Finding Biomarker Genes for Type 2 Diabetes Mellitus using Chi-2 Feature Selection Method and Logistic Regression Supervised Learning Algorithm

Hala M Alshamlan

Information Technology Department
College of Computer and Information
King Saud University, Ryiadh, KSA

Summary

Type 2 diabetes mellitus (T2D) is a complex diabetes disease that is caused by high blood sugar, insulin resistance, and a relative lack of insulin. Many studies are trying to predict variant genes that causes this disease by using a sample disease model. In this paper we predict diabetic and normal persons by using fisher score feature selection, chi-2 feature selection and Logistic Regression supervised learning algorithm with best accuracy of 90.23%.

Key words: *single nucleotide polymorphisms (SNP), Type 2 diabetes (T2D), support vector machine (SVM), cross-validation test*

I. Introduction

Diabetes mellitus is one of the most widespread diseases over the globe and on the rise significantly year by year. According to the International Diabetes Federation report [1], there are half a billion people live with diabetes worldwide, and an estimated 212 million people have undiagnosed diabetes. Also, Diabetes considers a major contributor to cardiovascular diseases and ranks eleventh common cause of disability worldwide [1]. In 2016, diabetes mellitus was the direct cause of more than one and a half million deaths [2]. Diabetes mellitus is subdivided into three main types: Type 1 diabetes, Type 2 diabetes, and Gestational diabetes, but type 2 diabetes accounts for the vast majority of diabetes cases.

Type 2 diabetes considers a complex disease that is resulted in a complex interplay between genetic, environmental and lifestyle factors [3]. Diabetes, in general, can be successfully managed and complications prevented when detected early. Since heredity affects the likelihood of getting the disease, early detection and successful recommendation for diagnosis become possible. Moreover, the greatly increased amount of data gathered in medical databases and the availability of historical data on complex diseases, traditional manual analysis has become inadequate and naturally leads to the application of machine learning techniques to discover interesting patterns.

Since the area of Biological Data Mining or Knowledge Discovery in Biological Data is more than ever necessary and important, we will propose a gene-disease prediction function to predict type 2 diabetes mellitus using diabetes gene expression. In order to achieve our goal, we will use Support Vector Machine (SVM) which is one of the most widely used supervised learning algorithms in predicting diseases field. Furthermore, most of the studies that had applied SVM algorithm achieved high accuracy results.

II. Background

Diabetes is a general term of heterogeneous disturbances of metabolism characterized by the presence of the chronic hyperglycemia [4]. Diabetes occurs when the pancreas is no longer able to make insulin (impaired insulin secretion), or when the body cannot take advantage of the produced insulin (impaired insulin action) [5]. Over the past two years, the number of people with diabetes mellitus has arisen 2% globally, making it one of the most important public health challenges to all nations. There are three main types of diabetes: type 1 diabetes, type 2 diabetes, and Gestational diabetes.

Type 1 diabetes is a result of pancreatic beta-cell destruction with consequent insulin deficiency, so the infected body needs a daily insulin injection to maintain blood glucose levels under control. Gestational diabetes

Manuscript received February 5, 2021
Manuscript revised February 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.2.2

(GDM) refers to high blood glucose during pregnancy [4]. GDM usually disappears after pregnancy, but the infected women and their children may develop to type 2 diabetes later in life [5].

Type 2 diabetes is the most common type of diabetes and accounts for around 90% of all diabetes cases [5]. In type 2 diabetes, hyperglycemia is the result of insulin resistance, which means that the body doesn't fully respond to the produced insulin. In some cases, impaired insulin action leads to eventually exhaust the pancreas that resulting in the body producing less insulin and causing higher blood sugar levels. Type 2 diabetes mellitus has a strong link between genetic and environmental risk factors. The genetic information of each cell in our body is stored in chemical form in DNA or RNA [6]. Genes are a sequence stretch of nucleotides that produce proteins that are necessary for normal functioning of the body. Each DNA carries thousands of genes. Single mutation or more in the mitochondrial DNA may cause any type of disease to appear. Therefore, identifying the candidate genes associated with type 2 diabetes, help in predicting the existence of the disease.

