Machine learning approaches to (i) predicting response to therapy in diabetes, (ii) data-driven diabetes subtype classification and (iii) synthetic data generation

Chris Sainsbury
Queen Elizabeth University Hospital, Glasgow
NRS Senior Fellow
Hon Senior Research Fellow, University of Birmingham

@csainsbury
http://glucose.ai
http://github.com/csainsbury/
• MD: endothelial function
• mathematical modelling of systemic/pulmonary circulation
• inpatient CBG data analysis
  • patient risk stratification
  • system performance analysis (clinical unit / individual operator)
  • just-in-time education for ward staff
• SCI diabetes data analysis
  • HbA1c variability and outcome in T1DM
  • multiscale / multiparameter variability and outcome
• Machine Learning
  • ANN vs logistic regression vs clinician prediction of DM type at diagnosis
  • myDiabetesIQ
• THINKINGAI group – University of Birmingham
SCI-Diabetes

‘we have (one of) the best datasets in the world, so there is an opportunity to build high quality algorithms to understand relationships within diabetes data…’

what questions to address?
initial questions / prediction problems chosen

i. suggest best next therapy / combination of therapies to achieve goals in multiple domains (hba1c reduction / blood pressure / mortality etc)

ii. predict complications (LLA, CV events)

iii. predict acute complications (hypoglycaemia etc)

iv. predict diabetes type at diagnosis, identify MODY etc
Innovate UK (Digital Health Technology Catalyst) 1M grant 2018-2021

Funding competition

Digital health technology catalyst 2017 round 1

UK businesses can apply for a share of up to £8 million to speed up development of new digital technology healthcare solutions.

Competition opens: Monday 31 July 2017
Competition closes: Wednesday 11 October 2017 12:00pm

Competition: Digital Health Technology Catalyst 2017 Round 1
Project Title: MyDiabetesIQ
Question 1

To generate a prediction of an individual’s response to any given clinically appropriate drug / combination of drugs - drawing on insight gained from the population over time.

or

what is/are the next best drug(s) for my patient?
what is the next best drug(s) for my patient?
virtual n = 1 drug trial

e.g. what drug should I prescribe to give this patient the best chance of having an HbA1c <60mmol/mol, with a reduction in blood pressure and BMI in 1 year?

taking into account their individual history of:

- HbA1c / BMI / blood pressure
- previously prescribed combinations of drug therapies
- how previous drugs have impacted on HbA1c / BMI / blood pressure
- sex
- age
- ethnicity

\[
\text{time series}
\]

\[
\text{stable over time}
\]
managing time series data – numerical data

hba1c

sbp

bmi
managing time series data – 2
numerical time series

managing missingness
- imputation
- lvcf / other
- interpolation – spline / linear
- masking

scaling / normalisation
managing time series data – prescription data
managing time series data – 3
drug combinations as words - for natural language processing approach

Drug Sentence: MF, GLP1_MF, GLP1_MF_SGLT2, GLP1_MF

Embedding -> numerical vector
input into RNN / LSTM
Recurrence Neural Network (LSTM) extracts information from sequence of input
multi-dimensional RNN extracts information from the interactions between input
sequences over time
schematic of RNN/LSTM based classifier
training, validation and withheld test sets
environments used:

- SCI diabetes data input
- visualisations etc

Languages and tools used:

- Python
- R
- Keras
HbA1c  
SBP  
BMI  
Drug Combinations

**runin period**

**test period**

- HbA1c
- SBP
- BMI
- Drug Combinations

**outcome measure of interest**

eg
change in HbA1c
change in SBP
change in BMI
(or final values for regressor version)

**training data structure**
Using the model to predict response

- run-in period
- test period

- HbA1c
- SBP
- BMI
- Drug Combinations

Predictions:
- change in HbA1c
- change in SBP
- change in BMI

(or final values for regressor)

Test drug 1:
- drug

Test drug 2:
- drug

Test drug 3:
- drug

Test drug 4:
- drug
bidirectional LSTM classifier

i. simple classification endpoint:
   
   eg
   
   • probability of a 10mmol/mol reduction in hba1c at 1y
   • probability of achieving HbA1c in range 48 - 60mmol/mol at 1y

ii. composite endpoints:
   
   eg
   
   probability of achieving HbA1c 48 - 60mmol/mol, with a SBP of <140mmHg
simple problem: which of 4 combinations most likely to reduce hba1c by 10mmol/mol?
more complex problem: which of 16 combinations most likely to reduce hba1c to <60mmol/mol without causing weight gain?

