallest anchors are large enough to be visible to a user.

Data Analysis, Model Training and Testing

Archimedes, a simulation-based diabetes prediction model, was chosen as the standard against which the models in this study were compared. Archimedes models biological processes related to diabetes, and its predictions have been validated by clinical trials [7]. A secondary dataset of consumer health information and corresponding Archimedes risk predictions was obtained from the ADA through its Diabetes PHD web application (www.diabetes.org/diabetesphd/). This website collects personal health information and uses Archimedes to give users probabilistic diabetes risk assessments. The dataset consists of individuals' health data and corresponding Archimedes predictions that were processed between late September and early December 2006. 821 observations for users with no past diagnosis of diabetes and no missing data values were retained. Univariate distributional analysis indicated that most of the variables in this dataset are reasonably representative of the U.S. population. Archimedes' 30-year percentage likelihood of diabetes onset was used for labeling each observation. Observations with 30-year predicted risk greater than the median risk level (13.11%) were labeled *high risk* and the others as *low risk*. Therefore, half of the observations were labeled *low risk* and half were labeled *high risk* using the results of a validated prediction model.

Model 1 (Figure 1) is a plot of the 821 observations mapped to the first two principal components calculated by PCA. The line through the data is an optimal linear separator of *low* and *high risk* as calculated by LDA on the two-dimensional data. The anchors are the circles surrounding the data points. Points plotted with "+" symbols have high risk labels while points with "-" symbols have low risk labels. It is important to note that in using the proposed methods, these labels would be used for model training but would not appear in practical instantiations unless Archimedes was separately used to predict each patient's risk. Visual inspection suggests that the first principal component (PC1), which explains the most variance in the data, also separates the *high* and *low risk* cases reasonably well. 10-fold cross-validation error from applying LDA to the 2-D space was 0.2983. The display also provides some sense of the degree of "confidence" in each prediction. Observations for which the classifier is more confident tend towards the left and right sides of the plot while those in the center are more uncertain. The variable weights that comprise the two principal components (Table 2) are depicted by the location and size of the attracting anchors. Higher levels of **systolic** and **diastolic blood pressure**, **BMI** and **physical activity level** (coded such that higher values indicate less activity) are positively correlated with each other and with high risk for diabetes. This is shown by relatively large anchors on the right side of the plot where PC1 is positive. As expected, **HDL** ("good cholesterol") is negatively correlated with these variables and with diabetes risk. The anchors indicate that strong risk factors for *high risk* patients in the lower right area of the plot are more likely to be high **BMI** and **family history**, as opposed to **LDL** or **diastolic blood pressure**. These patients are also more likely to be **older** and **non-smokers** since they are closer

to the **age** anchor and far from the **smoking** anchor. **Family history** is positively correlated with high risk, but the small anchor indicates it is not as influential as other variables.

Model 2 (Figure 2) was constructed by mapping the data to a space consisting of the first principal component and the linear discriminant (LD) computed by LDA. The linear separator in Model 2 is a horizontal line because the horizontal axis is the LDA discriminant computed in the

Table 2: PCA and LDA data transformations.

	Syst-olic BP	Diast-olic BP	LDL	HDL	Family History	Smoker	Regular Check Up	Physical Activity	Age	BMI
PC1	0.488	0.428	0.260	-0.371	0.081	0.134	-0.062	0.390	0.066	0.435
PC2	-0.002	0.130	0.111	-0.171	-0.097	0.443	0.668	0.009	-0.447	-0.301
LD	0.103	0.069	0.092	-0.322	0.190	-0.054	0.014	0.294	-0.242	0.823

original ten-dimensional space. Again, visual inspection shows the *high risk* observations separate from the *low risk*. For a physician, this display might allow quick identification of patients with extremely high risk who may need medical intervention. 10-fold cross-validated classification error from LDA was 0.2435. Note that the data is less scattered in Figure 2 due to correlation between PC1 and the LD. There was increased scatter in the first plot because, by definition, principal components maximize variance and are uncorrelated. So, while Model 2 improved risk stratification, it was less useful for separating the data more generally. Interpretations similar to those discussed for Model 1 about the confidence in predictions and the relevance of specific risk factors to the given population can also be made for Model 2.

