Managing Diabetes Therapy through Datastream Mining

Simon Fong, University of Macau, Macau SAR
Jinan Fiaidhi and Sabah Mohammed, Lakehead University, Canada
Luiz A.M. Moutinho, Dublin City University, Ireland

The authors’ datastream mining approach computationally derives real-time decision rules for formulating insulin-dependent diabetes mellitus therapy based on insulin prescription records and patients’ blood-glucose reactions.

According to the latest figures published by the World Health Organization, the number of people worldwide suffering from diabetes grew from 153 million in 1980 to 422 million in 2016 (www.who.int/diabetes/en). Around 3.7 million deaths are attributable to diabetes and high blood-glucose levels. Managing diabetic therapy—such as insulin-dependent diabetes mellitus (IDDM) therapy—can be challenging because of the need to adjust insulin dosages in response to fluctuating blood-glucose levels that are influenced by patients’ ever-changing physiological conditions.

Some common medical guidelines have been defined for the masses, but there is no adaptive equation for balancing blood-glucose levels and insulin intake dosages customized for every individual.

Due to their varying physiological characteristics, patients react differently to exogenous insulin. Moreover, bodily reactions to insulin can change over time, even for the same person, because of lifestyle alterations. Factors that are known to influence blood-glucose concentrations include body-mass ratio, hormone balance, mental state, diet, and physical exercise. The last
Related Work in Diabetes Therapy

It has been used to predict the correct insulin dosage to be administered at the right time and thus to account for fluctuating blood-glucose concentrations at different times of day. The aim is to keep glucose levels within some recommended range. Given the dynamicity, uncertainty, and variability of the complex interactions between the human body and insulin, along with unknown physiological factors, the decision process is far too complicated to be a manual task. One IT-enabled example is DIAdvisor, a personal blood-glucose control device. Once a user sets his or her body weight and age into the device, any deviation from the preset control is detected as a sign of hyper- or hypoglycaemia, as deduced by the decision rules. The rules, however, are generalized from a population. They are not patient-specific, and they do not adapt to the individual’s body, lifestyle, and eating habits.

Researchers have attempted to use time series forecasting methods to predict patients’ upcoming insulin intake cycles and their blood-glucose levels. Artificial neural networks are another prevalent tool for constructing a blood-glucose predictor. Other machine learning methods—including decision trees, probabilistic Bayes networks, fuzzy lattice reasoning, the sparse factor graph model, and fuzzy logic—have been applied and reported in the literature. In general, prediction models have been trained by supervised learning to recognize nonlinear mapping between multiple factors and their outcomes. Static training samples from historical data are often used. A common drawback is that the models are not adaptive because they were trained from static data. Recently, datastream mining methods have been evaluated as an alternative to conventional data mining methods in the case of a clinical decision support system. The results showed that datastream mining algorithms offer prediction accuracy that is on par with that of traditional data mining methods.

References
two variables, together with a patient’s course of insulin intakes, form a dynamic intervention pattern. One of the main objectives of diabetic therapy management is to determine how this intervention pattern should be dynamically adjusted, in both intensity and frequency, to a patient’s fluctuations in blood-glucose levels.

Here, we present a datastream mining approach to computationally derive real-time decision rules for formulating IDDM therapy based on insulin prescription records and patients’ blood-glucose reactions.

**Decision-Rule Generation with Datastream Mining**

This article proposes a datastream mining approach followed by decision-rule generation. The differences between the traditional data mining process and the incremental datastream mining process are illustrated in Figure 1. The traditional data mining process is a typical train-then-test approach in which a classification model must first be induced from a full batch of historical data before it becomes useful for testing with unseen samples (Figure 1a). Our proposed method of decision-rule generation is extended from a datastream mining process, as shown in Figure 1b. The steps in the high-level process flow are also sequentially labeled in the diagram. Note that in datastream mining, the test-and-train steps are intertwined.

The input data are taken as a continuous datastream that is loaded via a sliding window into the incremental learning process to induce a decision tree. The decision tree is built progressively by expanding tree branches that cover new conditions found in the fresh data. These fresh data are first tested by the decision tree, which predicts a result based on the samples that it has learned. Rather than reconstructing the whole
decision tree each time by reloading the entire set of training data, incremental learning refreshes the decision tree only when its prediction performance drops marginally. This incremental learning approach enables the timely prediction of the diabetic condition from the continuous input of blood-glucose measures and insulin doses. The input data are an unbounded time series—that is, streaming data.

