Automatic blood glucose classification for gestational diabetes with feature selection: Decision trees vs. neural networks.

E. Caballero-Ruiz G. García-Sáez M. Rigla, M. Balsells B. Pons M. Morillo E.J. Gómez and M.E. Hernando

Abstract— **Automatic blood glucose classification may help specialists to provide a better interpretation of blood glucose data, downloaded directly from patients glucose meter and will contribute in the development of decision support systems for gestational diabetes. This paper presents an automatic blood glucose classifier for gestational diabetes that compares 6 different feature selection methods for two machine learning algorithms: neural networks and decision trees. Three searching algorithms, Greedy, Best First and Genetic, were combined with two different evaluators, CSF and Wrapper, for the feature selection. The study has been made with 6080 blood glucose measurements from 25 patients. Decision trees with a feature set selected with the Wrapper evaluator and the Best first search algorithm obtained the best accuracy: 95.92%.**

Keywords— **Classification, decision support, diabetes, decision trees, neural networks**

I. Introduction

Pregnancy is associated with changes in insulin sensitivity which may lead to changes in plasma glucose levels [1]. Gestational Diabetes Mellitus (GDM) is defined as glucose intolerance with onset or first recognition during pregnancy. Approximately 7% of all pregnancies (ranging from 1 to 14%, depending on the population studied and the diagnostic tests employed) are complicated by GDM [2]. Several adverse outcomes are associated with it, as preeclampsia, fetal macrosomia, perinatal mortality or neonatal respiratory problems and metabolic complications. Although most cases resolve with delivery, the woman maintains a more elevated risk of developing type 2 diabetes in the future, and this chronic hyperglycemia is associated with long-term damage, dysfunction, and failure of different organs, especially the eyes, kidneys, nerves, heart, and blood vessels [2].

Improving maternal glycemic control can reduce the risk of GDM complications, so patients should self-monitor their blood glucose (BG) levels with a glucose meter, and write their measurements down in a control book, along with information about intakes for a clinician to check it over once a week. The specialist determines the best treatment which consists in nutritional prescription, recommendation to practice physical activity and, if it is necessary, insulin administration.

Telemedicine in combination with Decision Support tools (DST) can improve GDM outcomes [3,4] without increasing clinician's workload [5]. Our final research goal is to develop intelligent tools integrated in a telemedicine system that allows control of GDM automatically, guarantying glucose control objectives consecutions and unnecessary in person visits to the health care center. DST can improve GDM treatment by helping the specialist in the control book inspection. These tools, following the expert indications, can preprocess the monitoring data contained in the control book, and determine which patient is evolving satisfactorily and which one needs a deeper examination by the specialist. They can also be integrated into a telemedicine system since the current glucose meters allow data download. Patients can send their BG levels directly to the system to be analyzed, and according to this information, the specialist will decide the corresponding treatment.

Automatic analysis of glucose meter files have to deal with the problem of lack of intake information associated to the measurements. Any DST requires to know whether the measurement was taken in breakfast, lunch or dinner time and if it is a pre-prandial or a postprandial measurement. However, most of the available glucose meters do not allow registering this information, or even if they do, patients forget to introduce these data. This information is essential to the specialist in order to evaluate the state of the patient so an automatic blood glucose classifier should be developed.

This paper presents a comparison of two well known supervised machine learning algorithms [6,7]: Decision trees and Neural networks (NN) for automatic BG classification. Different feature selection (FS) methods have also been compared in order to select the optimum feature set.

II. Materials and methods

Both learning algorithms, decision trees and NN, have been combined with 9 different feature sets obtained applying 6 different FS methods to our data set. A total of 18 classifiers have been built, which performance has been evaluated testing their accuracy.

A. Glycemic data

The data set (DS) consists of 6080 BG measurements from 25 patients, who were told to measure their BG with a glucose meter at least 4 times a day, in a fasting state and after the 3 main meals: breakfast, lunch and dinner. However some patients control their BG levels more often: before or/and between meals, at night or repeat some of them. They also wrote down the results in their control book, during a period ranging from the diagnosis date until the delivery date.