Machine learning techniques had been used in order to make the detection or prediction of one disease. In the prediction stage, the supervised learning approach is used by training an algorithm with a labeled dataset of the target disease. The supervised learning algorithms, especially SVM had been employed extensively in gene-disease prediction field [7]. The reason behind it is the high accuracy result achieved by this algorithm compared to other algorithms.

III. Literature Review

In this chapter, an investigation of the published work in the field of the genetic T2D prediction will be conducted. Moreover, it will discuss what algorithms used in each study.

In a recent study in 2017 overviews, machine learning and data mining methods in diabetes research showed that clinical datasets were mainly used in machine learning algorithms for diabetes. From our search there are lots of studies in the field of T2D prediction using family, medical history and other risk factors datasets such: in recent studies in 2019 [8] and in 2017 [9] proposed a T2D Mellitus prediction model based on data mining. In addition, studies in 2014 [10] [11] and in 2015 [12] [13] which used the k-NN, SVM and decision tree classifiers to predict T2D. The data set in prior studies chosen for classification and experimental is based on the Type-2 diabetes Pima Indian Diabetic Set from the University of California, Irvine (UCI) Repository of Machine Learning Databases. Another study in 2016 [14] used data of the 6647 diabetic and nondiabetic people containing 21 independent variables, including demographic characteristics, physical activity, medical history, anthropometric measures, systolic and diastolic blood pressure, etc.

Our focus was on the studies that concern using the gene expression with a classified dataset to predict T2D disease. In 2010 Hyo-Jeong Ban and others propose an identification of Type 2 Diabetes-associated combination of SNPs using SVM [15]. They analyzed the importance of gene-gene interactions in T2D by investigating 408 SNPs in 87 genes in a sample of 462 T2D cases (positive samples) and 456 normal population controls (negative samples). Further, the normal control people do not have a history of diabetes, hypertension, and dyslipidemia. For each SNP, the p-value was calculated using a chi-squared test. They used SVM learning algorithm which learns a classifier from a set of positively and negatively labeled training vectors to discriminate T2D cases against normal. In addition, they employed a feature selection procedure to find the best combination of SNPs and 10-fold cross-validated classification accuracy was adopted in this work. Results showed that different SNP combinations have been identified with the prediction rates of 70.9% and 70.6% of subpopulation men and women data sets.

In 2017 [16] Atul Kumar and others used SVMRFE method which is a modification of SVM that ranks the genes based on their discriminatory power and eliminate the genes which are not involved in causing T2D disease. They predicted the most discriminatory gene target for type II diabetes. Three datasets were collected from GEO and DGAP repository all are 73 samples (36 normal, 34 disease) and genes size is 89133. Overall classification accuracy result was 83.9%.

In 2011 [17], Vasamsetty and others proposed a method for identifying differentially expressed genes causing Type -2 diabetes mellitus using microarray data for diabetes with parental history and healthy. This method focused on identifying multivariate and univariate outliers using Mahalanobis Distance, Minimum Co-variance Determinant (MCD) and other statistical methods. They used Microarray expression analysis for identifying and classifying genes causing (T2DM). Two samples were used in this study: one from diabetic with parental history and the other from healthy and identified 1579 genes which are differentially expressed. As a result, 579 differentially expressed genes were identified out of 39400 genes tested between healthy vs. diabetic with parental history.

In 2014 [18], Hui Liu1 and others provided a new way for understanding the pathogenic mechanism of T2D caused by epigenetic disorders. Analysis of Microarrays (SAM) method was used to identify the gene set (e.g. disease vs. control). Four datasets were used in this study, for T2D DNA methylation data: (GEO, accession number: GSE21232) was used which contained 11 normal samples and 5 T2D samples of T2D in human islet tissue. On the other hand, T2D expression data: (GEO, accession number: GSE38642) was used and contained 63 islet tissue samples, including nine T2D samples and 54 normal samples.

