Diabetes Prediction in Children Using the *Naive Bayes* Machine Learning Algorithm

Viswanath Kasireddy
Vandegrift High School

Mentors: Dr. PV Subba Reddy
Mr. Galen Goodreau

Abstract

Diabetes is one of the fastest growing long-term diseases, affecting millions of people worldwide, including children and teens. With no existing cure the only solution is to effectively predict the disease at an early stage in order to prevent it. The objective of this research paper is to develop a prediction algorithm to aid in the existing decision support systems for the prediction and diagnosis of Diabetes. The development of the prediction algorithm will be supported by data mining and machine learning techniques. Through the use of electronic data storage and maintenance systems, most modern hospitals maintain large amounts of patient data. The prediction algorithm attempts to use this data to assist in the diagnosis of the disease in a cost-effective way.

The prediction algorithm, which I designed and developed, is located at https://github.com/Viswa-Kasireddy/NaiveBayesSolver-to-predict-risk-of-Diabetes-

Introduction

Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose. Thirty years ago, type 2 diabetes mellitus was thought to be a rare occurrence in children and adolescents. Now it is increasing at an alarming rate in children and adolescents. A recent study found that type 1 diabetes in children aged up to 9 years increased by 21 percent between 2001 and 2009, while type 2 diabetes among youths aged 10-19 years rose by 30.5 percent in advanced countries like the US, UK, and Japan. This is a serious condition as these children will enter adulthood with increased health complications and a high risk of passing the disease on to offspring.

Our society will face major challenges due to the increasing prevalence of type 2 diabetes mellitus in children. If we do not make steps towards reducing the prevalence of diabetes, more and more individuals will be affected.

Methodology

Risk Factors and Sub-Classifiers

About Diabetes: The most common types of diabetes are type 1, type 2, and gestational diabetes. Chronic complications of diabetes include accelerated development of cardiovascular diseases, end-stage renal disease, loss of visual acuity, and limb amputations. These complications contribute to the excess morbidity and mortality in individuals with diabetes.

The attributes and values (Table 1) of patient records that are taken into consideration for a Diabetes diagnosis were chosen for their compatibility in use of prediction.

Table 1: Attributes and Values

Attributes	**Values**
Weight	Obese, Overweight, Normal, Underweight
Blood Sugar Levels	High (above 300 mg/dl blood glucose) Normal (70 to 100 mg/dL while fasting, non-diabetics) Low (below 40 mg/dl blood glucose)
Activity	High activity, Medium activity, Low activity
Family History	Sibling with disease, Parent with disease, None
Race	African-American, Hispanic, Native American, Asian, Pacific Islander, Caucasian
Age	Baby (0 -1 years) Toddler (1-3 years) Preschool (3-5 years) Grade School: (5-12 years) Teen (12- 18 years) Young Adult (18-21 years)
Sex	Male, Female
Birth Weight	Low, Normal, High
Gestational Diabetes	Yes, No
Viral Infections	Yes, No
Autoimmune Diseases	Yes, No

Risk Prediction using *Naive Bayes* Classifiers

Data Mining: An analytical process designed to explore large volumes of raw data in search of consistent patterns, and then validate findings by applying the detected patterns to new subsets of data. The goal of data mining is prediction. Disease prediction plays an important role in data mining. Data Mining is used intensively in the field of medicine to predict trends and highlight patterns.

Medical data: Consists of clinical data, lab results, physicians diagnosis, investigative results for specified medical conditions. This data, usually in a raw form, is most often incomplete, lacking attribute values and being mostly aggregated. For the sake of this project we need data transformed into a readable format for all the related to attributes for Diabetes in children.

Supervised Learning: Machine learning techniques for building predictive models from known inputs and response data. Supervised learning is a type of machine learning algorithm that uses a known dataset (called the training dataset) to make predictions. The training dataset includes input data and response values. Using the data set, the supervised learning algorithm seeks to build a model that can make predictions of the response values for a new dataset. A test dataset is often used to validate the model. Using larger training datasets often yield models with higher predictive power that can generalize well for new datasets. Supervised learning includes two categories of algorithms, Classification and Regression.