B. Data preprocessing.

a) Inputs

In order to procure the best classifier more features were obtained from the ones available in the glucose meter memory file, since in a previous study [8] we observed that classifiers accuracy improved with a large number of features. Directly from the memory file we acquire three features from each measurement: "*date*" and *"time"* when the measurement was taken and the *"bg"* concentration in mg/dL. We calculated another 17 features explained below.

From the *"date"* feature we calculate 5 more features: *"day", "month", "doy", "dow"* and *"workable"*, related respectively to the day and month of the date, the day of the year, day of the week and if the day is workable or not, as schedules and eating habits may change on weekends.

In a previous study [9], an expert determined that the insulin bolus administered close to a BG measurement was an important input to decide the measurement mealtime, which is our objective. We obtained 7 attributes related to insulin from the clinical history: *"insulin"*, a boolean indicating if the patient has insulin treatment or not, *"insulin_type"*, the type of the insulin treatment, *"rib", "ril", "rid", "si1"* and *"si2"* representing respectively the breakfast, lunch and dinner rapid insulin dose and the night and morning slow insulin dose. According to the same study, the time difference with the previous measurement was important too, so we calculated these 2 features related to that: *"interval_prev"*, and *"interval_post"*, which are the time difference with the previous and subsequent measurement. Counting on consecutive measurements BG values can provide information of whether an intake has taken place, so features *"bg_prev"* and *"bg_post"* have been calculated.

Finally we calculate the feature *"intake"*, which represents the most probable intake according to the patient schedules. It may have three possible values: breakfast, lunch or dinner and is obtained grouping each patient measurement in 3 subgroups according to *"bg"* and *"time"* attributes. This is done implementing a cluster in Octave [10] using the K-means algorithm [11].

In summary, we have a total of 20 features: *"date", "day", "month", "doy", "dow", "workable", "time", "interval_prev", "interval_post", "bg_prev", "bg", "bg_post", "insulin", "insulin_type", "rib", "ril", "rid", "si1", "si2" and "intake"*. We applied the "Remove useless" Weka filter to our DS to remove features that do not vary at all or that vary too much. The attribute *"slow ins 1"* is removed because none of our patients has that kind of treatment, leaving a total of 19 features.

b) Outputs

Ten different output classes have been used for the measurements classification: *"break-prep", "break-post", "lunch-prep", "lunch-post", "dinner-prep"* and *"dinner-post",* corresponding to main meals pre-prandial and post-prandial measurements; *"morning", "afternoon", "night" and "repeated"* corresponding to other measurements patients can make. The DS was labeled according to patients annotations contained in their control book.

C. Machine learning algorithms

The two learning algorithms used are the C4.5 Quinlan decision tree [12] and a Multilayer Perceptron (MLP) neural network [13]. The first one is characterized by good accuracy in a wide range of problems in addition to producing a comprehensible structure summarizing the knowledge it induces. It is also robust and fast, and it may degrade significantly its performance when dealing with many irrelevant features [14]. NN present lower accuracy than decision trees but are more robust. Another disadvantage of NN is that they work as a black box system where inputs and outputs are known but the output function is unknown. The architecture chosen has been MLP with 3 layers, input, hidden and output layers.

D. Feature selection

In order to select potentially relevant features from the ones we calculated, we have tested 6 different FS methods, combining three searching algorithms with two evaluators.

a) Evaluators

There are two main approaches for FS evaluation: wrappers and filters. We tested one of each approach, a wrapper and the Correlation based Feature Selection (CFS) which uses the filter approach:

Wrapper: Evaluates attribute sets by using a learning scheme. Cross validation is used to estimate the accuracy of the learning scheme for a set of attributes [14]. It is very computationally intensive.

CFS: Uses the Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred [15].

b) Searching algorithms

We tested three searching algorithms:

Greedy: Performs a greedy forward search through the space of attribute subsets. It starts with no attributes and stops when the addition of any remaining attributes results in a decrease in evaluation. It can also produce a ranked list of attributes [16].