The incremental learning algorithm in use, Hoeffding Tree Adaptive (HTA), was solicited with the aim of establishing the best-performing classifier and marking out the data subset from the datastream to induce the quality decision rule. According to the HTA model, the portions of relevant data increase in length as more training samples arrive; the accuracies of both classifiers mature and converge. This implies that the longer the model is in use, the better its efficacy in producing quality decision rules. We conducted an experiment to find the classifier that offers the best accuracy while taking the shortest training time, as discussed next.

Finally, using a simple method called bump hunting (BH), the decision rules, given the appropriate amount of data, can be efficiently generated. From the whole search space estimated from the input attributes, BH looks for a set of minimum subregions within which the value of the predicted output is maximized. The BH scheme follows the strategy of “patient” rule induction, which is in contrast to partitioning tree methods (which are used in most decision tree algorithms). The patient scheme constructs decision rules from minimized subregions by linking up the conditional tests of the data features or attributes.

Experiment
The objectives of our experiment were two-fold. Ultimately, the proposed model would be validated using empirical diabetes data. The validation would first be done by testing the proposed model and the data using several popular datastream mining algorithms. The first objective was to obtain the most appropriate datastream mining algorithm for inferring the decision rules. The second objective was to verify the decision-rule mechanism, using the BH method. Decision-rule generation was guided by the most accurate datastream mining model, which identified the relevant data portions from the datastream as the search spaces for BH to generate rules.

Data Setup
The dataset used in the experiment was acquired from the 1994 Spring Symposium on Artificial Intelligence in Medicine website (archive.ics.uci.edu/ml/datasets/Diabetes). The dataset was originally used for the Interpreting Clinical Data contest. It logged successive events about a typical course of IDDM-2 therapy being monitored. The data were collected from 70 diabetes patients over six months. The main events that were measured included different types of insulin doses, such as regular, neutral protamine Hagedorn (NPH), and blood-glucose concentration. Blood glucose was measured at least three times a day. Each patient was injected with normal insulin roughly three times a day and with fast-acting insulin about once a day. All injection times were approximate, given that there was no precise injection timing each day. There were occurrences of hypoglycemia recorded in the dataset. A label marked GTM (meaning greater than mean) was added. The label had a binary value of 0 or 1. The value was decided by comparing the current blood-glucose value to the weekly average. To capture temporal essences from the data, the dataset was transformed from raw to some converted attributes, such as how long (in minutes) since the last event occurred for each event.

We also assumed the following information for insulin therapy: 15–45 minutes for the onset period, a peak effect between 60–180 minutes, and a lasting effect between 240–360 minutes for regular insulin; and 60–180 minutes for the onset period, a peak effect between 240–360 minutes, and a lasting effect between 600–840 minutes for fast-acting insulin. These assumptions help regulate the prediction system’s behavior by imposing minimum and maximum lengths of time between successive blood-glucose measures. An instance of training data with values over or under these constraints is considered void. For example, glucose tests that were made too frequently in the minimum effective time or too sparsely beyond the maximum effective time are discarded. The raw dataset, which is in the form of a time series, is a complex pattern comprising the insulin doses and blood-glucose measures interleaved along the
temporal domain. This made it computationally challenging for the datastream mining algorithms to recognize the data. The time series is transformed into instances of records measuring the current blood-glucose level, how long ago the last insulin dose was injected, which type was injected, and the applied dosage. The final columns (BGM value, GTM, and Hypo) are used as verdicts or labels in the supervised training for model induction. The model tries to learn the relations or data mappings between the input attributes, such as insulin intakes, and the outcomes as consequences of medicine intakes. Once the relations are learned, the model can predict the likely outcomes (which are bodily reactions) given some insulin intake records.

