Comparative Analysis of Prediction Accuracy of General and Personalized Datasets Based Classification Model for Medical Domain

Omprakash Chandrakar
Associate Professor (Computer Science), Uka Tarsadia University, Bardoli, India
Email: opchandrakar@utu.ac.in
Dr. Jatinderkumar R. Saini
Director I/C & Associate Professor, Narmada College of Computer Application, Bharuch, Gujarat, India
Email: saini_expert@yahoo.com

----------ABSTRACT----------

With the advancement of Information and Communication Technology, electronic health services have been gaining importance. Because of the availability and accessibility of patient's electronic health records, various expert systems have been developed and being used with limited success from the diagnosis to risk assessment to treatment. This type of expert systems heavily depends on the patient's records collected globally. Health record of the patient, who is to be diagnosed or treated, is not taken into consideration. The authors envisage that if the patient's individual health records are given due weightage in building models, the actuary of such system will increase. This paper presents a comparative study of the effectiveness of classification techniques for predicting blood glucose level in both of the scenario, when the patient's own records are considered and not considered to build classification model.

Keywords - **Classification, Data mining, Diabetes, e-health, Personalized Dataset.**

I. INTRODUCTION :

Ith the advancement of Information and Communication Technology, electronic health services are gaining importance. Because of the availability and accessibility of patient's electronic health records, various expert systems have been developed and being used with limited success from the diagnosis to risk assessment to treatment.

The accuracy of such data mining based expert systems is not very high. This type of system uses historical data of the patients scattered around the country and even several times around the world. Authors are of the opinion that the use of global data in building the data mining model might be one of the reasons for less accurate prediction. This paper presents an experimental study to show the effectiveness of usage of patient's own medical records for building data model to predict blood sugar level in diabetic patient.

II. Significance of the problem

Diabetes is very fast growing disease in India and worldwide. Centers for Disease Control and Prevention, U.S., has projected in its press release that one third of the adult population in U.S. could have diabetes by 2050 if current trends continue [12]. There is no cure available for diabetes in Modern Allopathic System of Medicine. Modern treatment and medication only help in managing the blood glucose level to a certain level.

Although the role of the medical practitioner is important, unlike the other diseases like typhoid or cancer etc., the patients and their family members play a very crucial role in its treatment. So it becomes very important for the patient to keep their health record and keep themselves aware of their present health status.

Several research works have been reported that diagnose the diabetes with limited accuracy. The system uses various data mining techniques to build model. The system build a model based on the historical records. This model, than used, to predict whether a person is diabetic or not.

III. Literature Review:

After going through the published work, we have observed that:

1. All the methods discussed in the papers are heavily depends on the other patients records.
2. No method take individual patient's past records into consideration while assessing the health status and predicting any future problems.
3. All the systems are meant to be used by the experts, not for the patients.

Table 1: Literature Review

Ref.	Issues Discussed	Methods/Techniques Used
[1]	To improve the diagnostic accuracy of diabetes disease combining PCA and ANFIS	Principal component analysis and ANFIS
[2]	Diagnosis and classification of diabetes	Generalized discriminant analysis and least square support vector machine
[3-8]	Clustering/classification	CART/ classification tree
[9]	Clustering the leader genes and determining interactions among them with k-means algorithm	k-means
[10]	Feature selection for diabetes type 2	Naïve bayes
[11]	Heart Disease Prediction/ Heart Disease Diagnosis	Multi-Layer Neural Network - Backprpogation
[12]	Intelligent Heart Disease Prediction System (IHDPS)	Decision Trees, Naïve Bayes and Neural Network.
[13]	Decision Support System for diagnosis of Congenital Heart Disease	The Backpropagation Neural Networktrained by a supervised Delta Learning Rule.
[14]	Prediction of heart disease, blood pressure and sugar	Backpropogation algorithm
[15]	Diagnosis of heart and diabetes diseases	Modified K-means Algorithm is used for clustering based data preparation system for the elimination of noisy and inconsistent data and Support Vector Machines is used for classification

IV. Drawback of the current approach:

To build the model, the present system heavily depends on the historical records of the thousands of patients. It does not take individual patient's past records into consideration while assessing the health status and predicting any future problems.

Diabetes is highly personalized disease. There are various contributing factors of diabetes. Some of the significant factors are:

Age, Gender, Race, Duration of diabetes, Personal history, Family history, Hemoglobin level, Blood Pressure, Total cholesterol, Low Density Lipoprotein, High Density Lipoprotein, Body Mass Index, Life style, Stress level, Diet, Obesity, Metabolism, Physical activity, Insulin production, Resistance to insulin etc.

And the most crucial (and hard part) is that effect of change in various parameters to blood glucose highly varies from person to person. Means if we make similar changes in the diet of the two patients, keeping all other contributing factors same, their change in blood glucose level varies from person to person. It highly depends on the person's metabolism. Even the effectiveness of the drug varies from person to person, unlike other disease where the effectiveness of a particular drug does not varies much from person to person. So it is not very realistic and accurate to predict the health status of the patient based on the records of the other patients.

