EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Support Vector Machine based eHealth Cloud System for Diabetes Classification

Chandrashekhar Azad[1,*], Ashok Kumar Mehta[1], Dindayal Mahto[2], Dharmveer Kumar Yadav[3]

[1]Departement of Computer Applications, National Institute of Technology Jamshedpur, India
[2]School of Computing, SASTRA Deemed University, Thanjavur, India
[3]Departement of Computer Science & Engineering, Katihar Engineering College, Katihar, India

Abstract

INTRODUCTION: Diabetes is a major health issue because it leaves people with physical disabilities. Therefore, methodologies with a reduced error rate must be used to diagnose this dangerous disease. Data Mining techniques such as Artificial Neural Network are common tools adopted for the classification of diabetes and one of the core components of the eHealth system. Data Mining techniques aim to provide reliable and timely diagnostic outcomes during the diagnosis of the disease.

OBJECTIVE: The objective of the research work is to propose a Support Vector Machine based eHealth Cloud System for Diabetes Classification. This work aims to improve the diagnostic accuracy of computer-assisted diagnostic systems.

METHOD: The proposed methodology implemented in two-phase, In the first phase system is trained using different Support Vector Machine (SVM) kernel functions and in the second phase effectiveness of the system is tested in terms of classification accuracy and error. Different SVMs have the ability to diagnose this disease. PIMA Indian Diabetes Dataset (PIDD) has been used in our experiments for training and testing. Kernel functions are usually used to refer to the kernel trick, a method of using a linear classifier to solve a non-linear problem.

RESULT: In this classification accuracy and classification error are used for performance evaluation. It is worth mentioning that the system giving remarkable accuracy of 77.50% in Coarse Gaussian SVM in 10-fold validation whereas fine Gaussian SVM gives 98.8% accuracy in No validation set.

CONCLUSION: This paper introduces the SVM eHealth Cloud System for Diabetes Classification. The system is trained using the PIDD. Such a system can be used as “Application-as-a-Service” in cloud computing. It is therefore believed that the system will enhance the process of clinical decision-making and also assist physicians concerning Diabetes Diagnosis. It is worth to mention that the SVM kernel-based system performed well in comparison to the different systems.

Keywords: Data Mining, eHealth, Medical Mining, PIDD, Support Vector Machine.

Received on 29 November 2019, accepted on 14 May 2020, published on 20 May 2020.

doi: 10.4108/_____

1. Introduction

People are susceptible to different health issues based on their lifestyle and work behaviours. Some health issues are avoided and reduce it with changing diet, way of living, and atmospheres. Lifestyle health issues describe a certain

*Corresponding author. Email: csazad.ca@nitjsr.ac.in

Table 6. Configuration &complexity in No Validation

Parameter	Linear SVM	Quadratic SVM	Cubic SVM	Fine Gaussian SVM	Medium Gaussian SVM	Coarse Gaussian SVM
Prediction Speed (obs/sec)	~17000	~19000	~21000	~14000	~18000	~19000
Training Time	8.0958	7.2740	7.1833	7.0803	6.9617	6.5533
Kernel Function	Linear	Quadratic	Cubic	Gaussian	Gaussian	Gaussian
Kernel Scale	Automatic	Automatic	Automatic	0.71	2.8000	1

5.3 K-Fold Validation: **Defend in contrast to overfitting by dividing the data into the different folds and evaluating accuracy on each fold.**

The process is as follows:

- Randomly shuffle the set of data.
- Divide the data set into K folds.
- repeat K-times:
  - Take 1-fold as a testing set.
  - Take the remaining fold as a training set
  - Train a model on the training set and assess it on the testing set.
  - Retain the performance score and discard the model
- Summarize the model's performance using the list of model's performance scores.

Table 7. K-Fold validation

Classifier Type	Accuracy	Error
10-Fold		
Linear SVM	77.2	22.80
Quadratic SVM	75.5	24.50
Cubic SVM	72.5	27.50
Fine Gaussian SVM	65.4	34.60
Medium Gaussian SVM	77.1	22.90
Coarse Gaussian SVM	77.5	22.50
8-Fold		
Linear SVM	76.7	23.30
Quadratic SVM	75.9	24.10
Cubic SVM	72.5	27.50
Fine Gaussian SVM	65.5	34.50
Medium Gaussian SVM	76.4	23.60
Coarse Gaussian SVM	76.7	23.3
5-Fold		
Linear SVM	76.8	23.20
Quadratic SVM	75.1	24.90
Cubic SVM	71.1	28.90
Fine Gaussian SVM	65.8	34.20

