Predicting the Diagnosis of Type 2 Diabetes Using Electronic Medical Records

Oliver Bear Don't Walk IV, David Joosten, Tim Moon

December 12, 2014

1 Introduction

As of 2013, over 382 million people worldwide have diabetes [5]. Diabetes puts patients at a higher risk for blindness, kidney failure, heart disease, and stroke and it is especially prevalent in the United States in racial groups with low access to healthcare, such as Native Americans (15.9%), African Americans (13.2%) and Hispanics (12.8%) [6]. Although the onset of diabetes mellitus type 2 (DMT2) can be prevented or delayed with behavioral changes, e.g. physical activity or dietary changes, an estimated 27.8% of people with DMT2 in the United States are undiagnosed. In order to improve diagnosis methodologies, supervised and unsupervised machine learning algorithms trained on electronic medical records (EMR) were implemented and evaluated for effectiveness.

2 Features

2.1 Dataset and Feature Extraction

This study uses a publicly available EMR dataset released by Practice Fusion in 2012 for a Kaggle competition [4]. It consists of de-identified records for 9,948 patients, among whom 1,904 have been diagnosed with DMT2. The data was extracted from 17 database tables, which include diagnosis histories, medication histories, physician visits, lab reports, smoking histories, and demographic characteristics. We had four original features from the raw input (age, gender, weight, BMI), in addition to indicator variables for DMT2 diagnosis. Binary features were added in the form of indicator variables for medication prescriptions, diagnoses, and anomalous lab report results. However, upon inspection, it became clear that Practice Fusion stripped some data related to lab reports, possibly to de-identify patients or to make the Kaggle contest more challenging. Thus, features related to lab reports were not included in this study.

2.2 Feature Selection

The vast majority of features extracted from the dataset were binary features related to diagnoses and prescriptions. In order to reduce the number of features to a manageable number, filter feature selection was applied to find which of these binary features were most relevant. Specifically, the mutual information with diabetes diagnosis was computed for the 200 most common diagnoses and the 200 most common prescriptions. The 20 binary features with the greatest mutual information, listed in Table 1, were used for the learning algorithms.

3 Models

3.1 Unsupervised Learning

An unsupervised clustering algorithm was applied to provide insight into the distribution of positive DMT2 cases in the feature space. The gap statistic, as defined in Hastie, et al. [2], was computed for the data using k-means clustering with $k = 2$ to $k = 7$. This was implemented using the fpc package for R [3]. Local maxima in the gap statistic were interpreted as the optimal numbers of clusters.

3.2 Supervised Learning

Methodology The EMR dataset was used to train several supervised learning algorithms implemented in R [7]. In order to evaluate algorithm effectiveness, 10-fold cross-validation was applied to compute the test error, precision, and recall for each of the models. The train error was also computed to provide insight into the variance and bias of the models. Learning curves were obtained by varying the size of the training set and computing the test error, precision, and recall.

Naive Bayes In order to provide a baseline with which to compare the results of future models, the

Feature	Type	Mutual Information
272.2 (Mixed Hyperlipidemia)	Diagnosis	0.025
401.1 (Benign Essential Hypertension)	Diagnosis	0.021
Lisinopril	Prescription	0.009
401.9 (Unspecified Essential Hypertension)	Diagnosis	0.008
Zocor (Simvastatin)	Prescription	0.006
272.4 (Other and Unspecified Hyperlipidemia)	Diagnosis	0.006
585.3 (Chronic Kidney Disease, Stage III (Moderate))	Diagnosis	0.005
782.3 (Edema)	Diagnosis	0.005
401 (Essential Hypertension)	Diagnosis	0.004
Lipitor (Atorvastatin Calcium)	Prescription	0.004
Simvastatin	Prescription	0.004
414.00 (Coronary Atherosclerosis of Unspecified Type of Vessel)	Diagnosis	0.004
414.01 (Coronary Atherosclerosis of Native Coronary Artery)	Diagnosis	0.004
715.16 (Osteoarthrosis Localized Primary Involving Lower Leg)	Diagnosis	0.004
443.9 (Peripheral Vascular Disease Unspecified)	Diagnosis	0.004
781.2 (Abnormality of Gait)	Diagnosis	0.004
428.0 (Congestive Heart Failure Unspecified)	Diagnosis	0.004
Lasix (Furosemide)	Prescription	0.003
Coreg (Carvedilol)	Prescription	0.003
Cozaar (Losartan Potassium)	Prescription	0.003