last measured BMI at end year 4

last measured HbA1c at end year 4
probability of HbA1c reduction >10mmol/mol at 1y n = 1450 (initial HbA1c >60mmol/mol)
The best combination is **dpp4_mf** to achieve HbA1c <60mmol/mol, with reduction in BMI. The best combination for HbA1c <60mmol/mol, with reduction in SBP is **mf_sgl2_su**.
regressor

remove sigmoid activating function from final layer

-> returns value that can be mapped to predicted outcome value (eg HbA1c)
schematic of RNN/LSTM based regressor

Time Stable Data

- age
- sex
- ethnicity

standardisation / normalisation of data

merge

LSTM

Time Series Data

- HbA1c
- blood pressure (SBP / DBP)
- weight / BMI
- prescribed drug combination

standardisation / normalisation of data

n * (hidden layer)

outcome of interest (boolean)

main output (value)

auxiliary output (value)
model as regressor – prediction of HbA1c at 12 months vs actual

\[ r^2 = 0.66 \]
representation of predicted drug effects

predicted hba1c multiple possible drugs

comparator (baseline) drug combination
current performance

classifier – AUC (modelling actual drug therapy) 0.75 – 0.85
regressor – $r^2 \sim 0.65$

ongoing work

optimise neural network / layers used (LSTM vs convolutional etc)
explore alternative techniques (Gokhale/Tino bayesian inference approach etc)
some current issues

- imbalanced classes: downsampling / upsampling approaches
- stochastic variation: k-fold validation / averaging multiple runs
- hyperparameter tuning: increase computational power / use small samples

**diabetes classification issues**

- unsupervised reclassification

**data sharing**

- synthetic data generation
diabetes classification issues
Machine-Learning (neural network) driven algorithmic classification of Type 1 or Type 2 diabetes at the time of presentation significantly outperforms experienced clinician classification

CAR Sainsbury, R Muir, J Osmanska, GC Jones

glucose.ai Research Group, Department of Diabetes, Gartnavel General Hospital, Glasgow, Scotland

Background / Aims

Classification of type of diabetes at the point of diagnosis may not always be straightforward. We aimed to develop an algorithmic approach to the problem, using data available within the SCI-Diabetes dataset. Metrics that would be routinely available at the time of first presentation were chosen for inclusion. To assess the potential clinical utility - and to benchmark performance - a subset of individuals were presented to experienced clinicians who classified as either Type 1 or Type 2 diabetes. The accuracy of classifications generated for individuals within the subset by both algorithm and by clinicians were compared, using the established diagnosis within SCI diabetes as the comparator diagnosis. Individuals were included in the analysis only if they had a date of diagnosis at least 12 months from the data extraction date, in order to ensure a high likelihood of a stable and correct diagnosis being achieved.

Data was prepared for analysis using R(1), and analysis code was written in Python(2). An artificial neural network was chosen for the algorithmic approach, implemented in Tensorflow (3) (written using the Keras library (4))

A further subset of the test subset was generated for clinician classification. Data from individuals in this subset were presented in batches of 100 to experienced clinicians within our clinic. The proportion of individuals with a recorded diagnosis of Type 1 Diabetes within each batch was varied at random between 0.05 and 0.5 to reduce the possibility of clinicians inferring a diagnosis from the proportion already identified.

The ANN model was applied to the test subset, with the output being a probability for each individual of the correct diagnosis being Type 1 Diabetes. A Receiver Operator Characteristic (ROC) curve was generated using these probabilities. For Sensitivity / Specificity analyses a threshold of 0.2 was applied to the probability value – if the probability was above this threshold a diagnosis of Type 1 Diabetes was deemed to be predicted.

In the case of the subset of patients presented to clinicians for classification, a confusion matrix was generated for both algorithmic and clinician classification outcomes, and Sensitivity, Specificity, Positive Predictive Value and Negative Predictive value were calculated.
METHODS

- SCI-Diabetes NHS Greater Glasgow and Clyde data refinement
  - Type 1 and Type 2 patients only

- Parameters selected:
  - BMI, Systolic BP, Diastolic BP, HbA1c levels, Age, Gender and Ethnicity.

