Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner

Jianchao Han, Juan C. Rodriguze, Mohsen Beheshti
Department of Computer Science
California Statement University Dominguez Hills
jhan@csudh.edu, jrodriguez236@cp.csudh.edu, mbeheshti@csudh.edu

Abstract

Data mining techniques have been extensively applied in bioinformatics to analyze biomedical data. In this paper, we choose the Rapid-I's RapidMiner as our tool to analyze a Pima Indians Diabetes Data Set, which collects the information of patients with and without developing diabetes. The discussion follows the data mining process. The focus will be on the data preprocessing, including attribute identification and selection, outlier removal, data normalization and numerical discretization, visual data analysis, hidden relationships discovery, and a diabetes prediction model construction.

1. Introduction

Modern computers have made it so that every field of study is generating data at an unprecedented rate. Computers can process data in ways and speeds humans could never achieve. Data mining is the entire process of applying a computer-based methodology for developing knowledge from data.

Data mining is an iterative process in which progress is defined by discovery, through either manual or automatic methods. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined ideas about what will constitute an "interesting" outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data.

In practice, the two primary goals of data mining tend to be prediction and description [1]. Prediction involves using some attributes or fields in the data set to predict unknown or future values of other attributes of interest. On the other hand, description focuses on finding patterns for describing the data so that humans can interpret it. To achieve the goals of prediction and description one must follow a data mining process. There are many different versions of data mining processes and many opinions on how to approach them. This paper focuses on the RapidMiner software package to preprocess and analyze diabetes data and mine a diabetes prediction model.

"Kidney failure is a deadly complication of diabetes, and Pimas, so far as scientists can tell, have the world's heightest rate of type 2 diabetes." [3] The objective of this analysis is to understand any general relationships between different patient characteristics and the propensity to develop diabetes.

Rapid-I's RapidMiner software package supports all steps of data mining process [2]. It is a Java-based open-source software and can be used as a Java API. It also provides a simple and friendly GUI. RapidMiner uses internal XML representations to ensure standardized interchange format of data mining experiments.

Using RapidMiner, we will easily deliver an analysis report and a prediction model. The analysis report will summarize the data and their associations to developing diabetes. The prediction model will be a decision tree that should help in predicting whether a patient will develop diabetes using the data gathered.

The data set used in this project is excerpted from the UCI Machine Learning Repository [4]. The Pima Indians Diabetes Data Set contains 8 categories and 768 instances gathered from a larger databases belonging to the National Institute of Diabetes and Digestive and Kidney Diseases. The selection of these instances is as follow: All patients are females at least 21 years old of Pima Indian heritage.

2. Data preprocessing

Most of the data sets used in data mining were not necessarily gathered with a specific goal in mind. Some of them may contain errors, outliers or missing values. In order to use those data sets in the data mining process, the data needs to undergo preprocessing, using data cleaning, discretization and data transformation [5]. It has been estimated that data preparation alone accounts for 60% of all the time and effort expanded in the entire data mining process [6]

978-0-7695-3431-2/08 $25.00 © 2008 IEEE
DOI 10.1109/FGCN.2008.226

2.1. Feature identification and categorization

Attributes are usually described by a set of corresponding values. Features described by both numerical and symbolic values can be either discrete (categorical) or continuous. Discrete features concern a situation in which the total number of values is relatively small (finite), while with continuous features the total number of values is very large (infinite) and covers a specific interval (range) [7]. The following attributes can be gathered from the data set [2]:

- Pregnant: Number of times of pregnant
- Plasma-Glucose: Plasma glucose concentration measured using a two-hour oral glucose tolerance test. Blood sugar level.
- DiastolicBP: Diastolic blood pressure (mmHg)
- TricepsSFT: Triceps skin fold thickness (mm)
- Serum-Insulin: 2-hour serum insulin (mu U/mt)
- BMI: Body mass index (w in kg/h in m)
- DPF: Diabetes pedigree function
- Age: Age of the patient (years)
- Class: Diabetes onset within five years (0 or 1)

These characteristics need to be kept in mind as the data set is cleaned. Most of this work can be done in RapidMiner itself. After importing samples of the Pima Indian Data Set, changing default attribute titles, and renaming the values of attribute Class from (0, 1) to (No, Yes), one can obtain a proper categorized attribute table.

