Prediction of Glucose Level Conditions from Sequential Data

Natalia MORDVANYUK [a,1], Ferran TORRENT-FONTBONA [a] and Beatriz LÓPEZ [a]
[a] *eXiT research group, University of Girona, Spain*

Abstract.
In type 1 diabetes management, mobile health applications are becoming a cornerstone to empower people to self-manage their disease. There are many applications addressed to calculate insulin doses based on the current information (e.g. carbohydrates intake) and a few of them are accompanied by modules able to supervise postprandial conditions and recommend corrective actions if the user falls in an abnormal state (i.e. hyperglycaemia or hypoglycaemia). On the other hand, mobile apps favour the gathering of historical data from which machine learning techniques can be used to predict if user conditions will worsen.

This work presents the application of k-nearest neighbour on the historical data gathered on patients, so that given the information related to a sequence of meals, the method is able to predict if the patient will fall in an abnormal condition. The experimentation has been carried out with the UVA-Padova type 1 diabetes simulator over eleven adult profiles. Results corroborate that the use of sequential data improve significantly the prediction outcome when forecasts distinguish the type of meal (breakfast, lunch and dinner).

Keywords. health apps, diabetes, k-nearest neighbour, sequential data, hypoglycaemia, hyperglycaemia, self-management

1. Introduction

Type 1 Diabetes Mellitus (T1DM) is a chronic disease that demands a strict control of the Blood Glucose (BG) level of the patient. This BG control is required to avoid hypoglycaemia or hyperglycaemia events, which are associated to serious short-term and long-term complications, e.g. coma, blindness, sever kidney failure or even death [7]. Despite that, it was estimated that in 2003, 7.1% of hypoglycaemia events required emergency assistance [9], and it is also estimated that about 6-10% of all deaths in people with T1DM result from hypoglycaemia [2]. Therefore, there is a need to improve the control of BG in people with T1DM.

This paper studies how sequences of data (recorded by T1DM people) can be used to predict hypoglycaemia and hyperglycaemia events using the k-nearest neighbours (k-NN) method.

This paper is organized as follows: First we start by the explanation of the related work. Then, we describe and justify the methodology and the criteria used in data la-

[1]Corresponding Author:eXiT research group, University of Girona, Campus Montilivi, building EPS4, Girona, Spain 17071; E-mail: natalia.mordvanyuk@udg.edu

belling and sequence generation. Next, we explain our experimentation set-up and the results achieved. Afterwards results are discussed and conclusions are presented.

2. Related work

In depth research has been done on developing new methods to predict hypoglycaemia and hyperglycaemia events and most of these methods are based on the analysis of the Continuous Glucose Monitoring (CGM) as in case of [2,5,10,11,12].

The need of CGM data is also a limiting factor, due to compatibility and connection issues, when developing applications that need this information to predict BG levels and alert for possible hypoglycaemia or hyperglycaemia events. Considering this possible lack of information, [13] presents a system that uses only carbohydrates ingestion and insulin doses to feed physiological models in order to predict hypoglycaemia events using support vector machines. Nevertheless, the results obtained are far from those using CGM data. In the same way, [14] studies the use of different artificial intelligence techniques to predict hypoglycaemia and hyperglycaemia events in type 2 diabetes patients using a few (e.g. two samples) BG readings per day.

Regarding the use of sequences in a k-NN system, one of the first attempts is [8] for solving an intrusion detection problem. Moreover, the authors in [3], similarly to this paper, represent sequences by concatenating records. However, the authors apply the methodology to predict the failure of complex medical equipment, instead of hypoglycaemia or hyperglycaemia events.

In the field of diabetes, case-based reasoning with sequential information regarding the actions of the subject has been studied in [1]. However, the objective of the work is to provide insulin doses recommendations, instead of predicting BG levels.

3. Methodology

KNN is a very well established methodology used for class prediction: given a set of historical examples $e_i \in E$, each of them consisting of a situation s_i and a class $c_i \in C$, and a new query situation q, KNN returns the class $c_q \in C$ predicted for q, according to the similarity of q with the historical data E. The most common representation of data in s_i are sets of attributes, e.g. $s_i = \{(temperature, high), (age, 34)\}$. In some occasions KNN has been used to handle sequential information meaning that information of a case is structured as follows:

$$e_i = \langle (s_i^1, s_i^2, \ldots s_i^n), c_i \rangle. \tag{1}$$

In the problem faced in this paper, the following information is available: Time (T) in minutes; Carbohydrate (CH) intakes (mg); Bolus insulin dose (B);CGM readings (mg/dL). This information is used to create an event for each meal.

