Machine Learning Models

Abstract:

Introduction- The aim of this practical was to build machine learning models using metabolomic data from different sample types to diagnose Crohn's disease. The two machine learning models used for each sample type were support vector machine (SVM) and random forest (RF) classifier.

Method- Data from all sample types (blood, breath, urine and faeces) were imported, read and wrangled into a panda's data frame. A bootstrap using SVM and RF was done for each sample. The SVM and RF models were then tested to see the predictions they gave and its average accuracy score. A histogram was plotted to see distribution accuracy across bootstraps. Specificity and sensitivity were calculated. Then a t-test was done comparing the RF and SVM classifier for each sample type. Lastly permutation testing was done on the model with highest accuracy to see how the model performed compared to random chance. A t-test was done once more to see if the chosen model performed significantly better than random chance.

Results and discussion- Out of all sample types, faeces had the best average accuracy of 76%, then urine with 58% accuracy, then breath with 52% accuracy and lastly, the worst performing sample type was blood with an average accuracy of 46%. In the faeces and breath sample, the RF classifier was significantly better than the SVM classifier (student's t-test, $p<0.01$). The urine and blood sample types showed no significant differences between the classifiers. The chosen model was RF classifier using the faeces sample with 86% accuracy, when compared to random chance it proved to be significantly better (student's t-test, $p>0.01$).

It is likely the faeces sample performed the best because Crohn's disease affects the gastrointestinal tract. Therefore, there are more metabolomic differences between Crohn's disease and healthy samples in faeces compared to other sample types. More metabolomic differences between diseased and control samples makes it easier for machine learning models to accurately classify and diagnose samples of

an unknown class. The RF classifier was also significantly better than SVM in two sample types, this could be because SVM is more suited to smaller data sets that are linearly separable whereas RF is suited to larger dataset, with variables going into the thousands as our data did.

Conclusion- The best sample type was faeces when training machine learning models to diagnose Crohn's disease. RF classifier is also better suited to large datasets compared to SVM. The most accurate model also performed significantly better than it would by random chance.

Introduction:

The aim of this practical was to build accurate machine learning models using metabolomic data from different sample types and use these models to diagnose Crohn's disease. Crohn's disease is an inflammatory bowel disease that causes inflammation of the gastrointestinal tract. Studies have shown there are notable difference in metabolomic profiles in patients with Crohn's disease compared to healthy patients. Metabolomic data was collected from diseased and controlled patients from the blood, faeces, urine, and breath using gas chromatography mass spectrometry (GCMS). During GCMS, the samples are first heated to release volatile compounds. They are then detected by mass spectrometry which determines the mass of the metabolites as they elude from the chromatography column. Further fragmentation is then done to give more specific information about the structure of the metabolites. Using information about the mass and structure, the metabolites can then be identified.

The two machine learning models used in this experiment were support vector machine (SVM) and random forest (RF) classifier. SVM model works by creating an optimal margin boundary between classes. Random forest classifier contains many decision trees which give a class prediction, each class is predicted by majority voting of these trees.

Methods:

The first dataset explored was from the faecal sample set. From the available data archive (gcms_data_zipped.zip), 'BWG_FA_CDvCTRL.mat' was uploaded, read, and wrangled into a panda’s data frame. The training data was made and assigned to a variable. The panda’s data frame was then transposed and assigned to another variable; this was our test data.

Bootstrap method using SVM was carried out. In which the data was split into 70% to train our model and 30% to test our model. 100 bootstraps were carried out, with the data randomly split into test and training data differently each iteration. An average accuracy score of the classifier was produced and a histogram was plotted to show the accuracy distribution across 100 bootstraps. A confusion matrix was also created showing the predictions the model made.

Bootstrap evaluation was carried out again for the same sample, in the exact same way with the only difference being that a RF classifier was used to fit the model instead of SVM. Once again, an average accuracy score was produced, as was a histogram showing accuracy distribution across 100 bootstraps and a confusion matrix.

All the above steps were carried out again using data ‘'BWG_BL_CDvCTRL.mat' for blood sample type, 'BWG_UR_CDvCTRL.mat' for urine sample type and 'BWG_BR_CDvCTRL.mat' for breath sample type.

