Research and Analysis of Predictive Models for Diabetes Mellitus Using Machine Learning: Active Learning vs. Random Sampling

Chentao Ye

Donald Bren School of Information & Computer Sciences, UC Irvine
Irvine, CA, United States, 92697
498267405@qq.com, chentay2@uci.edu

ABSTRACT

Diabetes Mellitus is one of the most significant chronic metabolic diseases characterized by persistent hyperglycemia affecting human beings in the last few decades. Currently, many research groups are working for building a predictive model to help prevent diabetes. However, the laboratory data and results are usually expensive to collect and are hard to access to. Thus, to reduce the size of the required dataset, this paper analyzes whether active learning can achieve higher effectiveness than random sampling when constructing a predictive model with fewer data to identify people at risk of having Diabetes Mellitus before diagnosis.

Testing accuracy and the area under the receiver operating characteristic curve (AROC) were used to evaluate the discriminatory capability of these models. Through modeling experiments and comparisons of the two measurements, the results show that models of active learning methods can perform better than the ones using random sampling while only teaching them a small-size of the dataset.

CCS CONCEPTS

• Computing methodologies~Machine learning~Learning settings~Active learning settings•Applied computing~Life and medical sciences~Computational biology~Recognition of genes and regulatory elements

KEYWORDS

Machine Learning, Active Learning, Random Sampling, Diabetes Mellitus

1 Introduction

Diabetes mellitus (DM), characterized by persistent hyperglycemia, is one of the most common and well-known

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

ISAIMS 2021, October 29–31, 2021, Beijing, China

© 2021 Copyright is held by the owner/author(s).

ACM ISBN 978-1-4503-9558-8/21/10.

<https://doi.org/10.1145/3500931.3500957>

chronic metabolic diseases in humans. Theoretically, the synergistic effect of chronic hyperglycemia and other metabolic abnormalities in diabetic patients can lead to damages on a great number of organ systems, resulting in disability and even life-threatening health complications.

In 2017, there were 270,702 deaths certificates with diabetes listed as the underlying or contributing cause of death in the US[7]. The death number spiked in 2019 and 2020, which is due to the fact that the presence of diabetes was associated with a 1.8-fold to 2.86-fold increased risk of dying from COVID-19 compared to individuals without diabetes[4]. The DM is a lifelong disorder and has no absolute cure in medicine at present. There are more than 1 out of every 10 Americans of all ages (10.5% of the US population) and more than 1 out of every 4 elders aged 65 years or older (26.8%) diagnosed as DM (The National Diabetes Statistics Report, CDC, 2020) [7].

Notably, if patients are aware of the possibility of the disease before diagnosis and comprehensively take scientific prevention, type-2 and Gestational Diabetes are ideally preventable or delayable. Currently, many research groups are working on building a predictive model to help prevent DM. However, the laboratory data and results required for building the model are usually expensive and are hard to access to. None of the previous researches has taken into account the optimization that active learning methods can bring to the model. Thus, to improve the problem of massive data requirements, this paper analyzed whether active learning would provide a more efficient construction of the predictive model with fewer data than random sampling. The significance of this article lies in that through using a more effective model that requires a smaller size of data, prevention and treatment can be done earlier according to predicted likelihood, effectively cutting down the complications of diabetes and improving the prognosis of the latent patients.

2 Methods

2.1 Dataset Description

The dataset applied in this research was collected from the UCI machine learning repository (originally from the National Institute of Diabetes and Digestive and Kidney Diseases). Diagnostic records of 768 Pima Indian females aged at least 21 years are included in the dataset. Each record consists of eight

medical predictor variables and one target variable. Predictor variables include: Glucose, Body Mass Index (BMI), Age, Insulin, Diabetes Pedigree Function (DPF), Blood Pressure, Pregnancies (number of pregnancies), and Skin Thickness. The target variable, named Outcome, is in the binary form representing the DM diagnosis of the patient.