Given the set of meal events, these are sorted according to the time attribute and processed to create sequences of meals where each sequence contains all the ordered meals of a time window of d days.The class of each sequence of meals is labelled according to the postpandrial status (PS) of the last meal of the sequence. These sequences are then

used to predict the class of a given sequence of meals using KNN. The similarity measure that we used is a generic Euclidean distance. To manage the unbalanced distribution of hypo and hyper labels, two separated case bases have been defined: hyper and no hyper, and hypo and no hypo.

Therefore, postprandial states have been labelled following two procedures:

- PS is labelled as hypoglycaemia if a CGM reading between 2 and 6 hours after the bolus administration is below 70 mg/dl [17], otherwise it is labelled as non-hypoglycaemia.
- PS is labelled as hyperglycaemia if CGM readings are above 180 mg/dl during at least 60 minutes between 2 and 6 hours after the bolus administration [17], otherwise it is labelled as non-hyperglycaemia.

4. Experimentation and results

The experimentation has been carried out using the UVA-Padova type 1 diabetes simulator [16] and it compares the use of k-NN with sequential data with the k-NN without sequential data, only *one-shot* data.

4.1. Dataset

The dataset consists of 11 virtual adults and the corresponding meal information and label along 500 simulated days. Virtual subjects took three meals per day (breakfast, lunch and dinner).The hypoglycaemia case bases had 3646 hypoglycaemia entries and 12726 non-hypoglycaemia. The hyperglycaemia case bases had 2833 hyperglycaemia entries and 13539 non-hyperglycaemia.

4.2. Experimental set-up

It has been empirically decided to use 4-day long sequences, which means that each sequence has the corresponding meals of four days. Since the variability of the results was not significant depending on the number of neighbours k of the k-NN method, and also in order to avoid over-fitting, k has been chosen following the rule of thumb[2] for each case. Moreover, three different scenarios have been implemented to analyse different hypothesis:

- All meals together: in this scenario the dataset contains all types of meals together (breakfasts, lunches and dinners).
- Meal-based recommendation: the original dataset is split into three different subsets, one per type of meal (breakfast, lunch and dinner) in order to study if this split improves the results since the context of each meal is different.
- Personalisation: the original dataset is split into as many subsets as virtual subjects in order to study if the use of a personal set of past experiences, instead of using cases from other subjects, improves the accuracy of the predictions.

[2]The rule of thumb proposes k as the square root of the number of points in the training data set [6]

G.C.	One-shot				Sequences			
	Accuracy (%)	TPR	FNR	FPR	Accuracy (%)	TPR	FNR	FPR
Hyper	68.14 ± 0.39	0.7451	0.2549	0.3823	70.29 ± 0.45	0.7186	0.2814	0.3128
Hypo	68.72 ± 0.33	0.7070	0.2930	0.3325	67.03 ± 0.61	0.6642	0.3358	0.3237

Table 1. All meals together. Results obtained using temporal data with the proposed methodology (sequences) and without using temporal data (one-shot), where G.C. is the glucose condition.

G.C.	Meal	One-shot				Sequences			
		Accuracy (%)	TPR	FNR	FPR	Accuracy (%)	TPR	FNR	FPR
Hypo	B	75.25 ± 0.45	0.7758	0.2242	0.2707	81.00 ± 0.49	0.8555	0.1445	0.2354
Hypo	L	73.65 ± 0.41	0.7296	0.2704	0.2566	77.31 ± 0.60	0.7350	0.2650	0.1887
Hypo	D	76.19 ± 1.06	0.7413	0.2587	0.2175	83.09 ± 0.44	0.8878	0.1122	0.2261
Hyper	B	78.07 ± 0.57	0.7789	0.2210	0.2177	83.64 ± 0.77	0.8258	0.1742	0.1530
Hyper	L	72.80 ± 0.66	0.8132	0.1867	0.3572	83.24 ± 0.44	0.7994	0.2006	0.1345
Hyper	D	64.09 ± 0.63	0.6947	0.3053	0.4129	73.89 ± 0.60	0.7652	0.2348	0.2874

Table 2. Meal-based recommendation. Results of the meal-based recommendation, where G.C. is the glucose condition, B is the breakfast, L is the lunch, and D is the dinner

G.C.	One-shot				Sequences			
	Accuracy (%)	TPR	FNR	FPR	Accuracy (%)	TPR	FNR	FPR
Hyper	71.57 ± 7.05	0.7394	0.2606	0.3080	56.04 ± 4.57	0.5650	0.4350	0.4448
Hypo	74.96 ± 6.91	0.7978	0.2022	0.2986	55.62 ± 2.82	0.5386	0.4613	0.4262

Table 3. Personalisation. Results per patient using sequences and without using them (one-shot), where G.C. is the glucose condition.