Best First: Searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allows controlling the level of backtracking done. It considers all possible single attribute additions [16].

Genetic: Performs a search using the simple genetic algorithm described by Goldberg [17].

E. Evaluation method

FS and classifiers performance evaluation have been executed in an Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz using the Weka 3.6.9 [16] tool, because it provides the algorithms implementation we needed. Cross validation evaluation method [7] has been used for both tasks, with 3 folds for FS due to the wrapper execution time, and 10 folds for classifiers evaluation.

III. Results

A. Classifiers accuracy

Table 1 shows classifiers accuracy with each feature subset. The first three rows show results with the 3 initial features available in the glucose meter, with the features selected by the expert and with all the features we calculated in the preprocessing. In the first column appears the code to identify the FS method used, in the second column the FS evaluator, in the third column the FS search algorithm and in the last columns the C4.5 and MLP accuracy with the features selected by each FS method.

Table 1 Classifiers accuracy

Code	FS		Learning Algorithm	
	Eval.	Search Alg.	C4.5	MLP
GlucoMeter.	-	-	90,905%	86,875%
Expert	-	-	92,007%	87,823%
All	-	-	94,885%	93,273%
I	CFS	Greedy	95,395%	93,470%
II	CFS	Best First	95,395%	93,470%
III	CFS	Genetic	95,395%	93,470%
IV	Wrapper	Greedy	**95,921%**	94,340%
V	Wrapper	Best First	**95,921%**	**94,408%**
VI	Wrapper	Genetic	95,839%	94,079

Table 2 shows the number of features contained in each feature sets obtained with the feature selection methods showed above. Last columns shows which features are the ones selected.

Table 2 Feature selection (date[1], day[2], month[3], doy[4], dow[5], workable[6], time[7], interval_prev[8], interval_pos[9], bg_prev[10], bg[11], bg_post[12], insulin[13], insulin_type[14], rib[15], ril[16], rid[17], si2[18], intake[19])

Code	Nº of Features		Features selected	
	C4.5	MLP	C4.5	MLP
GlucoMeter	3	3	1,7,11	1,7,11
Expert	4	4	7,8,11,13	7,8,11,13
All	19	19	DS	DS
I	6	6	1,7-9,11,19	1,7-9,11,19
II	6	6	1,7-9,11,19	1,7-9,11,19
III	6	6	1,7-9,11,19	1,7-9,11,19
IV	**12**	11	**1,2,7-9,11,13-17,19**	1,7-11,14,15,18,19
V	**12**	**13**	**1,2,7-9,11,13-17,19**	**1,7-15,17-19**
VI	10	13	1,7-9,11,13-16-19,19	1,7-16,18,19

IV. Discussion

We observed that C4.5 achieves higher accuracy than MLP in all cases. Adding features to the initial three available in the glucose meter increases accuracy in both learning algorithms, though in MLP the improvement is higher. Adding 1 feature, we observed an improvement of 1.1% in C4.5 and 1% in MLP which rises to 4% in C4.5 and 6.4% in MLP when adding the rest of features we calculated in the preprocessing. Applying feature selection (Wrapper + Best

First) to these features we achieved a total improvement of 5% in C4.5 and 7.5 % in MLP.

CFS evaluator selects the same features regardless of the search algorithm and achieves less accuracy than wrapper (95.3% vs. 95.7% for C4.5 and 93.5% vs. 94.2% for MLP, in average). Wrapper execution time for MLP is very high, 72hours, while the execution of different alternatives for the C4.5 or the CFS takes less than 10 minutes.

BestFirst and Greedy achieved same results for C4.5 because they behave similarly, though the first one is a bit more thorough search technique.

The best accuracy results in both learning algorithms have been obtained with the feature selection method consisting in the Wrapper evaluator and the BestFirst search algorithm. In our case, increasing the feature search effort improved classifiers performance, though it is not always like that because of the bias-variance tradeoff [14].

V. Conclusion

C4.5 achieves higher accuracy than MLP, it is much faster to train and for feature selection, in addition of being more understandable for clinicians. The FS method consisting in wrapper evaluator and Best first search algorithm has proved to find the optimum feature set for the C4.5 and to achieve the best accuracy for our data set: 95,92%.