2.2. Outlier removal and feature selection

Using the Data View in RapidMiner, the attributes can be sorted in different ways to view patterns and values. Those rows with missing value should be removed for the attributes Plasma-Glucose, DiastolicBP, and BMI.

Outliers are extreme values that lie near the limits of the data range or go against the trend of the remaining data. Identifying outliers is important because they may represent errors in data entry. In addition, even if an outlier is a valid data point and not in error, certain statistical methods are sensitive to the presence of outliers and may deliver unstable results [5]. The Density Plotter in RapidMiner helps to easily discover outliers like one with value 99 found in the TricesSFT. The DistanceBasedOutlierDetection operator can be used to verify it.

The Plot View can also be used to view the data using different plotters. One plotter used was the Scatter Matrix, which helps to visually observe the correlation between two attributes. This plotter illustrates the need to focus on the correlation between BMI vs. TricepsSFT. One should take care to avoid feeding correlated attributes to one's data mining and statistical models. At best, using correlated variables will overemphasize one data component; at worst, it will cause the model to become unstable and deliver unreliable results [5]. TricepsSFT has 227 missing values (zero values). Therefore, this attribute could be removed based on the number of missing values and its relationship to BMI. The variable Serum-Insulin was also removed from the data set because of the number of missing values. The Histogram Plotter can be used to check the rest of the attributes for correlations.

2.3. Data normalization

Numerical attributes usually have ranges that vary greatly from each other. This data set is no exception. For example, compare the range of variables of BMI, which range from 18.2 to 67.1, to those of Plasma-Glucose, which range from 44.0 to 199.0. These types of differences in the ranges can lead to a tendency for the variable with greater range to have undue influence on the results. Therefore, data sets should normalize their numerical variables, to standardize the scale of effect each variable has on the results [5].

The RapidMiner software provides tools to quickly normalize numerical attributes without the risk of human error. In this project, the Min-Max normalization model was applied to transform the attribute's values to a new range, 0 to 1. The formula used to normalize attribute X is as follows:

$$Normalized\ X = (X-min(X))/(max(X)-min(X)).$$

2.4. Numerical data discretization

After normalizing the numerical attributes, it becomes clear that it would be easier to associate diabetes with other attributes by grouping them, which can be done by discretizing these attributes. The discretization of numerical attributes can be performed before or after normalization. For easy understanding, we choose to discretize numerical attributes before normalization or disabling normalization.

Discretization of numerical attributes can be either manually done based on user specification, called UserBasedDiscretization, or automatically done, called BinDiscretization. We discretize the attributes Pregnant, Plasma-Glucose, DiastolicBP and BMI manually, while DPF and Age automatically.

The attribute Pregnant is discretized into three bins: low (0,1), medium (2, 3, 4, 5), and high (>6). Using the information on WebMD, one can discretize the following attributes easily: Plasma-Glucose is grouped into low (<95), medium (95-140), and high (>140) [8]. DiastolicBP is divided into normal (<80), prehypertension (80-89), and high (>90) [9]. BMI is mapped into three intervals: low (<18.5), healthy (18.5-24.9), overweight (25-29.9), obese (30-34.9), and severely-obese (>35) [10]. DPF is discretized using the average value as follows: low (<0.42), medium (0.42-0.82), and high (>0.82). Finally, Age is automatically discretized into three ranges: range1 (<41), range2 (41-61), and range3 (>61).

3. Data analysis and hidden relationships discovery

Data analysis that can be done with the RapidMiner software includes graphs and tables, as well as various charts and plots. The RapidMiner Histogram Color Matrix was used to visually compare the values of the attributes and see the relationships with the Class attribute values (Yes, No), which finds that the patients with higher Plasma-Glucose values are very likely to develop diabetes and most with low Plasma-Glucose values do not develop diabetes within five years.