V.Experiments and Result

A conclusion section must be included and should indicate clearly the advantages, limitations, and possible applications of the paper. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

The authors envisage that if the individual patient's medical record is given due weightage in building model, the actuary of such system will increase. This paper presents a comparative study of the effectiveness of classification technique in both of the scenario, for the prediction of patient's blood sugar level, when the patient's own records are considered and not considered to build classification model.

Hypothesis:

If Individual Patient's Past Health Records are taken into consideration and given due importance, the prediction will be more accurate.

Testing the Hypothesis:

1. Prepared 16 data sets each corresponding to one patient. [D1, D2…D16]
2. Built and Test Model to predict Blood Glucose Level [Normal, High] for every patient.
3. Built a combined data set consisting of all 16 patients records. [D]
4. Built and Test Model to predict Blood Glucose Level [Normal, High] for every patient.
5. Results are compared.

Dataset Description:

The data set is taken from UCI Machine learning repository and provided by Michael Kahn, MD, PhD, Washington University, St. Louis, MO.

Table 2: Dataset Description

Data Set Characteristics:	Multivariate, Time-Series	Number of Instances:	N/A	Area:	Life
Attribute Characteristics:	Categorical, Integer	Number of Attributes:	20	Date Donated	N/A
Associated Tasks:	N/A	Missing Values?	N/A	Number of Web Hits:	84614
Data Set Characteristics:	Multivariate, Time-Series	Number of Instances:	N/A	Area:	Life
Attribute Characteristics:	Categorical, Integer	Number of Attributes:	20	Date Donated	N/A
Associated Tasks:	N/A	Missing Values?	N/A	Number of Web Hits:	84614

The database contains data of Outpatient Monitoring and Management of Insulin Dependent Diabetes Mellitus (IDDM). It is provided for the AAAI Spring Symposium on Interpreting Clinical Data.

The data set consists of 70 files. Each file contains approximately 1000 records of one patient. So total we have more than 70 thousand records of 70 patients. Diabetes files consist of four fields per record. Each field is separated by a tab and each record is separated by a newline.

File Names and format:

(1) Date in MM-DD-YYYY format
(2) Time in XX:YY format
(3) Code
(4) Value

The Code field is deciphered as follows:

33 = Regular insulin dose
34 = NPH insulin dose
35 = UltraLente insulin dose
48 = Unspecified blood glucose measurement
57 = Unspecified blood glucose measurement
58 = Pre-breakfast blood glucose measurement
59 = Post-breakfast blood glucose measurement
60 = Pre-lunch blood glucose measurement
61 = Post-lunch blood glucose measurement
62 = Pre-supper blood glucose measurement
63 = Post-supper blood glucose measurement
64 = Pre-snack blood glucose measurement
65 = Hypoglycemic symptoms
66 = Typical meal ingestion
67 = More-than-usual meal ingestion
68 = Less-than-usual meal ingestion
69 = Typical exercise activity
70 = More-than-usual exercise activity
71 = Less-than-usual exercise activity
72 = Unspecified special event

Processed Dataset:

```
@relation DATA_PID_01

@attribute Reg_Insulin-BF {G,N}
@attribute Reg_Insulin-BF-LR {G,N}
@attribute Reg_Insulin-BF-DR {G,N}
@attribute Reg_Insulin-BF-BR {N,G}
@attribute NPH_Insulin-BF {G,N}
@attribute NPH_Insulin-LR {N,G}
@attribute NPH_Insulin-DR {N,G}
@attribute NPH_Insulin-BR {G,N}
@attribute U_Insulin-BF {G,N}
@attribute U_Insulin-LR {N,G}
@attribute U_Insulin-DR {G,N}
@attribute U_Insulin-BR {N,G}
@attribute Hypoglycemic_sy {NP,P}
@attribute Typical_Meal {NP,P}
@attribute More-than-usual_Meal {NP,P}
@attribute Less-than-usual_Meal {NP,P}
@attribute Typical_Exercise {NP,P}
@attribute More-than-usual_Exercise {NP,P}
@attribute Less-than-usual_Exercise {NP,P}
@attribute Unspecified_Event {NP,P}
@attribute Class {N,H}

@data
G,G,G,N,G,N,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,N
G,G,G,N,G,N,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,H
G,N,G,N,G,N,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,H
G,G,N,G,G,N,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,H
G,G,G,G,G,N,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,H
G,G,G,N,G,N,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,H
N,G,G,N,N,G,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,H
G,N,G,N,G,N,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,N
G,G,G,N,G,N,N,?,?,?,?,?,?,NP,?,?,?,?,?,?,?,H
G,G,G,N,G,N,N,?,?,?,?,?,?,NP,?,?,?,?,?,?,?,N
G,G,G,N,G,N,G,?,?,?,?,?,?,NP,?,?,?,?,?,?,?,N
G,G,G,N,G,N,N,?,?,?,?,?,?,NP,?,?,?,?,?,?,?,H
G,G,G,N,G,N,N,?,?,?,?,?,?,NP,?,?,?,?,?,?,?,N
G,G,G,N,G,N,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,N
G,G,G,G,G,N,N,?,?,?,?,?,?,NP,?,?,?,?,?,?,?,H
G,G,G,N,G,N,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,H
G,G,G,N,G,N,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,N
G,G,G,N,G,N,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,N
G,G,G,N,G,N,N,?,?,?,?,?,?,?,?,?,?,?,?,?,?,N
G,G,G,N,G,N,N,?,?,?,?,?,?,NP,?,?,?,?,?,?,?,H
```