Table 8. Configuration &complexity in K-fold

Parameter	Linear SVM	Quadratic SVM	Cubic SVM	Fine Gaussian SVM	Medium Gaussian SVM	Coarse Gaussian SVM
10-Fold						
Prediction Speed (obs/sec)	~10000	~13000	~13000	~9900	~10000	~9900
Training Time(Sec)	10.85	11.068	14.459	10.539	10.495	10.283
Kernel Function	Linear	Quadratic	Cubic	Gaussian	Gaussian	Gaussian
Kernel Scale	Automatic	Automatic	Automatic	0.71	2.8000	1
8-Fold						
Prediction Speed (obs/sec)	~19000	~30000	~39000	~16000	~19000	~18000
Training Time(Sec)	1.344	1.9993	4.4785	1.0344	0.98975	0.91754
Kernel Function	Linear	Quadratic	Cubic	Gaussian	Gaussian	Gaussian
Kernel Scale	Automatic	Automatic	Automatic	0.71	2.8000	1
5-Fold						
Prediction Speed (obs/sec)	~48000	~51000	~63000	~35000	~34000	~33000
Training Time(Sec)	1.3936	1.0935	2.4907	1.029	0.98425	0.93487
Kernel Function	Linear	Quadratic	Cubic	Gaussian	Gaussian	Gaussian
Kernel Scale	Automatic	Automatic	Automatic	0.71	2.8000	1

Table 9. Performance Comparison

Algorithm	Accuracy	Error
K-fold and hold out validation*		
Linear SVM	77.40	22.60
Quadratic SVM	77.10	22.90
Cubic SVM	72.50	27.50
Fine Gaussian SVM	65.4	34.60
Medium Gaussian SVM	77.1	22.90
Coarse Gaussian SVM	77.5	22.50
No validation*		
Linear SVM	77.3	22.70
Quadratic SVM	80.5	19.50
Cubic SVM	87.5	12.50
Fine Gaussian SVM	98.8	01.20
Medium Gaussian SVM	82.7	17.30
Coarse Gaussian SVM	78.4	21.60
From Literature		
Sim [15]	75.29	24.71
Sim+F1 [15]	75.84	24.16
Sim+F2 [15]	75.97	24.03
FMM[16]	69.28	30.72
FMM-CART[16]	71.35	28.65
FMM-CART-RF[16]	78.39	21.61
Binary – coded GA[17]	74.80	25.20
BP[17]	73.80	26.20
Binary-coded GA[17]	77.60	22.40

*Our Trained System

5.4 Discussion

Our SVM Kernel-based expert system for Diabetes classification is simulated in MATLAB, Windows 10, 8 GB RAM, Intel Core i7-8700 CPU @3.20 GHz processor. The proposed system is trained using different SVM Kernel functions using PIDD. The system is trained with hold out validation, no validation, and different k-fold validation methods and performance and recorded in table 3-9. Table 1 shows the comparison of different kernel functions, table 2 shows the statistics of the dataset. Table 4,6, and 8 shows the configuration and complexity like prediction speed, Training Time, Kernel scale, Kernel function. Table 3 shows the performance of various kernel functions with hold out validation (80-20 and 70-30). Table 5 shows the performance of various kernel functions with No validation. Table 7 shows the performance of various kernel functions with K-fold cross-validation (K=10,8 5). In no validation set the best performance is observed on fine gaussian SVM and the values of various performance measures such as accuracy and error are 98.8 % and 01.20 respectively.

In Holdout validation set the best performance is observed on Liner SVM and the values of various performance measures such as accuracy and error are 77.40% and 22.60% respectively.

In K-Fold validation best performance is observed on Coarse Gaussian SVM in 10 fold validation and the values of various performance measures such as accuracy and error are 77.50% and 22.50% respectively.

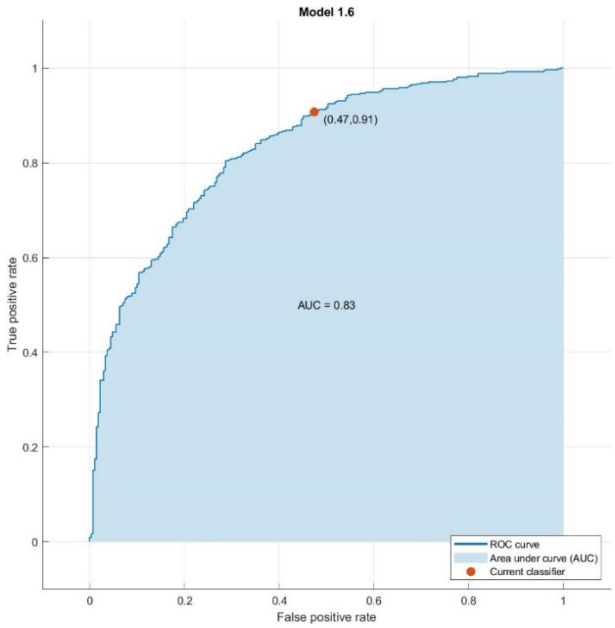

Figure 8. ROC of Coarse gaussian SVM

Figure 8. shows the ROC of Coarse Gaussian SVM, it is observed that the AUC is 0.83 in 10-fold validation.

