ABSTRACT

Diabetes mellitus, a life-threatening disease of our time was critically studied in this paper with the aim of designing a model that can be used to detect this disease so that people on their own and right in their homes can determine their health status as pertaining to this disease without consulting a physician. The model was developed using Minimum Covariance Determinant (MCD) classifier and it has 1 percent of misclassification error.

Keywords: Diabetes mellitus; Mahalanobis distance; MCD estimator; Fasting blood sugar and non-Fasting blood sugar.

1. INTRODUCTION

Diabetes is usually a lifelong (chronic) disease in which there is high level of sugar in the blood. Symptoms of high blood sugar include frequent urination, increase thirst, and increase hunger. Diabetes is due to either that the body does not produce enough insulin or because body cells do not properly respond to the insulin that is
produced. Insulin is a hormone produced in the pancreas which enables body cells use sugar (glucose) from carbohydrate in the food that one takes for energy or to store glucose for future use. If the body cells do not absorb the glucose, the glucose accumulates in the blood (hyperglycemia) leading to vascular, nerve and other complications. If left untreated, diabetes can cause many complications. Acute complications include diabetic ketoacidosis and nonketotic hyperosmolar coma [1]. Serious long-term complications include cardiovascular disease, stroke, kidney failure, foot ulcers and damage to the eyes [2].

There are three main types of diabetes mellitus (DM)

- Type 1 DM results from the body’s failure to produce enough insulin. This form was previously referred to as “insulin-dependent diabetes mellitus” (IDDM) or “juvenile diabetes.” The cause is unknown [2]
- Type 2 DM begins with insulin resistance, a condition in which cells fail to respond to insulin properly. As the disease progresses, a lack of insulin may also develop [3]. This form was previously referred to as “non insulin-dependent diabetes mellitus” (NIDDM) or “adult-onset diabetes.” The primary cause is excessive body weight and lack of enough exercise [2]
- Gestational diabetes, is the third main form and occurs when pregnant women without a previously history of diabetes develop a high blood glucose level [2].

Other forms of diabetes mellitus include congenital diabetes which is due to genetic defect of insulin secretion, cystic-fibrosis related diabetes, steroid diabetes induced by high doses of glucocorticoids and several forms of monogenic diabetes.

Prevention and treatment involves a healthy diet, physical exercise, not using tobacco and having a normal body weight. Blood pressure control and proper foot care are also important for people with the disease. Type 1 diabetes must be managed with insulin injections [2]. Type 2 diabetes may be treated with medications with or without insulin [4]. Insulin and some oral medications can cause low blood sugar [5]. Weight loss surgery in those with obesity is an effective measure in those with type 2 DM [5].

Gestational diabetes usually resolves after the birth of the baby.

As of 2014, an estimate 387 million people have diabetes worldwide [6] with type 2 diabetes making up about 90% of the cases [7], this is equal to 8.3% of the adult population, with equal rates in both women and men. In the year 2012 to 2014, diabetes is estimated to have resulted in 1.5 to 4.9 million deaths per year [4]. Diabetes at least doubles the risk of death [2]. The number of people with diabetes is expected to rise to 592 million by 2035 [8]. The global economic cost of diabetes in 2014 was estimated to be $612 billion [4]. In the United States; diabetes cost $245 billion in 2012 [9].

Due to the ever increasing prevalence rate of diabetes, this paper aimed at building a classification model using robust discriminant classifier with very high predictive power. With the model, every individual can test his state of health with regards to diabetes right in his home.

2. MINIMUM COVARIANCE DETERMINANT (MCD) ESTIMATOR

The MCD (Minimum Covariance Determinant) based LDF and QDF is given in [10]. The classifier which is based on robust version of the Lawley-Hotelling Test uses the spatial median estimators of [11], and the related scatter of [12] given by [13]. [10] used the re-weighted MCD estimator of multivariate location and scale because of its good statistical properties and FAST MCD algorithm which provides an efficient algorithm for computing estimates for large data set. For the X sample the MCD estimator is defined as the mean \( \hat{\mu}_{x,0} \) and the covariance matrix \( \hat{S}_{x,0} \) of \( n_x \) observations out of \( n_x \) observations whose covariance matrix has the lowest determinant. The quantity should be larger than \( [(n_x-p+1) = 2] \), where \( p \) is the number of the variables and \( n_x-h_x \) should be smaller than the number of outliers in the X population. With this choice the MCD attain its maximum breakdown value \( [(n_x-p+1) = 2] = 50\% \) The breakdown value of an estimator is defined as the largest percentage of contamination it can withstand. If one suspects less than 25% contamination in the X sample, it is advised to use \( h_x \approx 0.75 n_x \) as this yields higher finite sample efficiency. Based on the initial estimates \( \hat{\mu}_{x,0} \) and \( \hat{S}_{x,0} \), one can compute for each observation \( x_i \) its (preliminary) robust distance.
This distance asymptotically follows Chi-squared distribution with \(p\) degrees of freedom at 97.5 percentile value. The weight ‘1’ should be assigned to \(x_i\) if \(RD^0 \leq X^2_{p,0.975}\) and weight ‘0’ otherwise. The re-weighting step increases the finite sample efficiency of the MCD estimator considerably, whereas the breakdown value remains the same. This can be used to flag off outlier and so can be used to detect outliers.