Results:

Table 3: Predication Accuracy

Exp. No	Model Built with	Test Data Set	Prediction		
				Incorrect	
			Accuracy (%)	FALSE POSITIVE	FALSE NEGATIVE
1	D1	D1	45	2	9
2	D1	D2	40	12	0
3	D1	D3	10	18	0
4	D2	D1	45	0	11
5	D2	D2	60	0	8
6	D2	D3	90	0	2
7	D3	D1	45	0	11

8	D3	D2	60	0	8
9	D3	D3	90	0	2
10	D1 & D2	D3	90	0	2
11	D1 & D3	D2	60	0	8
12	D2 & D3	D1	45	0	11
13	D1 & D2 & D3	D1	45	0	11
14	D1 & D2 & D3	D2	60	0	8
15	D1 & D2 & D3	D3	90	0	2
16	D1 & D2 & D3	CR	65	0	21
17	D1 & D2	CROSS VALIDATION	52.5	0	19
18	D1 & D3	CROSS VALIDATION	67.5	0	13
19	D2 & D3	CROSS VALIDATION	75	0	10

Table 4: Comparative Predication Accuracy

Exp. No	**Test Data Set**	**Average Accuracy (%)**		
		Data set Involved	Data set NOT Involved	Difference
1	D1	45	45	0
2	D2	60	53.33	6.67
3	D3	90	63	27

Table 5: Comparative Predication Accuracy: False Positive and False Negative

No	**Test Data Set**	**Data set Involved**	**Data set NOT Involved**	**Difference**
1	FALSE POSITIVE	1.5	3.33	1.83
2	FALSE NEGATIVE	1.43	5.8	4.37

VI. Conclusion:

From the above result, it can be concluded that when patient's records are taken into consideration during model building:

1. Prediction is more accurate. Our experimental results show 11.21% improvement in accuracy.
2. False Negative cases, which are very crucial in medical diagnosis, are significantly less. 4.37% less False Negative cased are observed.

References

[1] Polat, K. and S. Güneş, An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing, 2007. 17(4): p.702-710.

[2] Polat, K., S. Güneş, and A. Arslan, A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine. Expert Systems with Applications, 2008. 34(1): p. 482-487.

[3] Breault, J.L., C.R. Goodall, and P.J. Fos, Data mining a diabetic data warehouse. Artificial Intelligence in Medicine, 2002. 26(1): p. 37-54.

[4] Duhamel, A., et al., A preprocessing method for improving data mining techniques. Application to a large medical diabetes database. Studies in health technology and informatics, 2003. 95: p. 269.

[5] Miyaki, K., et al., Novel statistical classification model of type 2 diabetes mellitus patients for tailor-made prevention using data mining algorithm. Journal of epidemiology/Japan Epidemiological Association, 2002.12(3): p. 243.

[6] Sigurdardottir, A.K., H. Jonsdottir, and R. Benediktsson, Outcomes of educational interventions in type 2 diabetes: WEKA data-mining analysis. Patient education and counseling, 2007. 67(1): p. 21-31.

[7] Toussi, M., et al., Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes. BMC Medical Informatics and Decision Making, 2009. 9(1): p. 28.

[8] Buja, A. and Y.-S. Lee. Data mining criteria for tree-based regression and classification in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. 2001. ACM.

[9] Covani, U., et al., Relationship between human periodontitis and type 2 diabetes at a genomic level: a data-mining study. Journal of periodontology, 2009. 80(8): p. 1265-1273.

[10] Huang, Y., et al., Feature selection and classification model construction on type 2 diabetic patients' data. Artificial intelligence in medicine, 2007. 41(3): p. 251-262.

[11] Chaitrali S. Dangare , Sulabha S. Apte , A data mining approach for prediction of heart disease using neural networks. 2012, pp. 30-40.

[12] http://www.cdc.gov/media/pressrel/2010/r101022.html

Omprakash Chandrakar completed BSc (Maths) in 1997, MCA in 2000 and pursuing PhD. At present working as an Associate Professor at Department of Computer Science and Technology, UkaTarsadia University, Bardoli, Surat, Gujarat, India

Dr. Jatinderkumar R. Saini completed BSc and MCA, PhD. Currently he is Director I/C & Associate Professor at Narmada College of Computer Application, Bharuch, Gujarat, India.