Mansoura University

Faculty of Computers and Information Sience

Machine learning algorithms in big data analyses identify determinants of insulin gene transcription

By Ziyad Abdulrahman Mustafa, Supervised by Prof.Sara El-Sayed El-Metwally.

Abstract

This paper will show analyzing over 490,000,000 data points to compare 10 different ML algorithms in a large (N=11,652) training dataset of human pancreatic single-cell transcriptomes to identify features (genes) associated with the presence or absence of insulin transcript(s). Prediction accuracy/sensitivity of models were tested in a separate validation dataset (N=2,913) and the performance of each ML-workflow assessed. Random Forest ML algorithm delivered high predictive power in a receiver operator characteristic (ROC) curve analysis at the highest sensitivity (0.98), compared to other algorithms. The top-10 features, (including IAPP, ADCYAP1, LDHA and SST) common to the three Ensemble ML workflows were significantly dysregulated in scRNA-seq datasets from Ire-1$\alpha^{\beta-/-}$ mice that demonstrate de-differentiation of pancreatic β-cells as well as in pancreatic single cells from individuals with Type 2 Diabetes.

Used Algorithms

ML Algorithm	Build	Description
Random Forest	Ensemble	Random Forest classifier is a meta-estimator that fits several decision trees on various sub-samples of datasets and uses an average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement. Reduction in over-fitting and Random Forest classifier is more accurate than decision trees in most cases.
Gradient Boosting	Ensemble	The idea behind "gradient boosting" is to take a weak hypothesis or weak learning algorithm and make a

		series of tweaks to it that will improve the strength of the hypothesis/learner. This type of Hypothesis Boosting is based on the idea of Probability Approximately Correct Learning (PAC).
Adaboost	Ensemble	For AdaBoost, many weak learners are created by initializing many decision tree algorithms that only have a single split. The instances/observations in the training set are weighted by the algorithm, and more weight is assigned to instances that are difficult to classify. More weak learners are added into the system sequentially, and they are assigned to the most difficult training instances. In AdaBoost, the predictions are made through majority vote, with the instances being classified according to which class receives the most votes from the weak learners.
Ridge Classifier	Regularization	Ridge classifier first converts binary targets to [-1, 1] and then treats the problem as a regression task, optimizing the same objective as above. The predicted class corresponds to the sign of the regressor's prediction. For multiclass classification, the problem is treated as multioutput regression, and the predicted class corresponds to the output with the highest value.
Logistic Regression	Regression	Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function. It is most useful for understanding the influence of several independent variables on a single outcome variable.
Naive Bayes	Bayesian	Naive Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many realworld situations such as document classification and spam filtering. This algorithm requires a small amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely fast compared to more sophisticated methods.
Decision Tree Classifier	Decision Tree	Given data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data. Decision Tree is simple to understand and visualise, requires little data preparation, and can handle both numerical and categorical data
K-Nearest Neighbours	Instance-based	Neighbours based classification is a type of lazy learning as it does not attempt to construct a general internal model but simply stores instances of the training data. Classification is computed from a simple

		majority vote of the k nearest neighbours of each point. This algorithm is simple to implement, robust to noisy training data, and effective if training data is large.
Linear Discriminant Analysis (LDA)	Dimensionality Reduction	It is a classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule. The model fits a Gaussian density to each class, assuming that all classes share the same covariance matrix. The fitted model can also be used to reduce the dimensionality of the input by projecting it to the most discriminative directions, using the transform method.
Linear Support Vector Classifier	Support Vector Machines (SVM)	Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Result

Ensemble ML workflows to identify genes associated with insulin transcription.

The scRNA-seq datasets obtained from public databanks of human pancreatic single-cell transcriptomes were classified as insulin-transcribing (1) or those with no insulin (0) As described earlier, all the three Ensemble ML workflows presented with an AUC that was better than any of the other ML workflows tested in our ROC curve analysis. Ensemble workflows also presented with high accuracy (>87%), precision (>0.89), and sensitivity (>0.95), which was comparable to other popular workflows such as logistic regression. As Ensemble ML workflows such as Random Forest use a collection of decision trees (forest), we decided to compare the performance of the top three (Ensemble) workflows to a single (Decision tree) algorithm. The relative contribution of the top 10 features (genes) from each of these ML workflows are presented as radar plots. To compare the expression of these features (genes) identified through each of the Ensemble and Decision Tree classifier, we examined the expression of genes identified to be associated with insulin transcription to those in a separate islet β-cell datasets.

Insulin-associated genes are dysregulated during β-cell dedifferentiation.

Dedifferentiation of β-cells, characterized by the loss of expression of key β-cell maturation marker genes with an accompanying reduction in insulin secretion, has been observed in mouse models of type 1 (T1D) and type 2 (T2D) diabetes, as well as in individuals with diabetes. We questioned if the expression of gene variables identified and validated (in silico) as being predictive of insulin gene transcription are dysregulated in a mouse model of T1D with evidence of islet dedifferentiation. Transient dedifferentiation of (which was not certified by peer review) is the author/funder. islet β-cells was recently reported in an established T1D preclinical mouse model upon β-cell-specific deletion of a key stress response gene, Ire1α, (Ire1α$^{\beta-/-}$) . These mice also demonstrated reduced β-cell number as well as diminished expression of insulin transcripts in β-cells compared to control (Ire-1α$^{fl/fl}$) mice. Therefore, we evaluated the expression of 25 gene transcripts that made up the top-10 features across the four different ML. Twelve of these features were not significantly regulated between Ire1α$^{\beta-/-}$ and Ire1α$^{fl/fl}$ islets. However, the remaining thirteen features were significantly dysregulated in β-cells of Ire-1α$^{\beta-/-}$ mice that were undergoing dedifferentiation. De-differentiating β-cells showed significant downregulation of five key genes; Iapp, MafA, Pcsk1n, Atp5e and Ldha, whilst all other insulin-associated gene transcripts showed significantly higher levels. In Type 2 diabetes (T2D) it is known that INS transcript expression is reduced, therefore we validated the top gene features common (IAPP, SST, MAFA, ADCYAP1 and LDHA) in three of the ML workflows analysed using a separate publicly available single-cell RNA-seq dataset from non-diabetic (ND) vs T2D adult human pancreas (GSE154126). Four of the five genes (IAPP, SST, MAFA, ADCYAP1), were significantly lower in T2D insulin-transcribing cells compared to ND insulin-transcribing cells.