For each scenario, we have compared the one-shot (entries with only one meal) with our sequential data approach (sequences of four days). Stratified 10-fold cross validation has been performed and the resulting sets have been balanced randomly sub-sampling the majority class. The following metrics was used to compare the results: The accuracy[3] because it provide us the overall information of how often the classifier is correct, true positive rate (TPR)[4] also known as recall or sensitivity, false negative rate (FNR)[5], and false positive rate (FPR)[6].

4.3. Results and discussion

The results are summarized in Tables 1,2,3 and 4, which show the average TPR, FNR and FPR and the average is standard deviation of the accuracy along the cross-validation folds.

The results of all meals together (Table 1) show that the use of sequences does not provide an improvement in terms of either accuracy, TPR, FNR or FPR, because the context of each type of meal is so different that predicting the class of a meal (e.g. breakfast) relying on other type of meals (e.g. dinner) is useless.

On the other hand, when datasets are divided according to the type of meal (see Table 2), the overall performance increases (TPRs up to 0.88 and accuracies up to 83% have

[3] Accuracy = (True Positives + True Negatives)/ total number of instances

[4] TPR = True Positives/(True Positives+False Negatives)

[5] FNR = False Negatives/(True Positives+False Negatives)

[6] FPR = False Positives/(False Positives+True Negatives)

G.C.	Meal	One-shot				Sequences			
		Accuracy (%)	TPR	FNR	FPR	Accuracy (%)	TPR	FNR	FPR
Hypo	B	56.41 ± 0.0467	0.5190	0.4809	0.3907	52.11 ± 0.0349	0.4042	0.5958	0.3619
Hypo	L	62.31 ± 0.0577	0.6070	0.3930	0.3607	55.51 ± 0.0523	0.4891	0.5109	0.3788
Hypo	D	62.94 ± 0.0650	0.6024	0.3976	0.3435	57.78 ± 0.0716	0.5695	0.4305	0.4139
Hyper	B	52.35 ± 0.0549	0.4544	0.5456	0.4073	49.64 ± 0.0555	0.4354	0.5646	0.4425
Hyper	L	57.82 ± 0.1041	0.5430	0.4570	0.3865	51.81 ± 0.0681	0.4758	0.5242	0.4396
Hyper	D	59.55 ± 0.0579	0.6087	0.3913	0.4177	53.77 ± 0.0377	0.5135	0.4865	0.4382

Table 4. Personalisation. Results per patient of the meal-based recommendation, where G.C. is the glucose condition, B is the breakfast, L is the lunch, and D is the dinner

been achieved, while FNR and FPR are significantly lower, from 0.25 to 0.11), which reinforces the previous conclusion. Furthermore, in this scenario, the use of sequential data clearly outperforms the k-NN with one-shot data meaning, which highlights the relevance of past information to predict hypo- and hyperglycaemia events.

The results achieved using personal datasets (Tables 3 and 4) are worse than the achieved in the previous scenario, except when the dataset is not divided by the type of meal and sequences are not used (see Table 3), for which results are slightly better. These results mean that an over-split of the data available finally leads to a low performance due to a lack of data. Moreover, if we consider that 500 days have been simulated and a greater amount of data could be difficult to be available for real subjects, we can conclude that the use of personal datasets with the proposed methodology is highly inefficient.

5. Conclusions

The use of continuous glucose monitoring is crucial to help people with diabetes to prevent hyperglycaemia and hypoglycaemia events that could convey severe health problems. However, continuous glucose monitors are not always available. This paper presents a prediction methodology based on sequencing patient data and forecasting their status using k-NN has been presented towards this end. The proposed methodology achieves a TPR up to 0.88 considering only carbohydrates intakes, bolus dose and preprandial blood glucose on UVA/Padova type 1 diabetes simulator, depending on the context. k-NN, however, does not provide an explicit pattern of the patient behaviour regarding the disease in order to provide medical evidence for clinicians. In a future work, other eager mechanisms for sequence learning [15], or even its hybridization with sequence learning as done in [4] should be explored, or even a fuzzy approach in the labelling of the instances in order to avoid the crisp borders.