Further analyzing this relationship between Plasma-Glucose and Class by using a box plot affirms the above observation. To help clarify whether the observation might be of value, a Naïve Bayes learning tool is applied, where the attributes are considered to be random variables, and the data are considered to be known. The parameters are regarded as coming from a distribution of possible values, and Bayesians look to the observed data to provide information on likely parameter values [5]. This verifies that the initial observation appears to be correct.

After the data preprocessing, our next goal is to find generally associations in the data in order to understand the relationships between the attributes and whether the patients go on to develop diabetes. With the discretization of numerical attributes, we will focus on the sub-groups (bins) created instead of the individual values of the attributes to minimize the complexity of the analysis without losing accuracy.

RapidMiner provides a very useful tool, BasicRuleLearner, for helping narrow observation down, which sifts through the data and finds general relationship rules, such as

If Plasma-Glucose = high then Yes (124/60)
If Pregnant = medium the No (28/65)
If DPF = low the No (26/50)

One may realize that some rules have very low accuracy, thus may be misleading. It would be incorrect to simply look at one attribute and then look at the result. It must be better to see the results when two or more attributes are combined, not to mention to combine all attributes. This can be achieved using CHAID (Chi-squared Automatic Interaction Detector) decision tree [11]. CHAID detects interaction between variables in the data set by identifying discrete groups of respondents, and seeks to predict what the impact will be on the dependent variables by taking their responses to explanatory variables. Since CHAID requires statistics data, it is not necessary to discretize numerical variables.

4. Constructing the Prediction Model

Data analysis and hidden relationship reveal that have been made so far can be used to either modify the attribute or help get a better understanding of the attribute values. Now we need to construct a predictive model to estimate whether a patient will develop diabetes within an acceptable percentage of certainty, instead of simply shedding light about the data itself. RapidMiner provides means for this purpose. We choose two main options: the ID3 Algorithm and the Decision Tree.

A decision tree can be learnt by splitting the source data set into subsets based on an attribute value test [12]. This process is repeated on each derived subset in a recursive manner. The recursion is completed when splitting is either non-feasible or a singular classification can applied to each element of the derived subset. A random forest classifier uses a number of decision trees, in order to improve the classification rate. The decision is not only helpful in representing the current data relationships, but also able to apply other data to the algorithm and test how well it works at predicting the outcome.

RapidMiner supports to generate a decision tree. A part of the decision tree automatically produced by RapidMiner is shown in Figure 1, where the Plasma-Glucose attribute is chosen as the root node. This further reinforces our original observation in the prior section. The decision tree tells us that Plasma-Glucose is the main attribute that will lead us to knowing whether a patient will develop diabetes. The actual data set states that 248 patients develop diabetes and 475 do not. This decision tree predicts that there are 200 patients that develop diabetes and 523 that do not. Of those 200 patients, 19 are wrong, which brings the correct predictions down to 181. This means the decision tree has 72% of accuracy.

One can also choose the ID3 Algorithm to build the prediction model. The ID3 Algorithm adopts a greedy (i.e., non-backtracking) approach where decision trees are recursively constructed in a top-down divide-and-conquer manner [13]. It starts with a training set of tuples and associated class labels. The training set is recursively partitioned into smaller subsets as the decision tree is being built. The ID3 decision tree predicts that there 231 patients that develop diabetes and 492 that do not. Of those 231 patients, 33 are wrong, which brings the correct predictions down to 198, having 80% of accuracy.

Consider the attributes that are predicted incorrectly. We begin with the false positives. We already know that according to the data set, diabetes is often associated with increased levels of Plasma-Glucose. In the false positive examples, the patients have a lower level of Plasma-Glucose than expected. Other characteristics are similar to the average for the data set. This hints at the possibility that we may be missing important attributes to classify these particular observations correctly. There might be other risk factors that the data collections did not consider, like level of exercise, cholesterol, etc.