Both Models 1 and 2 performed reasonably well in stratifying observations based on risk. As expected, Model 2 classified significantly better (χ^2=6.69, p=.01), and the most important risk factors aligned with what is typically thought to be the most important predictor of diabetes risk, obesity (measured by **BMI**). Model 1, on the other hand, scattered the data generally, allowing more insight into what factors explain differences (variance) among patients. Table 3 shows the 10-fold cross validated classification accuracy of some common classifiers applied to the original 10-dimensional dataset. The visual classifiers tended to classify the data with error comparable to other methods. Even Model 1, based on PCA, did not perform considerably worse, despite not using the data labels for training.

Table 3: 10-fold cross-validation error of other classifiers on the Diabetes PHD data.

Method	10-Fold CV error
Naïve Bayes	0.2728
Logistic Regression	0.2448
1-NN	0.3448
10-NN	0.2996
200-NN	0.3130

Discussion and Future Work

The methods presented in this paper provide a basis for future research on automated methods and data visualization that may be integrated with clinical information systems. Though dimensionality reduction creates a loss of information, as described in an example above,

visualization may help recover some clinically relevant risk information. In the future, the proposed methods may be integrated with interactive capabilities in a GUI interface. For example, the plotted points could be "clickable" and linked to electronic medical records of individual patients. In a dynamic system, risk factors could be added, removed or re-weighted interactively. Given a training set of labeled data, clinicians could train and apply their own data transformations to better fit their practice population. We are currently testing models for data from patients already diagnosed with type II diabetes in order to assess risk of diabetes complications such as heart disease. Once these models are developed and integrated into an interactive system, a sample of physicians will formally evaluate the visualization approach to risk assessment.

Conclusion

Dimensionality reduction and information visualization methods can be integrated with health data to produce two-dimensional data plots which (1) approximate the risk predictions of a validated model and (2) provide interpretability that makes the predictions more transparent for a primary care physician user interested in assessment of individual patients and populations.

Acknowledgements

The authors thank the ADA for use of the Diabetes PHD data, and Drs. Michelina Fato, MD and David Eibling, MD for their clinical input.

References

1. CDC, 2004, National Diabetes Fact Sheet: General Information and National Estimates on Diabetes in the United States, 2003, Rev Ed., U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA.
2. Yarnall, K.S.H., Pollak, K.I., Ostbye, T., Krause, K.M., and Michener, J.L., 2003, "Primary Care: Is There Enough Time for Prevention?," American Journal of Public Health. 93(4), 635-641.
3. Kahn III, C., Ault, T., Isenstein, H., Potetz, L., and Van Gelder, S., 2006, "Snapshot of Hospital Quality Reporting and Pay-for-Performance under Medicare.," Health Affairs. 25(1), 148-162.
4. Shneiderman, B., 2002, "Inventing Discovery Tools: Combining Information Visualization with Data Mining," Information Visualization. 1(1), 5-12.
5. Card, S.K., Mackinlay, J.D., and Shneiderman, B., 1999, Information Visualization: Using Visualization to Think, Morgan Kaufmann, San Francisco.
6. Hastie, T., Tibshirani, R., and Friedman, J., 2001, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics, Springer, New York.
7. Eddy, D.M. and Schlessinger, L., 2003, "Archimedes: A Trial-Validated Model of Diabetes," Diabetes Care. 26(11), 3093-101.
8. Knowler, W.C., Pettitt, D.J., Saad, M.F., and Bennett, P.H., 1990, "Diabetes Mellitus in the Pima Indians: Incidence, Risk Factors and Pathogenesis," Diabetes Metab Rev. 6(1), 1-27.
9. Olsen, K.A., Korfhage, R.R., Sochats, K.M., Spring, M.B., and J.G., W., 1993, "Visualization of a Document Collection: The Vibe System," Information Processing and Management. 29(1), 69-81.