Table 1: Patient Information Sheet Example

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>BMI</th>
<th>SBP</th>
<th>DBP</th>
<th>HbA1c</th>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>White - Scottish</td>
<td>25.83</td>
<td>118</td>
<td>76</td>
<td>84</td>
<td>78.49</td>
<td>male</td>
</tr>
</tbody>
</table>

Table 1 illustrates an example of the information sheet provided to physicians. The same information was then programmed into both the logistic regression and ANN. The information was given as shown to ensure no unfair bias was given to any model during diagnosis.
Table 1. Confusion Matrix Results

<table>
<thead>
<tr>
<th></th>
<th>Physician</th>
<th>ANN</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (CI 0.95)</td>
<td>0.86 (0.78, 0.92)</td>
<td>0.93 (0.85, 0.96)</td>
<td>0.91 (0.83, 0.96)</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.77</td>
<td>0.85</td>
<td>0.81</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.93</td>
<td>0.96</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table 1 illustrates the median confusion matrix values collected from the 6 physician forms analysed. The table details accuracy (CI 0.95), specificity and sensitivity for each model which was then utilised in the development of the ROC curve analysis.
Figure 2 illustrates a full dataset analysis of the ANN model (left) and the logistic regression (right). On each model relevant performance of the physician (red dot) versus the model (blue dot) are also plotted to illustrate model superiority.
Figure 3. ANN and patient characteristic association

Table 2. Statistical Summary

<table>
<thead>
<tr>
<th>Factors</th>
<th>General cohort</th>
<th>Type 1 patients</th>
<th>Type 2 patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort no.</td>
<td>49,995</td>
<td>3,222</td>
<td>46,773</td>
</tr>
<tr>
<td>Sex</td>
<td>F: 22,124</td>
<td>F: 1,388</td>
<td>F: 20,736</td>
</tr>
<tr>
<td></td>
<td>M: 27,871</td>
<td>M: 1,834</td>
<td>M: 26,037</td>
</tr>
<tr>
<td>Mean Age (years)</td>
<td>56.7 (48.0, 67.2)</td>
<td>28.5 (15.1, 39.9)</td>
<td>58.6 (49.8, 67.8)</td>
</tr>
<tr>
<td>Mean BMI</td>
<td>31.8 (27.2, 35.6)</td>
<td>23.4 (19.5, 26.4)</td>
<td>32.4 (27.8, 35.9)</td>
</tr>
<tr>
<td>Mean hba1c</td>
<td>66.0 (49.0, 79.0)</td>
<td>86.1 (63.0, 107.0)</td>
<td>64.9 (49.0, 77.0)</td>
</tr>
<tr>
<td>Mean sbp</td>
<td>136.5 (124.0, 147.0)</td>
<td>120.8 (110.0, 130.0)</td>
<td>137.5 (125.0, 148.0)</td>
</tr>
<tr>
<td>Mean dbp</td>
<td>80.2 (72.0, 87.0)</td>
<td>72.6 (63.0, 80.0)</td>
<td>80.7 (73.0, 88.0)</td>
</tr>
</tbody>
</table>

Figure 3 illustrates the association between examined patient characteristics and ANN model predicted diagnosis. BMI (top left), HbA1c (top right), Age (bottom left) and SBP (bottom right) are illustrated.
problem

are labels accurate?

<table>
<thead>
<tr>
<th>Factors</th>
<th>General cohort</th>
<th>Type 1 patients</th>
<th>Type 2 patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort no.</td>
<td>49,995</td>
<td>3,222</td>
<td>46,773</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F: 22,124</td>
<td>F: 1,388</td>
<td>F: 20,736</td>
<td></td>
</tr>
<tr>
<td>M: 27,871</td>
<td>M: 1,834</td>
<td>M: 26,037</td>
<td></td>
</tr>
<tr>
<td>Mean Age (years)</td>
<td>56.7 (48.0, 67.2)</td>
<td>28.5 (15.1, 39.9)</td>
<td>58.6 (49.8, 67.8)</td>
</tr>
<tr>
<td>Mean BMI</td>
<td>31.8 (27.2, 35.6)</td>
<td>23.4 (19.5, 26.4)</td>
<td>32.4 (27.8, 35.9)</td>
</tr>
<tr>
<td>Mean hba1c</td>
<td>66.0 (49.0, 79.0)</td>
<td>86.1 (63.0, 107.0)</td>
<td>64.9 (49.0, 77.0)</td>
</tr>
<tr>
<td>Mean sbp</td>
<td>136.5 (124.0, 147.0)</td>
<td>120.8 (110.0, 130.0)</td>
<td>137.5 (125.0, 148.0)</td>
</tr>
<tr>
<td>Mean dbp</td>
<td>80.2 (72.0, 87.0)</td>
<td>72.6 (63.0, 80.0)</td>
<td>80.7 (73.0, 88.0)</td>
</tr>
</tbody>
</table>
problem / opportunities