Table 1: Binary features with the greatest mutual information with diabetes diagnosis. Diagnoses are indicated with ICD-9 codes. Generic names for brand name prescriptions are indicated.

naive Bayes algorithm was applied to the classification of DMT2 diagnosis. As the only generative learning model applied in this paper (calculating $p(y|x)$ through estimating $p(x|y)$ and a prior $p(y)$), naive Bayes makes the strong assumption that features are conditionally independent. It is implemented in the e1071 package for R [1].

Logistic Regression Following on the results of the naive Bayes model, logistic regression was investigated since it make weaker assumptions concerning the conditional probability distribution of features. Specifically, as a generalized linear model based upon the conditional mean $p(y|x)$ subject to the Bernoulli distribution, it does not require features to be conditionally independent nor multivariate normal. Further, the data provides a sufficient number of samples (almost 10,000) for the effective use of logistic regression using the glm package for R [7].

Support Vector Machines (SVM) Support vector machines (SVMs) are powerful classifiers that involve constructing a hyperplane decision boundary in the feature space that maximizes the functional margin with the data. They are especially well-suited for modeling nonlinear behavior since one can use kernels to project data into high-dimensional (possibly infinite-dimensional) feature spaces. Since the data was not expected to be separable, ℓ_1 regularization was applied. SVMs are implemented in the e1071 package for R [1].

k-Nearest Neighbors (KNN) Since it is possible that patients with DMT2 are present in clusters throughout the multi-dimensional feature space, k-nearest neighbors (KNN) is a reasonable alternative to the parametric models explored thus far. By using an odd number of patients mapped in the feature space nearest (using Minkowski distance with parameter 2) to a sample that requires a class assignment, KNN can model highly-localized phenomena and nonlinear behavior. KNN is implemented with the kknn package for R [8].

Decision Trees Finally, in order to better communicate the hierarchy of indicators of DMT2, this paper explores a single white box approach using decision trees. By applying this non-parametric greedy algorithm whose objective it is to maximize information gain in a top-down search of features, this paper is able to provide visualization of some of the decision boundaries with respect to individual features. While this approach is unlikely to capture feature in-

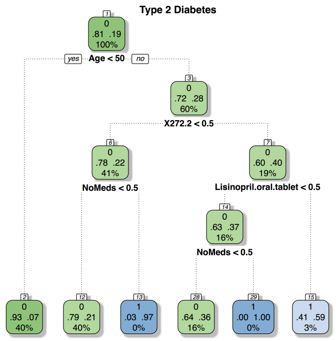


Figure 1: Decision tree (pruned) showing diagnosis, medication and demographic features.

teraction (since each feature's decision boundary is calculated in isolation), it should provide insight into the feature space unavailable from the other models explored. This is implemented with the rpart package for R [9]. See Figure 1 for an example decision tree.

4 Results

4.1 Clustering Analysis

Results for the gap statistic analysis are shown in Figure 2. The gap statistic for varying k applied to the k-Means Cluster algorithm indicates no optimal number of clusters between $k = 2$ through $k = 7$. Specifically, the gap statistic for varying k applied to the k-Means Cluster algorithm did not yield any local maximum, and this indicates no optimal number of clusters between $k = 2$ through $k = 7$. Further, the cluster sizes when $k = 2$ were observed to be approximately equal, which does not correspond to the relative sizes of samples with (1,904) and without (8,044) a positive diagnosis of DMT2.

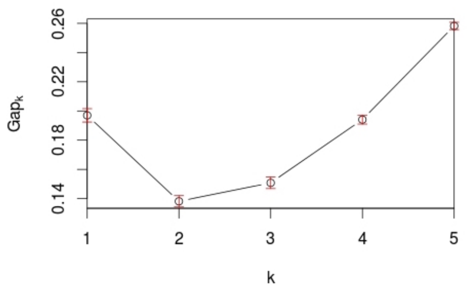


Figure 2: Gap statistic indicates data cannot be meaningfully separated into two classes.