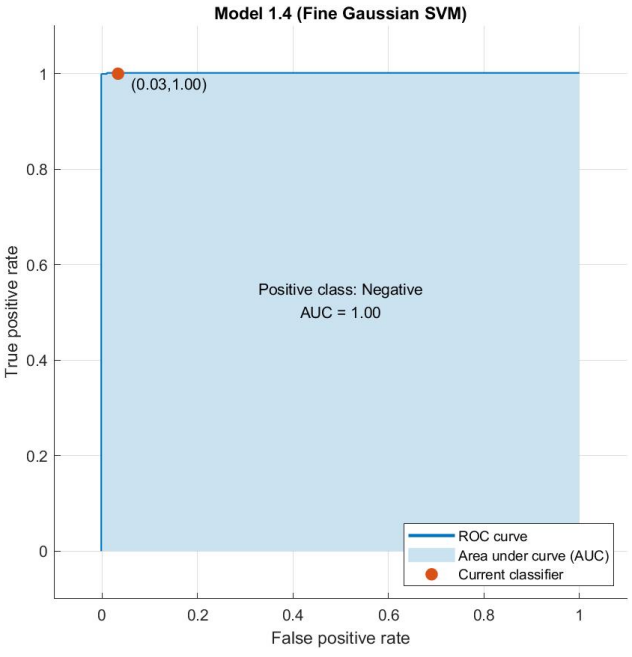

Figure 9. ROC of Fine gaussian SVM

Figure 9. shows the ROC of Fine Gaussian SVM, it is observed that the AUC is 1.00 in no validation, in this case, accuracy is very high but the system may have a problem of overfitting.

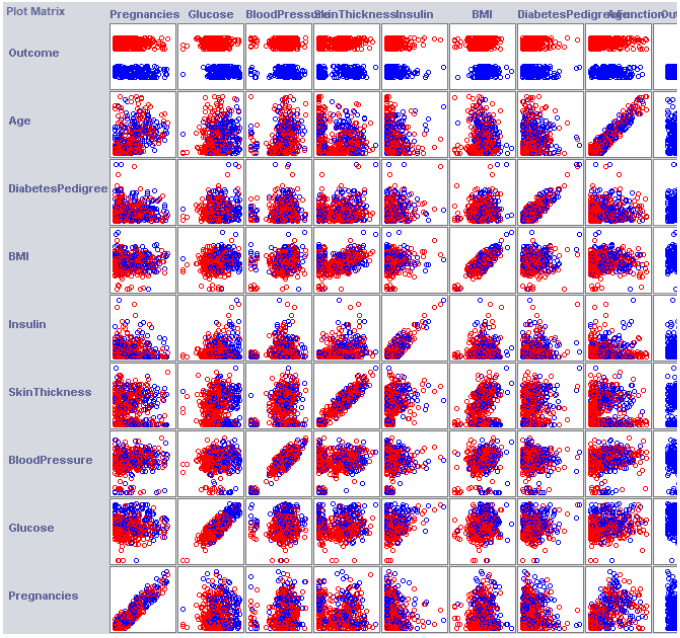

Figure 10. Visualization of Dataset

6. Conclusion

ICT advances lead to the use of data mining techniques in different fields, including medical sciences. By using data mining techniques, we can design and implement complex medical processes. In this, we proposed the SVM eHealth Cloud system for diabetes prediction. The system is trained using the PIDD. Such a system can be used as "Application-as-a-Service" in cloud computing, also, it is useful for various medical science fields such as diagnosis and assisting surgeons, doctors, etc. It is worth to mention that the system giving remarkable accuracy 77.50% by coarse gaussian SVM in 10-fold validation and fine gaussian SVM gives 98.8% accuracy in No validation set. In the future, we will use some optimization techniques like GA, PSO, ACO, etc. along with the machine learning algorithm. It is worth to mention that such a machine learning-based eHealth system may be used for the early prediction of diabetes. Such a system enables health institutions to modernize many of their procedures and more efficiently and cost-effectively deliver services.