1 to clean existing data

2 to investigate a data-driven, time series based classification

3 to find ‘missing’ diagnoses (MODY etc)
unsupervised approach

self organising map

ANN that allows dimensionality reduction

eg
2 dimensional representation of multidimensional input

in this context – can make assessment of probability of the correct label being applied
combined supervised / unsupervised approaches

+ add human into the loop

clean existing data
input dataset

supervised analysis algorithm 1

prediction 1

unsupervised analysis algorithm 2

assess likely accuracy of prediction 1

above threshold

output classification

below threshold

feedback & relabel for re-training

'clinical' prediction 2

send to reviewer

clean existing data
data-driven, time series based classification
Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables


Figure 2: Cluster characteristics in the ANDIS cohort
Distributions of HbA1c, age at diagnosis, and BMI, HOMA2-B, and HOMA2-IR at registration, in the ANDIS cohort for each cluster. k-means clustering was done separately for men and women: pooled data are shown here for clusters 2-5. SAD=severe autoimmun diabetes. SIRD=severe insulin-deficient diabetes. SIRD=severe insulin-resistant diabetes. MOD=mild obesity-related diabetes. MARD=mild age-related diabetes. HOMA2-B-homoeostatic model assessment 2 estimates of β-cell function. HOMA2-IR-homoeostatic model assessment 2 estimates of insulin resistance. ANDIS=All New Diabetics in Scania.
2. clustering approach using TS data

- HbA1c
- BP
- BMI
- biochem
- drugs / drug response

Example: pattern mining for BMI trajectory analysis (THIN data)
3. identify missing diagnoses

difficult as small sample size to work from (but may be confident in label accuracy)

upsampling ideal to maximise value from data

synthetic data

potential to help with another problem – data sharing
requirements for synthetic data generator

- generate multi-dimensional time series data
- reflect distributions of individual parameters
- include interactions/associations between parameters over time
- allow training models on synthetic data that will perform well on real data
Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio

Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

Abstract

We propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model $G$ that captures the data distribution, and a discriminative model $D$ that estimates the probability that a sample came from the training data rather than $G$. The training procedure for $G$ is to maximize the probability of $D$ making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions $G$ and $D$, a unique solution exists, with $G$ recovering the training data distribution and $D$ equal to $\frac{1}{2}$ everywhere. In the case where $G$ and $D$ are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitative evaluation of the generated samples.

arXiv:1406.2661
structure of the given data, without specifying a target value. Generative models learn the intrinsic distribution function of the input data $p(x)$ (or $p(x,y)$ if there are multiple targets/classes in the dataset), allowing them to generate both synthetic inputs $x'$ and outputs/targets $y'$, typically given some hidden parameters.
Differentially Private Generative Adversarial Network

Liyang Xie1, Kaixiang Lin1, Shu Wang2, Fei Wang3, Jiayu Zhou1

1Computer Science and Engineering, Michigan State University
2Department of Computer Science, Rutgers University
3Department of Healthcare Policy and Research, Weill Cornell Medical School
[xieyan,linkaixj]@msu.edu,sw498@cs.rutgers.edu,few2001@med.cornell.edu,jiayuzhou@msu.edu

ABSTRACT
Generative Adversarial Network (GAN) and its variants have recently attracted intensive research interests due to their theoretical foundation and excellent empirical performance in generative models. These tools provide a promising direct solution to the problem where data availability is limited. One concern is that the defense of the learned generative model could concentrate on the training data points, meaning that it can easily remember training samples due to the high density of deep networks. This becomes a major concern when GANs are applied to private or sensitive data such as medical records, and the concentration of distribution may lead to sensitive critical patient information. To address this issue, in this work, we propose a differentially private GAN (DPGAN) model that can achieve differential privacy in GANs by adding careful noise to gradients during the learning procedure. We prove that our method can generate high-quality data points that are both private and available.