TABLE I. A SUMMARY OF THE PRESENTED STUDIES IN THE LR

Year	Paper reference	Techniques	Datasets	Results
2017	[16]	Fischer score and t-test were used for selection and SVM-RFE for classification	Three datasets were collected from GEO and DGAP (37 normal and 34 T2D cases) all identified 89133 genes	The algorithm classified the genes (with a classification accuracy of 83.9%)
2014	[18]	weighted human DNA methylation network (WMPN), T2D-related subnetwork (TMSN), Analysis of Microarrays (SAM) method	*GSE21232*: 11 normal samples and 5 T2D samples of T2D, *GSE38642*: nine T2D samples and 54 normal samples	It is found that there existed genes with a variant expression level that can affect and promote the development of T2D
2011	[17]	Microarray expression analysis and classification, Lowess Normalization method (LOWESS), Mahalanobis Distance, Minimum Covariance Determinant (MCD)	Two samples: one from diabetic and the other from healthy and identified 1579 genes which are differentially expressed	579 differentially expressed genes were identified out of 39400 genes tested
2010	[15]	(RBF)-kernel SVM, cross-validation test	One dataset of 408 SNPs in 87 genes in a sample of 462 T2D cases and 456 normal controls	Prediction rates of 70.9% and 70.6%

IV. Data Processing and Methods

Data mining is defined as gaining important information from available huge data sets. Analyzing different sets of data can be done using two important data mining techniques: clustering and classification. Classification is a supervised methodology in which assigning the different objects or groups to suitable classes. In this study we use two groups of people as control and case of diabetes disease. Data processing and retrieving information from genetic data requires more accuracy supporting techniques to avoid major problem on human health. That is the main reason of applying data mining techniques for processing diabetes classification problem.

A. Data and Data Pre-processing

We used two datasets: GSE38642 (54 normal vs. 9 disease) [19], and GSE13760 (11 normal vs. 10 disease) from Hematology Department in Roskilde Hospital [20]. These samples have been taken from the Gene Expression Omnibus database (GEO), In this study, all the people of case and control were more than 20 years of age., Table II summarizes the datasets description.

TABLE II. DATASETS DESCRIPTION USED FOR THE STUDY

Dataset name	# of Samples	# of Genes	# of Control samples	# of Case samples	Female	Male	Age
GSE38642	63	18808	54	9	27	36	20-87
GSE13760	21	22277	11	10	-	-	-

Anaconda software with Jupiter package that support Python and R languages was used to implement the proposed methods to select suitable gene subset and then classification is performed on the selected gene subset.

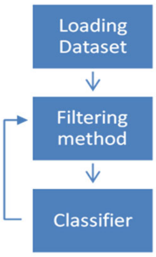

Fig. 1. Development stages

B. Feature selection methods

The most common challenge in bioinformatics is in the process of selecting relevant and removing the redundant genes from the dataset. However, in microarrays the classification is time consuming due to the fact the sample size is very small and huge size of genes. The process of feature selection performed before classification reduces the running time and also increases the accuracy of prediction.

In this study the process of feature selection is performed using the Fisher score and chi-square approaches. The total selected number of genes ranges from 1800-2700. This range is selected based on another study [15] that selects 8% of the total genes in its dataset. The number of relevant genes selected by using the Fisher score and chi-square approaches are shown in Table III below:

TABLE III. NUMBER OF GENES SELECTED USING FISHER SCORE AND CHI2 METHODS

Dataset name	Number of Genes in dataset	Number of genes selected by fisher score and chi ^{2}
GSE38642	18808	[1800-2700]
GSE13760	22277	

C. Classification methods

We used SVM and logistic regression to calculate the classification accuracy and compared them.

- Support vector machine (SVM): the most common classifier used.
- Logistic regression: used to model the probability of certain class the used in biological sciences.

After genes are selected then it subjected to classification methods: logistic regression and SVM classifiers. SVM classification does not provide good results and it shows some bias between classes this may refers to the huge genes with lack balancing between the controls and cases, so we discard this classifier. The classification accuracy is compared on both datasets among the feature selection methods and the classifiers in Table IV V

TABLE IV. THE ACCURACY OF LOGISTIC REGRESSION ON TWO DATASETS BASED ON FISHER SCORE FEATURE SELECTION

k	*GSE38642*	*GSE13760*
2700	88.57%	42.85%
2500	**90.23%**	47.61%
2400	88.57%	47.61%
2300	87.14%	47.61%
2000	85.47%	52.38%
1800	85.47%	**61.90%**