Performance Comparison

The dataset was subject to the datastream mining software program Massive Online Analysis (MOA; moa.cms.waikato.ac.nz), which is Java-based and open source for experimentation. On the MOA benchmarking platform, different datastream mining algorithms, programmed in Java, can be applied on some supplied dataset to evaluate and compare algorithm performance. We selected six datastream mining algorithms to data-mine the diabetes dataset to predict hypo cases (five out of the six were popular choices that were recently invented): Hoeffding tree (HT), implementing the classical very fast decision tree (VFDT); k-nearest neighbor (KNN); online regression trees with options (ORTO); stochastic gradient descent (SGD); support vector machine (SVM)-based logistic regression; and naïve Bayes (NB). The five popular algorithms were compared with HTA, which is an improved version of HT.3

The experiment predicted whether hypoglycemia would happen given patients’ last taken insulin doses and their timings. The prediction model was incrementally built according to the method depicted in Figure 1b: fresh data were tested first and then deferred for use in refreshing the model, if necessary. There were no cascading errors because the model refresh in each cycle was independent. Rather, the quality of the classifier increased as more incoming data were used to continuously train the model. The performance dropped only when the new data contradicted current concepts. The sliding window size was 100. Two major aspects of the performance—accuracy and time consumption—are plotted in Figures 2 and 3, respectively, to compare the six algorithms.

As observed in Figure 2, HT and HTA were the only two algorithms that could reach an equilibrium of accuracy at high values—95.96 and 96.93 percent, respectively. However, HTA attained
steady-state top performance earlier than HT, at about the 35,400th instance. HT took slightly longer and attained top performance at about the 49,500th instance. This shows a phenomenon in which an almost perfect classifier was possible, but where time was needed to train the classifier with a sufficient amount of samples because the training samples arrive quickly in a streaming fashion. It proved our hypothesis that the longer the training went on, the more stable the prediction results if HT or HTA were used. HTA had an edge over HT because of its dual classifier design; the main HTA classifier could adapt to the data streams while its supplementary classifier was monitoring fluctuations.

The accuracy curves for the other algorithms appeared to be stationary time series that zigzagged up and down along the way in time, without any increasing trend. Although KNN, an incremental learning algorithm that is also known as a lazy learner, outperformed HT and HTA in the early stages, its performance did not improve further. Given enough warm-up time and samples for training HT and HTA, their performances exceed that of KNN later on. The accuracies of the different algorithms were ranked as HTA > HT > KNN > ORTO > SGD > NB. NB performed the worst among the six with a mean of 42.6 percent, possibly because NB assumed that the data attributes were totally independent. In reality, we know that any injection of insulin into the body would have a certain effect on lowering blood glucose in a due measurement. The low accuracy of NB probably counter-proved that the attributes of the IDDM-2 dataset were indeed dependent on one another to a certain extent.

Speed was another performance indicator that Figure 3 shows as time consumption in seconds. In datastream mining, an algorithm that yields the highest accuracy, yet mines the data in the shortest time, is ideal. In Figure 3, SGD is the fastest, then NB, which took almost zero seconds in processing. However, SGD and NB were disqualified because of their very low accuracy and close to zero kappa values. HT (at 0.3 seconds) and HTA (at 0.74 seconds) were the next fastest candidates. HTA outperformed its original version of HT in terms of accuracy. The improvement in accuracy, however, cost extra runtime via the additional mechanism of monitoring the datastream while keeping the main classifier adaptive. ORTO took an average of 0.96 seconds, and KNN was the slowest and consumed an average of 6.06 seconds.

**Decision-Rule-Generation Ability**

In addition to quantitative experimentation in comparing the prediction algorithm’s performance, we tested the algorithm’s ability in
decision-rule generation. The latest subset of training data, as found by HTA (which is supposed to offer the highest accuracy) was extracted from the datastream and put into the BH method for rule generation. Rule generation spanned two parts, with one part per class label. In the case of IDDM-2, there was a set of decision rules generated that described the conditions (with propensities), leading to the verdict of Hypo = Yes. Likewise, another set of rules was generated that described the safe situations in which the patient would not enter hypoglycemia—that is, Hypo = No.

The decision rules that described both conditions (Hypo = Yes/No) are visualized as a branching tree in Figure 4.

Figure 4a shows that there were 14 occurrences of hypoglycemia out of the most recent and relevant input data subset (of window size 100). Tracing the decision rules in Figure 4a from left to right, if a sample was randomly chosen from all 100 samples, it was 14 percent of the propensity at the root. However, when a conditional test was added, such as last_regular_dose ≤ 5, the chance of hitting Hypo = Y was enhanced to 18.75 percent. The chance of the verdict Hypo = Y increased as more conditional tests were added. It increased all the way up to 62.5 percent (rightmost in the figure) with the presence of all necessary conditional tests that test the attributes of an unseen sample along this sequence.

In this way, these decision rules could help a user to map out guidelines for preventing Hypo = Y—for instance, by ensuring that the last regular insulin dose taken was no less than five units and that the time of insulin intake since the previous one was no longer than 0.1639 per day, and so forth regarding the NPH.