Acknowledgements

This work has received funding from the EU Horizon 2020 research and innovation programme under grant agreement No 689810 (PEPPER), and from the University of Girona under the grant MPCUdG2016 (Ajut per a la millora de la productivitat científica dels grups de recerca), and the Spanish MINECO under the grant number DPI2013-47450-C21-R. This work has been developed with the support of the research group SITES awarded with distinction by the Generalitat de Catalunya (SGR 2014-2016).

References

[1] D. Brown. *Temporal case-based reasoning for insulin support*. PhD thesis, Oxford Brookes University, 2015.

[2] S. L. Cichosz, J. Frystyk, O. K. Hejlesen, L. Tarnow, and J. Fleischer. A novel algorithm for prediction and detection of hypoglycemia based on continuous glucose monitoring and heart rate variability in patients with type 1 diabetes. *J. of Diab. Sci. and Tech.*, 8(4):731–737, jul 2014.

[3] M. Compta and B. López. Integration of sequence learning and CBR for complex equipment failure prediction. In *Case-Based Reasoning Research and Development*, pages 408–422. Springer, 2011.

[4] P. Gay, B. López, and J. Meléndez. Sequential learning for case-based pattern recognition in complex event domains. In *Proceedings of the 16th UK Workshop on Case-Based Reasoning*, pages 46–55, 2011.

[5] E. I. Georga, J. C. Principe, D. Polyzos, and D. I. Fotiadis. Non-linear dynamic modeling of glucose in type 1 diabetes with kernel adaptive filters. In *2016 38th Annual Int. Conf. of the IEEE Eng. in Med. and Biology Society (EMBC)*, aug 2016.

[6] A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat. Solving the problem of the k parameter in the knn classifier using an ensemble learning approach. *arXiv preprint arXiv:1409.0919*, 2014.

[7] J. Hippisley-Cox and C. Coupland. Diabetes treatments and risk of amputation, blindness, severe kidney failure, hyperglycaemia, and hypoglycaemia: open cohort study in primary care. *BMJ*, 352, 2016.

[8] T. Lane and C. E. Brodley. Temporal sequence learning and data reduction for anomaly detection. *ACM Trans. Inf. Syst. Secur.*, 2(3):295–331, 1999.

[9] G. P. Leese, J. Wang, J. Broomhall, P. Kelly, A. Marsden, W. Morrison, B. M. Frier, and A. D. Morris. Frequency of severe hypoglycemia requiring emergency treatment in type 1 and type 2 diabetes. *Diabetes Care*, 26(4):1176–1180, 2003.

[10] X. Mo, Y. Wang, and X. Wu. Hypoglycemia prediction using extreme learning machine and regularized. In *2013 25th Chinese Control and Decision Conf. (CCDC)*, 2013.

[11] S. Oviedo, J. Vehí, R. Calm, and J. Armengol. A review of personalized blood glucose prediction strategies for t1dm patients. *Int. J. for Numerical Methods in Biomedical Eng.*, page e02833, 2016.

[12] S. M. Pappada, B. D. Cameron, and P. M. Rosman. Development of a neural network for prediction of glucose concentration in type 1 diabetes patients. *J. of Diab. Sci. and Tech.*, 2(5):792–801, 2008.

[13] K. Plis, R. Bunescu, C. Marling, J. Shubrook, and F. Schwartz. A machine learning approach to predicting blood glucose levels for diabetes management. *Modern Artificial Intelligence for Health Analytics. Papers from the AAAI-14*, 2014.

[14] B. Sudharsan, M. Peeples, and M. Shomali. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. *J. of Diab. Sci. and Tech.*, 9(1):86–90, nov 2014.

[15] R. Sun and C. L. Giles. *Sequence learning: Paradigms, algorithms, and applications*. Springer, 2003.

[16] R. Visentin, C. Dalla Man, B. Kovatchev, and C. Cobelli. The university of virginia/padova type 1 diabetes simulator matches the glucose traces of a clinical trial. *Diab. Tech. & Therapeutics*, 16(7):428–434, 2014.

[17] C. Zecchin. *Online Glucose Prediction in Type 1 Diabetes by Neural Network Models*. PhD thesis, Università degli Studi di Padova, 2014.