TABLE V. THE ACCURACY OF LOGISTIC REGRESSION ON TWO DATASETS BASED ON CHI-2 FEATURE SELECTION

k	*GSE38642*	*GSE13760*
2700	88.81%	33.33%
2600	87.14%	33.33%
2300	88.81%	33.33%
1800	**88.81%**	**38.09%**

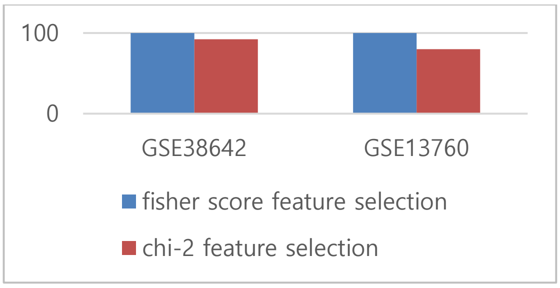

Fig. 2. Classification accuracy comparizon

The classifier accuracy in Table IV above shows that Logistic regression produces the highest accuracy with fisher score for GSE38642 dataset with 90.23% and GSE13760 dataset with 61.90% Fig. 2 shows the average accuracy of the classification.

V. Conclusion

Diabetes mellitus is one of the most widespread complex diseases over the world. We predict diabetic and non-diabetic people by processing two datasets. Feature selection with logistic regression classification were used. The obtained accuracy result of logistic regression on two datasets based on fisher score feature selection was higher than Ch-2 feature selection. The accuracy results of two data were 90.23% and 61.90% respectively.

References

[1] International Diabetes Federation, IDF DIABETES ATLAS, 8th ed. 2017.

[2] “Diabetes.” .

[3] D. J. Hunter, “Gene-environment interactions in human diseases,” Nat. Rev. Genet., vol. 6, no. 4, pp. 287–298, Apr. 2005.

[4] Z. Punthakee, R. Goldenberg, and P. Katz, “Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome,” Can. J. Diabetes, vol. 42, no. Supplement 1, pp. S10–S15, Apr. 2018.

[5] “International Diabetes Federation - What is diabetes.” .

[6] L. Panawala, “Difference Between Gene and Genome,” Feb. 2017.

[7] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research,” Comput. Struct. Biotechnol. J., vol. 15, Jan. 2017.

[8] N. Sneha and T. Gangil, “Analysis of diabetes mellitus for early prediction using optimal features selection,” J. Big Data, 2019.

[9] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, “Informatics in Medicine Unlocked Type 2 diabetes mellitus prediction model based on data mining,” Informatics Med. Unlocked, vol. 10, no. August 2017, pp. 100–107, 2018.

[10] R. C. Anirudha, R. Kannan, and N. Patil, “Genetic Algorithm Based Wrapper Feature Selection on Hybrid Prediction Model for Analysis of High Dimensional Data,” 2014 9th Int. Conf. Ind. Inf. Syst., pp. 1–6.

[11] V. V. V, “Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus,” vol. 95, no. 17, pp. 12–16, 2014.

[12] S. Habibi, M. Ahmadi, and S. Alizadeh, “Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree : Results of Data Mining,” vol. 7, no. 5, pp. 304–310, 2015.

[13] W. Chen, S. Chen, and H. Zhang, “A Hybrid Prediction Model for Type 2 Diabetes Using K-means and Decision Tree,” no. 61272399.

[14] A. Ramezankhani, O. Pournik, and J. Shahrabi, “The Impact of Oversampling with SMOTE on the Performance of 3 Classifiers in Prediction of Type 2 Diabetes,” no. 24, pp. 137–144, 2016.

[15] H. Ban, J. Y. Heo, K. Oh, and K. Park, “Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine,” 2010.

[16] A. Kumar, D. J. Sundara, and S. Singh, “Genomics Data SVMRFE based approach for prediction of most discriminatory gene target for type II diabetes,” Genomics Data, vol. 12, pp. 28–37, 2017.

[17] C. S. Vasamsetty, I. Member, S. R. Peri, A. A. Rao, and K. Srinivas, “Gene Expression Analysis for Type-2 Diabetes Mellitus – A Case Study on Healthy vs Diabetes with Parental History,” vol. 3, no. 3, 2011.

[18] H. Liu et al., “Detection of type 2 diabetes related modules and genes based on epigenetic networks,” vol. 8, no. Suppl 1, pp. 1–16, 2014.

[19] GEO, “GSE38642.” [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38642. [Accessed: 05-Oct-2019].

[20] GEO, “GSE13760.” [Online]. Available: 5-10-2019.