4.2 SVM Analysis

10-fold cross-validation was used to compute the test error of SVMs with several kernels. Results are summarized in Table 2. Note that results are similar with the linear, polynomial, and radial kernels. The radial kernel was chosen for the SVM since it projects into an infinite-dimensional feature space, and hence may better reproduce the nonlinear behavior of the data. The test error, precision, and recall of the SVM were also calculated as the cost function parameter for ℓ_1 regularization was varied. Results are shown in Figure 3. The error was minimized by choosing the cost function parameter to be $C = 1$, although it did not vary significantly.

Kernel	Equation	Test Error
Linear	$u^T v$	0.183
Polynomial	$\left(\gamma u^T v + c_0\right)^d$	0.183
Radial	$e^{-\gamma \|u-v\|^2}$	0.183
Sigmoid	$\tanh\left(\gamma u^T v + c_0\right)$	0.254

Table 2: Results from 10-fold cross-validation on SVMs with several kernels. The cost function parameter is $C = 1$. The parameter values are $\gamma = 1/24$ and $c_0 = 0$.

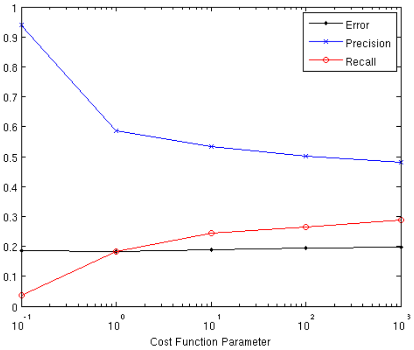


Figure 3: Test error, precision, and recall of an SVM with different values of the cost function parameter.

4.3 Cross-Validation

Results from 10-fold cross-validation are summarized in Table 3.

Model	Test Error	Train Error	Precision	Recall
Naive Bayes	0.47	0.46	0.18	0.39
Logistic Regression	0.18	0.18	0.60	0.24
SVM	0.18	0.15	0.59	0.18
k-Nearest Neighbors	0.22	0.07	0.42	0.31
Decision Trees	0.19	0.09	0.57	0.11

Table 3: Results from 10-fold cross-validation on supervised learning algorithms.

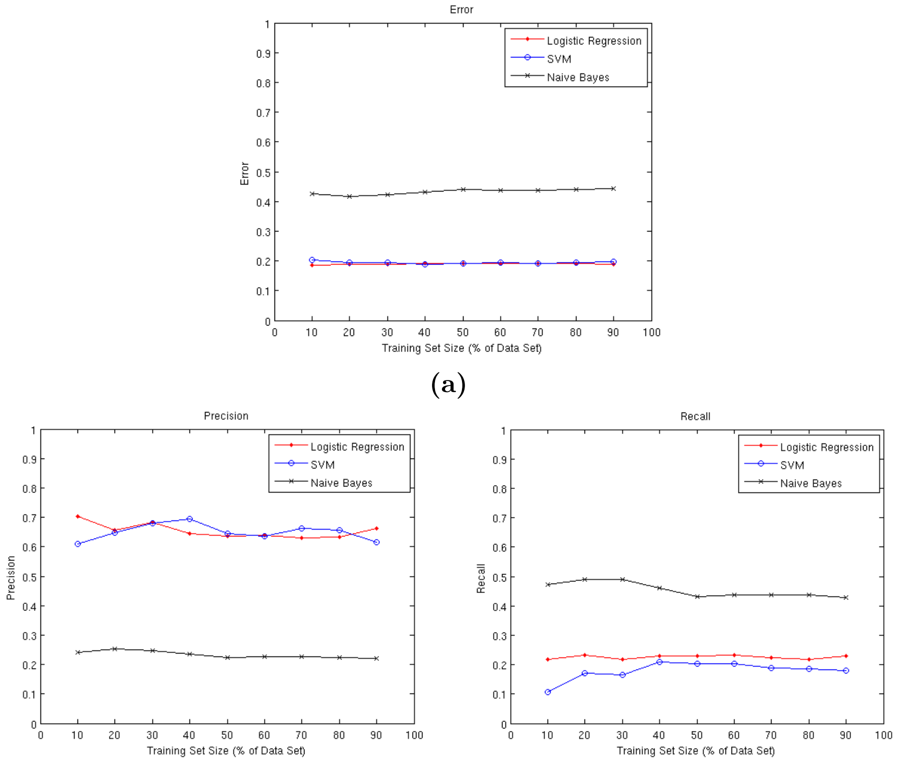
4.4 Learning Curves

Several learning curves for the naive Bayes, logistic regression, and SVM classifiers are shown in Figure 4.