Classification: for categorical response values, where the data can be separated into specific "classes"
Regression: for continuous-response values
Common classification algorithms:

Support vector machines (SVM)
Neural networks
Naïve Bayes classifier
Decision trees
Discriminant analysis
Nearest neighbors (kNN)

Laplace Smoothing:

If a given class and feature value never occur together in a training set, then the frequency based probability estimate will be zero. This is problematic since it nullify the total estimate. Therefore all the existing probabilities should be adjusted such that no probability is zero. This is because while this event may not have occurred presently, there is a chance it will occur in the future. Laplace smoothing is applied for Maximum Likelihood estimates. We can assume that our training set is large enough that adding one to each count would only make a negligible difference in the estimated probabilities, and eliminate the possibility of a probability of zero.

Bayes Theorem: Bayesian reasoning is applied to decision making and inferential statistics that deal with probability. It uses the knowledge of prior events to predict future events and their probability. The Bayes theorem finds the probability of an event occurring given the probability of another event that has already occurred.

Let X = {x1, x2, . . . , xn} be a sample, whose components represent values made of a set of n attributes. In Bayesian terms, X is considered "evidence". Let H be some hypothesis, such as the data X belongs to a specific class C. For classification problems, our goal is to determine P(H|X), the probability that of hypothesis H, given that it holds the "evidence", (i.e. the observed data sample X). In other words, we are looking for the probability that sample X belongs to class C, given that we know the attribute description of X.

According to Bayes' theorem, the probability that we want to compute P (H|X) can be expressed in terms of probabilities P(H),P(X|H), and P(X) as P (H|X) = P(X|H) P(H)/P(X), and these probabilities may be estimated from the given data.

Naïve Bayes Classifier: a classifier which uses the Bayes Theorem. It predicts membership probabilities for each class (the probability that a given record or data point belongs to a particular class). The class with the highest probability is considered as the Maximum A Posteriori (MAP), or most likely class. The Naive Bayes classifier assumes that all the features are unrelated to each other. Presence or absence of a feature does not influence the classes.

The naive Bayesian classifier works as follows:
1. Let T be a training set of samples, each with their class labels. There are k classes, C1, C2, … Ck. Each sample is represented by an n-dimensional vector, X={x1, x2, . . . , xn}, depicting n measured values of the n attributes, A1, A2, . . . An respectively.

2. Given a sample X , the classifier will predict that X belongs to the class having the highest a posteriori probability, conditioned on X. That is X is predicted to belong to the class Ci if and only if $P(C_i|\mathrm{X}) > P(C_j|\mathrm{X})$ for $1 \leq j \leq m, j \mathrel{!=} i$

Thus we can find the class that maximizes P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem, P(Ci|X) =P(X|Ci)P(Ci)/P(X)

3. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample. Mathematically this means that

$$P(\mathbf{X}|C_i) \approx \prod_{k=1}^{n} P(x_k|C_i)$$

The probabilities $P(x_1|C_i),P(x_2|C_i),\ldots,P(x_n|C_i)$ can easily be estimated from the training set. Recall that here x_k refers to the value of attribute A_k for sample X. In order to predict the class label of X, $P(\mathrm{X}|C_i)P(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of X is C_i if and only if it is the class that maximizes $P(\mathrm{X}|C_i)P(C_i)$.

Why was the Naïve Bayes Classifier was chosen?
The Naïve Bayes classifier was chosen for this research because:

- Studies comparing classification algorithms have found the Naive Bayesian classifier to be comparable in performance with decision tree selected neural network classifiers.
- Bayesian classifiers exhibit high accuracy and speed when applied to large databases.
- It is faster to predict classes using this algorithm than other classification algorithms.
- The Naive Bayes Algorithm is simple and easy to implement.
- The Naive Bayes Algorithm makes probabilistic predictions.
- The Naive Bayes Algorithm is highly scalable. It scales linearly with the number of predictors and data points.
- The Naive Bayes Algorithm is easily updatable for incoming data
- Naive Bayes can be used for Binary and Multiclass classification. It also provides different variations, such as GaussianNB, MultinomialNB, BernoulliNB.
- The Naive Bayes Algorithm easily handles continuous and discrete data.
- The Naive Bayes Algorithm can be trained on small dataset
- The Naive Bayes Algorithm can deal with missing values easily. It is a great choice for medical predictions where data is inconsistent or missing.