Firstly, 80% of all patient samples were randomly chosen as the training set and the remaining 20% of data formed the testing set. There are 614 patient records in the training set and 154 records in the testing set. The training set was used to teach the predictive model and the testing set was used to evaluate the effectiveness of the constructed model.

To better understand each feature, the Decision Tree Classifier and Random Forest Classifier from sklearn were used, and a parameter of “feature_importance_” was applied to visualize the importance of each feature to the predicted results (figure 1).

Figure 1. Features' Importance to The Predicted Results

Based on the plot, the amounts of Insulin, DPF, Blood Pressure, Pregnancies, and Skin Thickness appear to play no role in the decision tree (feature importance of 0) and play insignificant roles in the random forest (feature importance less than 0.09). To reduce the model complexity and avoid over-fitting, these features from the dataset were dropped in turn. By using the 5-fold cross-validation method and AROC to observe the effectiveness after deleting specific feature(s), most models perform better with reasonable training accuracy and higher testing accuracy when the two least important features (Pregnancies and Skin Thickness) were dropped (Figure 2). Therefore, Pregnancies and Skin Thickness were excluded in the final analysis dataset with 6 predictor features.

Figure 2. Average Testing Accuracy of 5-Fold Cross-Validation Before/After Dropping The 2 Features

2.2 Model Construction

Many powerful machine-learning algorithms have been applied to construct the predictive model, such as K-Nearest Neighbor (KNN), Random Forest. Among all of them, Support Vector Machine (SVM) and Multilayer Perceptron (MLP) have shown considerable success in a wide range of practical applications and achieved the most obvious results in the DM model (as shown in Figure 2). In this research, the process and results of these two algorithms in model construction are analyzed.

Here is the basic logic of the two Machine-Learning Algorithms:

(a) Support Vector Machine:

SVM is a supervised machine-learning algorithm used for classification and regression problems. Basically, it separates different target classes in a hyperplane in n-dimensional space to create the best decision boundary with maximum margin so that correctly predicting the class of new data.

(b) Multilayer Perceptron:

MLP is a supervised machine-learning algorithm used for classification and regression problems. By training on a dataset, it can achieve a function that can be used to transfer the input data in n-dimension to a m-dimension output. It consists of three types of layers: 1) Input layer receives the input signal to be processed. 2) The arbitrary number of hidden layers acts as the computational engine. 3) Output layer performs the required predictive work of new data.

The optimal hyperparameters of each model were optimized by using the 5-fold cross-validation method, where it had the maximum average area under the receiver operating characteristic curve. Both SVM and MLP models were built on Active Learner from modAL (an active learning framework for Python3). In order to verify the hypothesis proposed at the end of the introduction, several different query strategies were applied and compared on the two ActiveLearner respectively. Due to the fact that there are 614 records in the training dataset, I randomly chose 14 of them as the initial training set and the remaining 600 samples as the pool set allowed for the Active

Learner to query and teach. The models were based on the following operation logic to achieve the experimental purpose:

1) 14 initial training records were fixed to ensure each model can start from the same point. 2) 600 pool records were queried and learned one after another by the Active Learner of pool-based sampling to train the predictive model. 3) AROC and Testing Accuracy values were recorded after every 20 extra samples were taught to visualize the results.

Several of the more classical and popular Query Strategies were applied in the research, and here is the basic logic of them:

(1) Random Sampling:

This is the query strategy used to compare with active learning methods below. When querying for labels based on this measure, the strategy selects the sample randomly from the pool dataset.

(2) Uncertainty Sampling:

The unlabeled example with maximum uncertainty is viewed as the most informative case. When querying for labels based on this measure, the strategy selects the sample with the highest uncertainty. The measure of the uncertainty in the research is defined by:

$$U(\mathbf{x}) = 1 - P(\hat{\mathbf{x}}|\mathbf{x})$$

(where $\mathbf{x}$ is the predictor features and $\hat{\mathbf{x}}$ is the most likely prediction)

(3) Margin Sampling:

When querying for labels based on this measure, the strategy selects the sample with the smallest margin (highest unsure). Classification margin is the difference in probability of the first and second most likely prediction.