5 Discussion

5.1 Method Evaluation

Inspecting Table 3, we see that SVMs and logistic regression are the methods that yield the smallest generalization error, approximately 18%. The SVM is particularly interesting because changing the



(a)

(b) **(c)**

Figure 4: Learning curves for naive Bayes, logistic regression, and SVM classifiers, showing the effect of training set size on (a) test error, (b) precision, and (c) recall.

cost function parameter for ℓ_1 regularization causes a tradeoff between precision and recall. Specifically, increasing the cost function parameter decreases the precision and increases the recall. This suggests that the cost function parameter can be adjusted to tune the precision and recall to match the needs of doctors.

5.2 Data Implications

Each model's performance may also indicate characteristics about the underlying data. Firstly, the poor performance of the naive Bayes model relative to the others may indicate that each feature is not conditionally independent of every other feature. Further, the relative weakness in the results of KNN combined with the results of our unsupervised clustering model indicate that the data is not best described by several independent clusters in the feature space. Instead, those models that applied a discriminative decision boundary, namely logistic regression and SVM, performed best in classifying patients.

5.3 Bias and Variance

One can estimate the relative contributions of bias (model limitations) and variance (overfitting) to the error of a model by comparing the test error and train error. From Table 3, we see that the test error and train error are very close for naive Bayes and logistic regression, suggesting that the bulk of the error is due to bias. This is corroborated by their flat learning curves, which indicates that there is some inherent error in the models even when the training set is large. On the other hand, the train errors for k-nearest neighbors and decision trees are fairly small compared to the test error, implying that the error is largely due to variance. Finally, the SVM has a train error that is moderately smaller than the test error, indicating that both bias and variance contribute to error in the SVM.

6 Conclusion

Electronic medical records (EMR) were used to train learning algorithms for DMT2 diagnosis. A variety of supervised learning algorithms were evaluated and it was found that SVMs and logistic regression produced the smallest error. SVMs are especially promising since one can adjust their behavior with different choices of kernel or cost function parameter to suit the needs of medical practitioners trading off false negatives and false positives.

Based upon the results, future studies should attempt to reduce the bias present in the logistic regression and SVM models. Specifically, new features such as genetic markers, lifestyle factors and more relevant lab tests (e.g. glucose, which was crucially missing) would provide additional dimensions along which to separate classes. Furthermore, future work should incorporate time series data, which is crucial for identifying the onset of DMT2 in a particular year. This will account for the possibility of internal structure that is currently not captured.

References

[1] Dimitriadou, Evgenia, et al. "Misc functions of the Department of Statistics (e1071), TU Wien." R package (2008): 1-5. http://cran.r-project.org/web/packages/e1071/

[2] Hastie, Trevor, et al., "Estimating the number of clusters in a data set via the gap statistic", Journal of the Royal Statistical Society Series B, vol 2, no 63, p.411 - 423.

[3] Hennig, Christian. "fpc: Flexible procedures for clustering." R package version 2 (2010): 0-3. http://cran.r-project.org/web/packages/fpc/

[4] "Identify Patients Diagnosed with Type 2 Diabetes." Kaggle. July 10, 2012. Accessed October 3, 2014. https://www.kaggle.com/c/pf2012-diabetes.

[5] "IDF Diabetes Atlas, 6th Edition." International Diabetes Federation. 2014. Accessed October 11, 2014. http://www.idf.org/diabetesatlas/introduction.

[6] "National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States." Centers for Disease Control and Prevention. 2014. Atlanta, GA: U.S. Department of Health and Human Services. Accessed October 11, 2014. http://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf.

[7] R Core Team. "R: A language and environment for statistical computing." R Foundation for Statistical Computing. Vienna, Austria. 2014. http://www.R-project.org/

[8] Schliep, Klaus and Klaus Hechenbichler. "KKNN: Weighted k-Nearest Neighbors." R package. 2014. http://cran.r-project.org/web/packages/kknn/

[9] Therneau, Terry, et al. "RPart: Recursive Partitioning and Regression Trees." R package. 2014. http://cran.r-project.org/web/packages/rpart/