The re-weighted MCD estimator of multivariate location and scale \(\hat{\mu}_{x,MCD}\) and \(\hat{\Sigma}_{x,MCD}\) respectively are used to replace the parameters of Fisher’s linear discriminant function.

\[
D^*(X) = (\mu_x - \mu_y) \sum_{i=1}^{n} \{X - \frac{1}{2}(\mu_x + \mu_y)\} \quad (2)
\]

In fact, \((\mu_x - \mu_y) \sum_{i=1}^{n} \) is a row vector, \(a = (a_1, a_2, \ldots, a_p)\). Thus if \(x_1, x_2, \ldots, x_o\) denote the elements of \(X\),

\[
D^*(X_i) = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_p x_p \quad (3)
\]

where \(a_0 = -\frac{1}{2}(\mu_x - \mu_y) \sum_{i=1}^{n} (\mu_x + \mu_y)\).

Allocate \(X\) to population \(X (\Pi_x)\) if \(D^*(X) > \log_c b\), and otherwise to population \(Y (\Pi_y)\).

\[
\log_c b \quad \text{is always set to be equal to} \quad \log_e
\]

\[
\frac{c_x q_y}{c_y q_x} = \tau \quad (4)
\]

where \(c_x\) is the cost of misclassifying a unit of group \(X\) and \(q_x\) is the prior probability that \(X\) will belong to population \(\Pi_x\). Similar definitions apply for \(c_y\) and \(q_y\). In practice, \(q_x\), \(q_y\), \(c_x\) and \(c_y\) are often not known and therefore we set \(\tau = 0\) throughout this work, under the assumption that the prior probabilities are equal and that the costs of misclassification being equal for the two populations \((c_x = c_y = 1)\), the total probability of misclassification

\[
TPM = P^0 = \frac{1}{2} q_x + \frac{1}{2} q_y \quad (5)
\]

where

\[
q_x = p_{y|x} = p\{D(X) < O / X \sim F_x(x, \theta_i)\}
\]

\(p_{y|x}\) is the probability of misclassifying object of population \(X\) to \(Y\). And

\[
q_y = p_{x|y} = p\{D(X) > O / X \sim F_y(x, \theta_i)\}
\]

\(p_{x|y}\) is the probability of misclassifying object of population \(Y\) to \(X\).

The total cost of misclassification is then

\[
\pi c_1 q_1 + (1 - \pi) c_2 q_2.
\]

\(\pi\) is the prior probability that an observation comes from \(\Pi_x\).

### 3. NUMERICAL APPLICATION

#### 3.1 Source of Data

The Data for this analysis are secondary. The fasting and non-fasting blood sugar level (FBS and NFBS respectively) of diabetics and non-diabetics people coupled with their gender were randomly selected from the cases reported at Amaku General Hospital, Awka.

FBS tests are run very early in the morning on an empty stomach with the patient having had dinner at normal time while NFBS test are run at any time of the day.

The laboratory reference ranges for adults are:

- Glucose (fasting): \(9 - 3.4\) mmol/l
- Glucose (non-fasting): \(4 - 8\) mmol/l
3.2 Calculation of MCD Estimator

Through extensive simulation studies it is observed that \( h \) \( n \) observations out of \( n \) observations whose covariance matrix has the lowest determinant were those observations with smaller Mahalanobis distance [14]. The first thing done here was to calculate the Mahalanobis distance of each datapoint. With the assistance of this distance we were able to get dataset whose estimators attain their maximum breakdown value in a few numbers of iterations. More information on the calculation of the MCD estimators will be seen on [14].

4. RESULTS AND FINDINGS

When the MCD discriminant procedure was applied on real life data of diabetic and non-diabetic peoples we obtain the model for classifying patient into diabetic status as

\[
D^*(X) = 2.1629FBS + 0.0005age + 2.7402NFBS - 33.8286
\]

A person is classified as diabetic patient if this function is obtained positive, otherwise classify as non-diabetic. The model has probability of misclassification of 0.01; this means that out of 100 people studied only one of them was misclassified. The model shows age as not a strong factor that determines person's diabetic status.

5. CONCLUSION

With the predictive power of 99%, the obtained model above can correctly classify patients into their diabetic status. Physicians at Amaku general hospital Awka by this model are advised not to bother on collecting information on age while trying to find the health status of patients as regards Diabetes mellitus. Also once a person knows his FBS level and NFBS level, he can find his health status as regarding diabetes mellitus without consulting Physician. The model is optimal for classification of diabetes mellitus status.

COMPETING INTERESTS

Authors have declared that no competing interests exist.
14. Okonkwo EN, Onyeagu SI, Okeke JU. A comparative performance of several robust discriminant analysis procedures. Journal of the Nigerian statistical Association. 2011;23:57-66.

© 2015 Okeke et al.; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?id=1072&id=5&aid=8965