In this research, it is defined by:

$$M(\mathbf{x}) = |P(1|\mathbf{x}) - P(0|\mathbf{x})|$$

(where $\mathbf{x}$ is the predictor features and $1/0$ is the binary prediction)

(4) Query By Committee:

The committee allows the ActiveLearner to keep several sampling strategies about the data. When querying for labels based on this measure, the strategy selects the sample with the highest disagreement of the sampling strategies. In this research, both Uncertainty and Margin sampling were used in the Committee.

3. Results

3.1 Active Learning Methods Performed Better With Small-Portion of Data in SVM

The optimal set of hyperparameters obtained for the SVM model in this research is as follows: the regularization parameter (C) is 1, the kernel type (kernel) is "linear", the probability estimates (probability) is True, the training and testing x values are scaled by using the StandardScaler from sklearn.preprocessing. Through applying the 5-fold cross-validation method to verify this predictive model, the average testing accuracy of 77.61% and average AROC values of 87.6% were obtained, indicating high reliability of the model with the set of hyperparameters.

After the train-test split method was adopted, , the testing accuracy of the model trained with the initial 14 random samples was 59.09% (Figure 3). By querying and teaching the model with 20 extra samples, the active learning model of Margin Sampling can be improved significantly to an accuracy of 72.72% (13.63% of the increase), while Random Sampling can only achieve the testing accuracy of 59.98% (increase less than 1%). For the Margin Sampling model, the testing accuracy has exceeded 77% when given only about 15% of the sample data (~ 92 records). The same goal of testing accuracy in Random Sampling, however, it can only be achieved after learning about 35% of the sample data (~215 records). Although Random Sampling slightly overtakes Margin Sampling in the last 30% of the data, it can be seen clearly that the Margin Sampling has obvious advantages when only teaching them with the first 70% of the data, even reaching a gap of 12.74% at the maximum.

Another active learning method of Uncertainty Sampling was also applied to demonstrate the effectiveness of these models (Figure 4). Although the initial advantage is not as pronounced as Margin Sampling, it is easy to see from the chart that the accuracy of Uncertainty Sampling in the first 70 percent of data is much higher than that of Random Sampling.

Figure 3. Testing Accuracy of Active Learning and Random Sampling in SVM

Figure 4. Testing Accuracy of Active Learning and Random Sampling in SVM

A similar difference also appears in AROC values. After 34 samples are queried and taught, the AROC value of Random Sampling is 68% while the value of Margin Sampling has achieved 75% (Figure 5). Teaching 20 extra samples to each model, the AROC value of Margin Sampling is now dramatically increased to 90% while the one of Random Sampling can only get 81% (Figure 6).

Figure 5. ROC Curve of Active Learning and Random Sampling in SVM with 34 Data Samples

Figure 6. ROC Curve of Active Learning and Random Sampling in SVM with 54 Data Samples

3.2 Active Learning Methods Performed Better With Small-Portion of Data in MLP

The optimal set of hyperparameters for the MLP model is as follows: the Maximum number of iterations (max_iter) is 1000, the L2 penalty parameter (alpha) is 1, the random_state is 0, the training and testing x values are scaled. Through applying the 5-fold cross-validation method, the average testing accuracy of 77.09% and average AROC values of 89.3% were obtained, indicating a high reliability of the model with the set of hyperparameters.

Using the train-test split method with the ratio of 8:2, the testing accuracy of the model trained with the initial 14 random samples was 73.37% (Figure 7). By querying and teaching the model with 40 extra samples, active learning model of QueryByCommittee can be improved significantly to an accuracy of 79.81% (6.44% of the increase), while Random Sampling can

only achieve the testing accuracy of 74.62% (increase less than 1.5%). For the QueryByCommittee model, the testing accuracy has exceeded 80% when given only about 15% of the sample data (~92 records). The same goal of testing accuracy in Random Sampling, however, it can only be achieved after learning about 55% of the sample data (~338 records). Although Random Sampling catches up QueryByCommittee in the last 35% of the data, it can be clearly seen that QueryByCommittee has obvious advantages when only teaching them with the first 65% of the data, reaching the gap of 5.19% at the maximum.