Steps to implement the Naïve Bayes Classifier with Laplace Smoothing

Training Data: Pre-existing medical data is used to form classifiers. Each tuple or sample is assumed to belong to a predefined class as determined by the class label attribute. The set of tuples are then used to construct a model. This model (Figure 1) represents the classification rules.

Learning phase:

1. Input the Training data into Naïve Bayes classifier
2. Calculate the probabilities for each attribute for both classes for the Training data

3. Apply Laplace smoothing to calculate the smoothed probability
4. Configure the prediction algorithm

Test Phase: When a new patient record in the given sample data below is input, the algorithim will calculate the probability for both the values of Diabetes = “Yes” and “No”

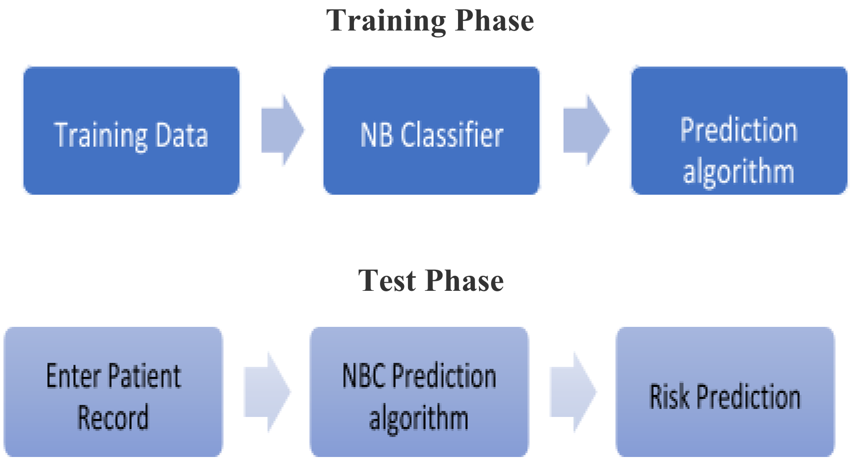
Steps for calculating the probability for both attributes of Diabetes:

1. Classes in which data can be classified:
   Patient at risk for Diabetes
   Patient not at risk for Diabetes
2. Enter the patient’s record (Test Data)
3. Set as input to the classifier
4. Calculate the maximized probability for both classes
5. Decide the class for the patient’s record

Calculations:

1. For Attribute = “Weight” and the value = “Normal”, Probability of Diabetes = “Yes” is calculated. Probability(Weight = Normal) for Class (Diabetes = Yes)
2. This process is continued for all the attribute values from the sample data
3. The product of the probabilities for all the attributes values is multiplied with the probability of the Diabetes = “Yes” overall. Probability of the occurrence of Diabetes => “Yes” = Probability(Weight = Normal) for Class (Diabetes = Yes) * Probability(Activity = Poor) for Class (Diabetes = Yes) * Probability (Diabetes = Yes)
4. Similarly the probability of Diabetes = “No” for all attribute values in the sample data is calculated. Probability(Weight = Normal) for Class (Diabetes = No)
5. The product of all probabilities for all attributes values is multiplied with the probability of the Diabetes = “No”. Probability of the occurrence of Diabetes => “No” = Probability(Weight = Normal) for Class (Diabetes = Yes) * Probability(Activity = Poor) for Class (Diabetes = No) * Probability (Diabetes = No)
6. Compare both probabilities and determine the occurrence of the event to be more likely. If the (Probability of the occurrence of Diabetes => “Yes”) > (Probability of the occurrence of Diabetes => “No”), then the prediction result is a high likelihood of developing Diabetes. Otherwise, the prediction result is a low likelihood of developing Diabetes.

Figure 1: Flowchart for implementation of classification on patient data:



Results

Table 2: Sample of Training Data

Weight	Activity	Family History	Race	Age	Sex	Birth Weight	Blood Sugar	Gestational Diabetes	Viral Infections	Autoimmune Disease	Diabetic
Obese	Low	Sibling	Pacific Islander	Child	Female	Low	High	No	No	No	Yes
Overweight	Low	Sibling	Pacific Islander	Child	Female	Low	Low	Yes	No	No	Yes
Obese	Medium	Sibling	Pacific Islander	Child	Male	High	Normal	No	No	No	Yes
Obese	Low	Sibling	Hispanic	Child	Female	Low	High	No	No	Yes	Yes
Normal	Low	Sibling	Pacific Island	Teen	Male	Low	Low	No	Yes	No	Yes

			er								
Overweight	Low	Parent	Asian	Child	Female	Low	Low	No	No	No	No
Obese	Low	None	Pacific Islander	Child	Female	Low	High	No	No	No	Yes