The decision rule in Figure 4b was also telling regarding conditions that led to Hypo = N. Safety guidelines could be formulated from these rules, highlighting the conditions, timings, and dosages to be observed to prevent a situation (hypoglycemia) from occurring. These decision rules were obtained in real time, given that datastream mining can take place in real time along with continuous input data feeds. To enable decision making, decision rules offer insights about the conditions in which a certain situation will (or will not) happen; a classifier foretells whether a situation will happen given current situations. These two decision support tools work hand in hand, complementing each other via different intelligence formats.

The inferred rules could be programmed into mobile apps on smartphones that serve to alert users to take insulin on time with the appropriate dosage. Because bodily reactions with medicine are different for every individual, and people’s lifestyles and diets do change over time, this rule-generation method helps derive the most suitable and up-to-date rules that can adapt fully to users. The rules would help guide users with a customized insulin intake course and prevent hypo- or hyperglycemia. Nevertheless, it is known that patients often have comorbidities and thus
might have numerous other clinical issues that could impact diabetes management. Our method shows a simple example by mapping rules that reflect only the relations between insulin intake and blood-glucose outcomes. More complex rules could be generated by extending the data mining model to include extra attributes or factors that impact outcomes.

Our new breed of computational methods—a paired classifier that can predict whether a situation could occur and decision rules that offer insights on preventing certain medical situations—are enabled by datastream mining. Future work includes programming these decision rules into an interactive system. Our proposed model could be extended by incorporating information about more factors than those tested here.

Our model can assist healthcare experts in finding a suitable dosage and the correct timing of insulin administration based on decision rules so that the fluctuation of blood-glucose concentrations can be regulated to a stable level. The decision rules are patient-specific and can be applied to some personalized diabetic advisor, customized to a patient’s lifestyle and health requirements. The rules could be applied to patients’ comorbidities in the form of recommended courses of insulin management or be programmed as a monitoring alarm running on smartphone mobile apps.

Acknowledgments
The authors are thankful for financial support from the following research grants: Temporal Datastream Mining by Using Incrementally Optimized Very Fast Decision Forest (iOVFDF), grant number MYRG2015-00128-FST offered by the University of Macau, FST, and RDAO; and A Scalable Datastream Mining Methodology: Stream-Based Holistic Analytics and Reasoning in Parallel, grant no. FDCT-126/2014/A3, offered by FDCT Macau.

References

Simon Fong is an associate professor in the Department of Computer and Information Science at the University of Macau, Macau SAR. His research interests include data mining, metaheuristics, and big data analytics. Fong received a PhD in computer science from La Trobe University. He is a member of IEEE and ACM. Contact him at ccfong@umac.mo.

Jinan Fiaidhi is a professor in the Department of Computer Science at Lakehead University, Canada, and an adjunct research professor at the University of Western Ontario. Her research interests include mobile and ubiquitous learning utilizing emerging technologies, such as cloud computing, calm computing, learning analytics, mobile learning, personal learning environments, social networking, enterprise mashups, big data, and the Semantic Web. Fiaidhi received a PhD in computer science from Brunel University London, UK. She is an IEEE Senior Member, a Member of the Chartered Institute for IT, an information systems specialist, and a member of the Society of Professional Engineers. Contact her at jinan.fiaidhi@lakeheadu.ca.

Sabah Mohammed is a professor in the Department of Computer Science at Lakehead University, Canada. His research interests include big medical data, web intelligence, security of ubiquitous data, data mining, cloud computing, and health informatics. Mohammed received a PhD in computer science from Brunel University London, UK. He is an IEEE Senior Member, a Member of the Chartered Institute for IT, an information systems specialist, and a member of the Society of Professional Engineers. Contact him at mohammed@lakeheadu.ca.

Luiz A.M. Moutinho is a professor of bio-marketing and Futures research in the DCU Business School at Dublin City University. His research interests encompass big data, social media and memetic algorithms, emotionomics and computer speech signals, neuroscience in marketing, evolutionary algorithms in marketing, fuzzy logic models in consumer behavior, and neuroscience in marketing. Moutinho is the founding editor of the Journal of Modelling in Management (JM2), has four associate editorships, and is on the editorial board of another 46 international academic journals. He received a PhD in marketing management from the University of Sheffield. Contact him at luiz.moutinho@dcu.ie.