To validate the previous results, the active learning method of Uncertainty Sampling was also applied to demonstrate the effectiveness of these models (Figure 8). Based on the plot, a very similar developing pattern of the predictive model is observed in UncertaintySampling.

Figure 7. Testing Accuracy of Active Learning and Random Sampling in MLP

Figure 8. Testing Accuracy of Active Learning and Random Sampling in MLP

Taking a further insight at AROC values. After 34 samples are queried and taught, the AROC value of Random Sampling is 74% while the value of QueryByCommittee has achieved 85% (Figure 9). Teaching 20 extra samples to each model, the AROC value of

QueryByCommittee is 84% while the one of Random Sampling can only get 80% at the same point (Figure 10).

Figure 9. ROC Curve of Active Learning and Random Sampling in MLP with 34 Data Samples

Figure 10. ROC Curve of Active Learning and Random Sampling in MLP with 54 Data Samples

4. Conclusion

Through the analysis of this paper, it is concluded that the active learning methods can provide more efficient construction on the DM predictive model with higher testing accuracy and AROC values when given a small-portion or small-size data than those applying random sampling.

By taking advantage of the feature of active learning that always learn the most informative sample first, the development of the predictive mechanism for diabetes is highly likely to be enhanced and accelerated by improving the problem of difficult access to laboratory data in the reality. The DM risk prediction system, indeed, can be built online using a computer program to help doctors or patients themselves to access the risk of developing diabetes before diagnosis, and accordingly, give scientific prevention or treatment in time.

However, SVM and MLP models are shown to be valid in the research but may not be the most effective among all machine learning algorithms. More attention should be paid to the technique discovery of the DM risk prediction model with optimized parameters. Also, the query strategies applied in the

research made few adjustments to the specific model. An interesting direction for further development of this model is to construct a set of query strategies tailored to the DM risk prediction system.

ACKNOWLEDGMENTS

I would like to thank professor Robert F. Murphy and teacher assistant Jinzhe Zhang for their instruction and support. Also, I would like to acknowledge Donald Bren School of Computer Science, UC Irvine, for the generosity in providing the dataset used in this research. I also thank the Editor and reviewers for their helpful comments, which have helped us to improve the manuscript.

REFERENCES

- [1] Muhammad, L.J., Algehyne, E.A. & Usman, S.S. Predictive Supervised Machine Learning Models for Diabetes Mellitus. SN COMPUT. SCI. 1, 240 (2020). Available: <https://doi.org/10.1007/s42979-020-00250-8>
- [2] Sneha, N., Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data 6, 13 (2019). Available: <https://doi.org/10.1186/s40537-019-0175-6>
- [3] Lai, H., Huang, H., Keshavjee, K. et al. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocr Disord 19, 101 (2019). Available: <https://doi.org/10.1186/s12902-019-0436-6>
- [4] Unnikrishnan, R., Misra, A. Diabetes and COVID19: a bidirectional relationship. Eur J Clin Nutr (2021). Available: <https://doi.org/10.1038/s41430-021-00961-y>
- [5] Yu, W., Liu, T., Valdez, R. et al. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak 10, 16 (2010). Available: <https://doi.org/10.1186/1472-6947-10-16>
- [6] Alian, Shadi, Juan Li, and Vikram Pandey. "A personalized recommendation system to support diabetes self-management for American Indians." IEEE Access 6 (2018): 73041-73051. Available: <https://ieeexplore.ieee.org/abstract/document/8539994>
- [7] National Diabetes Statistics Report, 2020. Centers for Disease Control and Prevention, CDC (2020). Available: <https://www.cdc.gov/diabetes/library/features/diabetes-stat-report.html>
- [8] Goyal, R, Jialal, I. Diabetes Mellitus Type 2. National Center for Biotechnology Information, NCBI (2021). Available: <https://www.ncbi.nlm.nih.gov/books/NBK513253/>.