Automatic blood glucose classification is essential for automatic glucose meter file inspection. This not only will save pregnant women from problematic and unnecessary displacements, as they can download their measurements and send them to the system at any time, but also avoid the risk of making mistakes or oversights when transcribing the glucose meter results. Automatic BG classification will contribute to the development of DSS that can help in BG data interpretation counting with more exact and more available data.

Acknowledgment

This work has been funded by the Spanish grant “SINEDiE” (PI10/01125), co-funded by FEDER and the FP7 project “MobiGuide” (FP7-287811). We would like to thank Hospital de Sabadell and Hospital Mutua de Terrasa for their collaboration.

References

1. IDF Clinical Guidelines Task Force. (2009). Global guideline on pregnancy and diabetes. Brussels: International Diabetes Federation.
2. American Diabetes Association (2012) Diagnosis and classification of Diabetes Mellitus. Diabetes Care, 35(1):64–71 DOI:10.2337/dc12-s064.
3. Dalfrà, M. G., Nicolucci, A., & Lapolla, A. (2009). The effect of telemedicine on outcome and quality of life in pregnant women with diabetes. J Telemed Telecare, 15, 5:238–42. DOI:10.1258/jtt.2009.081213.
4. Ferrara, A., Hedderson, M. M., Ching, J., Kim, C., Peng, T., & Crites, Y. M. (2012). Referral to telephonic nurse management improves outcomes in women with gestational diabetes. Am. J. Obstet. Gynecol. 6, 206:491. DOI:10.1016/j.ajog.2012.04.019
5. Klonoff, D. C., & True, M. W. (2009). The missing element of telemedicine for diabetes: decision support software.J Diabetes Sci Technol,5,3: 996–1001
6. Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2007). Machine learning: a review of classification and combining techniques. Artif Intell Rev, 3, 26:159–190. doi:10.1007/s10462-007-9052-3
7. Geurts, P., Irrthum, A., & Wehenkel, L. (2009). Supervised learning with decision tree-based methods in computational and systems biology. Mol Biosyst, 1–18.
8. Caballero-Ruiz E, Garcia-Sáez G, Rigla Cros M, Balsells M, Pons B, Gómez Aguilera EJ, Hernando Pérez ME (2012). Clasificación de medidas de glucemia en función de ingestas en diabetes gestacional. CASEIB Proc. XXX Congreso Anual de la Sociedad Española de Ingeniería Biomédica, San Sebastián, Spain.
9. García-Sáez G, Alonso JM, Molero J, Rigla M, Martínez-Sarriegui I, de Leiva A, Gómez E, Hernando M (2009) Mealtime Blood Glucose Classifier Vased on Fuzzy Logic for the DIABTel Telemedicine System. Artif Intell Med, Lecture Notes in Computer Science 5651:295-304 DOI 10.1007/978-3-642-02976-9_42
10. GNU Octave at http://gnu.org/software/octave
11. Hartigan JA, Wong MA. (1979). Algorithm AS: A k-means clustering algorithm. Appl Stat-J Roy St C 28: 100-108
12. Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.
13. Hykin S (1994). Neural networks: a comprehensive foundation, Macmillan College Publishing Company Inc, New York .
14. Ron Kohavi, George H. John (1997). Wrappers for feature subset selection. Artificial Intelligence. 97(1-2):273-324 DOI 10.1016/S0004-3702(97)00043-X
15. M. A. Hall (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand
16. Witten H, Frank E, (2005). Data Mining: Practical machine learning tools and techniques. (2nd ed.)Morgan Kaufmann Publishers, San Francisco.
17. David E. Goldberg (1989). Genetic algorithms in search, optimization and machine learning. Addison-Wesley.

Author: Estefanía Caballero Ruiz
Institute: Polytechnic University of Madrid
Street: Avd. Complutense nº 30, Ciudad Universitaria. 28040
City: Madrid
Country: Spain
Email: ecaballero@gbt.tfo.upm.es