In the case of false negatives, the patients do have characteristics that would go on to develop diabetes, which are high Plasma-Glucose levels and increased BMI. Again, this hints at the possibility that the data set is missing important fields for the classification of those patients. Considering that we have a well-defined set of false positives and negatives, the proper prediction model would need more data gathering.

5. References

[1] Kantardzic, M. (2002) Data Mining: Concepts, Models, Methods, and Algorithms, New Jersey: John Wiley & Sons.

[2] Rapid-I (2008), Interactive Design. Products: RapidMiner, http://rapid.com/content/view/13/69/lang.en/

[3] Wheelwright. J. (2005), Native America's Alleles, Discover Magazine, http://discovermagazine.com/2005/native-americas-alleles

[4] Asuncion, A., Newman, D. J. (2007) Pima Indians Diabetes Data Set, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabets, Irvine, CA: University of California, School of Information and Computer Science.

[5] Larose, D. T. (2006) Data Mining Methods and Models, Hoboken: John Wiley & Sons, Inc.

[6] Pyle, D. (1999) Data Preparation for Data Mining, San Francisco: Morgan Kaufmann.

[7] Cios, K. J., Pedrycz, W., Swiniarski, R.W., Kurgan, L. A. (2007) Data Mining: A Knowledge Discovery Approach, New York: Springer.

[8] Seibel, J. A. (2007) Diabetes Guide, WebMD, http://diabetes.webmd.com/guide/oral-glucose-tolerance-test.

[9] Stein, D. W. (2006) Hypertension / High Blood Pressure Guide, WebMD, http://www.webmd.com/ hypetension-diagnosing-high-blood-pressure

[10] Zelman, K. M. (2008), How Accurate is Body Mass Index, or BMI? WebMD, http://www.webmd.com/diet/features/how-accurate-body-mass-index-bmi.

[11] Kass, G. V. (1980) An Exploratory Technique for Investigating Large Quantities of Categorical Data. Journal of Applied Statistics 29(2): 119-127.

[12] Quinlan, J. R. (1992) C4.5: Programs for Machine Learning, San Francisco: Morgan Kaufmann.

[13] Han, J., Kamber, M. (2006) Data Mining: Concepts and Techniques, 2nd ed. San Francisco: Morgan Kaufman.

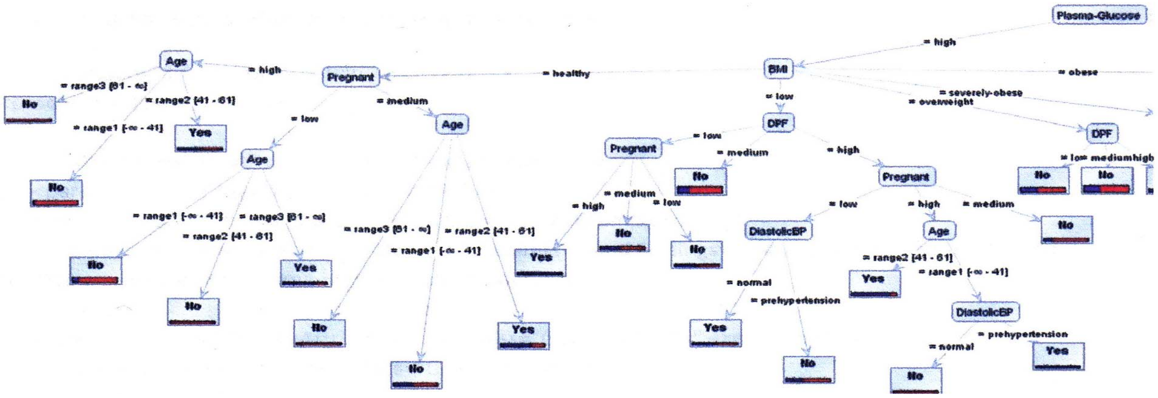

Figure 1: A part of the decision tree