Table 3: Sample Summation of Weight Attribute (Learning Phase)

Weight	**Diabetes = Yes**	**Diabetes = No**
Underweight	7	1
Normal	3	2
Overweight	4	5
Obese	8	9
Sum	22	17

Table 4: Sample Probability Calculation of Weight Attribute (Learning Phase)

Attribute/Classifier	**Probability of Yes**	**Probability of No**
Underweight	.3181	.0454
Normal	.1363	.0909
Overweight	.1818	.2272
Obese	.3636	.4090

Table 5: Sample Patient A Input

Weight	**Activity**	**Family History**	**Race**	**Age**	**Sex**	**Birth Weight**	**Blood Sugar**	**Gestational Diabetes**	**Viral Infections**	**Autoimmune Disease**
Normal	Low	None	Pacific Islander	Teen	Male	Normal	High	Yes	no	no

Table 6: Sample Patient A Output

Input Patient Records	**Prediction**
Patient A	High Risk for Diabetes

Using Training Data such as those featured in Table 2, the prediction algorithm was developed in the learning phase. The calculations of the learning phase led to the results of Tables 3 and 4. Upon completion of the prediction algorithm, it was fed input data in order to refine accuracy. Table 5 represents a sample input of the completed prediction algorithm, while Table 6 represents the algorithms response to the sample input.

Conclusion

The aim of this research project was to develop a prediction algorithm using the Naive Bayes Classifier. Given that there is a large amount of unstructured clinical data in different formats, it is challenging to accurately predict the risk of diabetes in children using it. However, the prediction algorithm can handle cases where data is missing, while accurately predicting the risk of diabetes in children. In doing so, preventative measures can be taken at an early stage. This algorithm can be used as a plugin to existing computer tools in decision support systems on any platform to further enhance diagnoses of Diabetes.

The algorithm was developed with the attributes of the sample data, but can incorporate new attributes to improve the accuracy of predictions quickly through its flexibility. Since the algorithm is capable of scaling easily with the increase in predictors and data points with newly adaptable data mining techniques and advances in machine learning, the acquired knowledge of heterogeneous data can be transformed into structured data suitable for machine learning. This leads to the accuracy of the prediction algorithm being able to improve significantly, as well as being modeled into an automated system.

For the future, I plan to perform additional experiments with more datasets to improve the classification accuracy of model to predict specific types of Diabetes (Type1 or Type2).

I would also like to make the NB Classifier tool to be integratable into existing Decision Support Platforms, through platform independence and cost effectiveness.

Acknowledgments

I would like to thank Dr. PV Subba Reddy for his help. Without his expertise, I would not have been able to choose which attributes the algorithm would focus on, as well as accurately analyze and refine the prediction results.

I would like to thank Mr. Goodreau for his insight and guidance of the algorithm's programming and execution on the JAVA platform, as well as advice on further steps for the project.

References

1. D. Ratnam, P. HimaBindu, V. MallikSai, S. P. Rama Devi and P. RaghavendraRao, "Computer-Based Clinical Decision Support System for Prediction of Heart Diseases Using Naïve BayesAlgorithm", International Journal of Computer Science and Information Technologies, vol. 5, no. 2 (2014) pp.23842388.
2. "Machine Learning" course by Stanford University on CourseEra. https://www.coursera.org/learn/machine-learning
3. Naïve Bayes Classifier: http://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/
4. Naïve Bayes Classifier: http://responsive.media.mit.edu/wp-content/uploads/sites/5/2014/01/Class-4-Naive-Bayes.pdf
5. Choosing a Machine Learning Algorithm: http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/
6. Naïve Bayes and Laplace Smoothing: https://classes.soe.ucsc.edu/cmps140/Winter17/slides/3.pdf
7. Naïve Bayes Classifier by K. Ming Leung, https://pdfs.semanticscholar.org/4632/2a2c113db0e28dd1b37aeabd5668adf77de6.pdf