References

[1] Sharma M, Majumdar PK. Occupational lifestyle diseases: An emerging issue. Indian journal of occupational and environmental medicine. 2009 Dec;13(3):109.

[2] Sobra J, Ceska R. Diseases of civilization from the aspect of the evolution of the human diet. Casopis learn ceskych. 1992 Apr;131(7):193-7.

[3] Key TJ, Allen NE, Spencer EA. The effect of diet on risk of cancer. Lancet. 2002;360:861–8.

[4] Diabetes,https://www.who.int/en/news-room/fact-sheets/detail/diabetes, Accessed on 27/11/2019.

[5] Tripathy BB, Chandalia HB, Das AK. RSSDI textbook of diabetes mellitus. JP Medical Ltd; 2012 Jan 15.

[6] Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE. Medical data mining: knowledge discovery in a clinical data warehouse. InProceedings of the AMIA annual fall symposium 1997 (p. 101). American Medical Informatics Association.

[7] Qasem SN, Shamsuddin SM. Radial basis function network based on time variant multi-objective particle swarm optimization for medical disease diagnosis. Applied Soft Computing. 2011 Jan 1;11(1):1427-38.

[8] Aslam MW, Zhu Z, Nandi AK. Feature generation using genetic programming with comparative partner selection for diabetes classification. Expert Systems with Applications. 2013 Oct 1;40(13):5402-12.

[9] Barakat N, Bradley AP, Barakat MN. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE transactions on information technology in biomedicine. 2010 Jan 12;14(4):1114-20.

[10] Dogantekin E, Dogantekin A, Avci D, Avci L. An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS. Digital Signal Processing. 2010 Jul 1;20(4):1248-55.

[11] Ephzibah EP. Cost effective approach on feature selection using genetic algorithms and fuzzy logic for diabetes diagnosis. arXiv preprint arXiv:1103.0087. 2011 Mar 1.

[12] Ganji MF, Abadeh MS. A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Systems with Applications. 2011 Nov 1;38(12):14650-9.

[13] Ganji MF, Abadeh MS. Using fuzzy ant colony optimization for diagnosis of diabetes disease. In2010 18th Iranian Conference on Electrical Engineering 2010 May 11 (pp. 501-505). IEEE.

[14] Kahramanli H, Allahverdi N. Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications. 2008 Jul 1;35(1-2):82-9.

[15] Luukka P. Feature selection using fuzzy entropy measures with similarity classifier. Expert Systems with Applications. 2011 Apr 1;38(4):4600-7.

[16] Seera M, Lim CP. A hybrid intelligent system for medical data classification. Expert Systems with Applications. 2014 Apr 1;41(5):2239-49.

[17] Örkcü HH, Bal H. Comparing performances of backpropagation and genetic algorithms in the data classification. Expert systems with applications. 2011 Apr 1;38(4):3703-9.

[18] https://in.mathworks.com/help/stats/choose-a-classifier.html

[19] Azad C, Jain S, Jha VK. Design and Analysis of Data Mining Based Prediction Model forParkinson's disease. Issues.;1(1):181-9.

[20] Azad C., Jha VK. Fuzzy min–max neural network and particle swarm optimization based intrusion detection system. Microsystem Technologies. 2017 Apr 1;23(4):907-18.

[21] PIDD: PIMA Indian Diabetes Dataset, https://www.kaggle.com/uciml/pima-indians-diabetes-database
[22] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia computer science. 2018 Jan 1;132:1578-85.
[23] Jangiti S, VS SS. Scalable and direct vector bin-packing heuristic based on residual resource ratios for virtual machine placement in cloud data centers. Computers & Electrical Engineering. 2018 May 1;68:44-61.
[24] Jangiti S, Ram ES, Sriram VS. Aggregated Rank in First-Fit-Decreasing for Green Cloud Computing. In Cognitive Informatics and Soft Computing 2019 (pp. 545-555). Springer, Singapore.
[25] Azad C, V. K. Jha. Fuzzy min–max neural network and particle swarm optimization based intrusion detection system. Microsystem Technologies. 2017 Apr 1;23(4):907-18.
[26] Azad C., V. K. Jha. Data mining in intrusion detection: a comparative study of methods, types and data sets. International Journal of Information Technology and Computer Science (IJITCS). 2013 Jul 1;5(8):75-90.
[27] Azad C., V. K. Jha. A novel fuzzy min-max neural network and genetic algorithm-based intrusion detection system. InProceedings of the Second International Conference on Computer and Communication Technologies 2016 (pp. 429-439). Springer, New Delhi.