CSC CONCEPTS
• Computing methodologies → Neural networks.
• systems organization → Neural networks.
• Security → Privacy-preserving protocols.

KEYWORDS
Deep Learning; Differential Privacy; Generative model

REAL-VALUED (MEDICAL) TIME SERIES GENERATION WITH RECURRENT CONDITIONAL GANS

Stephanie L. Hyland1
ETH Zurich, Switzerland
Tri-Institutional Training Program in Computational Biology and Medicine, Weill Cornell Medical College
stephanie.hyland@inf.ethz.ch

Cristóbal Esteban2
ETH Zurich, Switzerland
cristobal.esteban@inf.ethz.ch

Gunnar Rätsch2
ETH Zurich, Switzerland
ratsch@inf.ethz.ch

ABSTRACT
Generative Adversarial Networks (GANs) have shown remarkable success in training models to produce realistic-looking data. In this work, we propose a Recurrent GAN (RGAN) and Recurrent Conditional GAN (RCGAN) to produce realistic real-valued multi-dimensional time series, with an emphasis on their application to medical data. RGANs make use of recurrent neural networks (RNNs) in the generator and the discriminator. In the case of RCGANs, however, these RNNs are conditioned on auxiliary information. We demonstrate our model on a set of toy datasets, where we observe visually and quantitatively (using saliency heatmaps) that they can successfully generate realistic time-series data. We also describe novel evaluation methods for GANs, as we generate a synthetic labelled training dataset, and evaluate on real test data. Finally, we demonstrate that the performance of a model trained on the synthetic data, and vice-versa, illustrates with these metrics that RCGANs can generate time-series data for supervised training, with only minor degradation in performance on real data. This is demonstrated on digit classification from ‘serialised’ MNIST at training an early warning system on a medical dataset of 17,000 patients from intensive care unit. We further discuss and analyse the privacy concerns that arise when using RCGANs to generate realistic synthetic medical time series and demonstrate results from differentially private training of the RCGAN.

Privacy-preserving generative deep neural networks support clinical data sharing

Authors: Brett K. Beaulieu-Jones1, Zhiwei Steven Wu2, Chris Williams3, Casey S. Greene3a

1Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
2Computer and Information Science, School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia, Pennsylvania, USA
3Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

*To whom correspondence should be addressed: csgreene@upenn.edu

One Sentence Summary: Deep neural networks can generate shareable biomedical data to allow reanalysis while preserving the privacy of study participants.

Abstract: Though it is widely recognized that data sharing enables faster scientific progress, the sensible need to protect participant privacy hampers this practice in medicine. We train deep neural networks that generate synthetic subjects closely resembling study participants. Using the SPRINT trial as an example, we show that machine-learning models built from simulated participants generalize to the original dataset. We incorporate differential privacy, which offers strong guarantees on the likelihood that a subject could be identified as a member of the trial. Investigators who have compiled a dataset can use our method to provide a freely accessible public version that enables other scientists to perform discovery-oriented analyses. Generated data can be released alongside analytical code to enable fully reproducible workflows, even when privacy is a concern. By addressing data sharing challenges, deep neural networks can facilitate the rigorous and reproducible investigation of clinical datasets.
relative time-series data

synthetic time-series data

differential privacy inhibits data leakage

- quality vs security tradeoff
- privacy budgets
- computational limits
beaulieu-jones et al 2017. doi: https://doi.org/10.1101/159756
GAN vs true generated systolic BP distributions – 100 epochs with differential privacy
single time point, 100 ‘IDs’
GAN vs true generated HbA1c distributions – 100 epochs with differential privacy
single time point, 100 ‘IDs’
potential uses of GANs generated synthetic data

• balance classes (upsampling)
• with dp implemented – allow data sharing

• with further development can be used as classifier/regressors – may provide a general solution to all problems discussed
glucose.ai group, NHS GGC
Greg Jones
Debbie Morrison
Sean Harbison

myDiabetesIQ
Debbie Wake
Scott Cunningham
Nicky Conway
Sam Philip
Felix Agakov

Stratified Medicine MSc students
Ruth Muir
Alastair Irvine
Halah al Saadi

thinkingAI group, University of Birmingham
Krish Nirantharakumar
Krishna Gokhale
Tom Taverner
Peter Tino