The specificity and sensitivity were calculated for the SVM and RF models for each sample type using the following formula:

$$Specificty = no.of\ true\ negatives \div (no.of\ true\ negatives + no.of\ false\ positives)$$

$$Sensitivty = no.of\ true\ positives\ \div (no.of\ true\ positives + no.of\ false\ negatives)$$

In addition, for each sample type, the SVM and RF classifiers were compared to see which one was more accurate using Student’s t-test.

$$Ttest = \ m - \mu \div (s \div \sqrt{n})$$

$$m = mean$$

$$\mu = theoretical\ value$$

$$s = standard\ deviation$$

$$n = number\ of\ samples$$

Lastly, permutation testing was done to see how the chosen model (the one with the highest accuracy) compared to random chance. The training data for this sample type was randomly shuffled and a bootstrap of the classifier was done in the same way as previously. The key difference in this case was that the training data was replaced with the randomly shuffled training data. An average accuracy score was produced, as was a histogram showing accuracy distribution across 100 bootstraps and a confusion matrix. Specificity and sensitivity were also calculated and finally a t-test was performed to see how the chosen model compared to random chance.

Results and discussion:

Table 1- table showing SVM and RF classifier percentage accuracy in diagnosing Crohn’s disease and which model (if any) gives better accuracy (determined using student’s t-test statistical analysis).

Sample type	SVM (% accuracy)	RF (% accuracy)	Average accuracy	More statistically accurate model (SVM/RF/Neither) (p-value)
Faeces	65%	86%	76%	RF (p <0.01)
Specificity	100%	100%		
Sensitivity	50%	100%		
Urine	58%	57%	58%	Neither (p= 0.67)
Specificity	100%	83%		
Sensitivity	0%	0%		
Blood	46%	43%	46%	Neither (p= 0.08)
Specificity	100%	0%		
Sensitivity	0%	40%		

Breath	41%	63%	52%	RF (p <0.01)
Specificity	0%	80%		
Sensitivity	100%	17%		

For sample type ‘faeces’, the accuracy of the SVM classifier was 65% and the accuracy of the RF classifier was 86%. The RF classifier was statistically more accurate compared to the SVM classifier (student’s t-test, p-value<0.01). The urine sample type had a 58% accuracy for the SVM classifier, the RF classifier was 57%. Unlike the ‘faeces’ sample type, there was no significant difference between the two classifiers (student’s t-test, p=0.67). The blood sample type has 46% accuracy for the SVM classifier and 43% accuracy for the RF classifier. There was no difference in performance between the classifiers in this sample type (student’s t-test, p= 0.08). Lastly for the breath sample type, the accuracy of the SVM classifier was one of the lowest, with an average accuracy of 41%, the RF classifier was significantly better for this sample (student’s t-test, p>0.01) with an average accuracy of 63%.

From all the sample types, the one with the worst average accuracy percentage was ‘blood’ with 46%, then it was ‘breath’ (52%), then ‘urine’ (58%) and lastly the one with the best accuracy was ‘faeces’ with an average accuracy of 76%.

One of the possible reasons that the faecal sample had the best accuracy in predicting Crohn’s disease is because this disease affects the gastro-intestinal tract which can affect any part of the small intestine, large intestine and colon in which faeces is formed. The inflammation caused by this disease affects how the body digests food, absorbs nutrients, and gets rid of faecal waste. One study found the metabolic profile of faeces in patients with IBD (including Crohn’s disease) compared to healthy controls showed to have important metabolic biomarkers of disease related changes (De Preter et al., 2014). It is very much possible that because the faeces of a Crohn’s disease patient is more metabolically different to a healthy patient, it made it easier for the SVM and RF machine learning models to recognise different metabolomic patterns between the diseased and healthy samples and predict test samples accurately based on learned differences and patterns of the metabolome.

The findings of one study showed of 302 metabolites associated with lipid metabolism, 54% were significantly altered in Crohn's disease compared with a healthy subject (Scoville et al., 2017). Lipid metabolism takes place largely in the small intestine where fecal waste passes through. However, it is slightly unclear why the classifiers for the other sample types didn't perform as well when many studies have found metabolomic differences in blood, breath and urine in Crohn's disease patients compared to controls. One implication is that although the metabolome has proved to be affected in all the sample types in other studies, the faecal sample may contain the most metabolomic differences simply because of the nature of this disease and the fact that it primarily affects the intestines and colon.

Table 2- Table comparing RF classifier for faeces and RF permutation test for faeces in terms of accuracy, specificity, sensitivity and significance

Classifier type	Accuracy	Specificity	Sensitivity	T-test
RF classifier for faeces	86%	100%	100%	P<0.01
RF permutation test for faeces	57%	60%	50%	

Overall, the RF forest classifier for the faeces sample type had the highest accuracy of 86%, therefore it was subject to permutation testing to see how it compared to random chance. Table 2 shows the chosen classifier (RF classifier for faeces sample type) is significantly more effective at accurately predicting diagnosis than random chance.

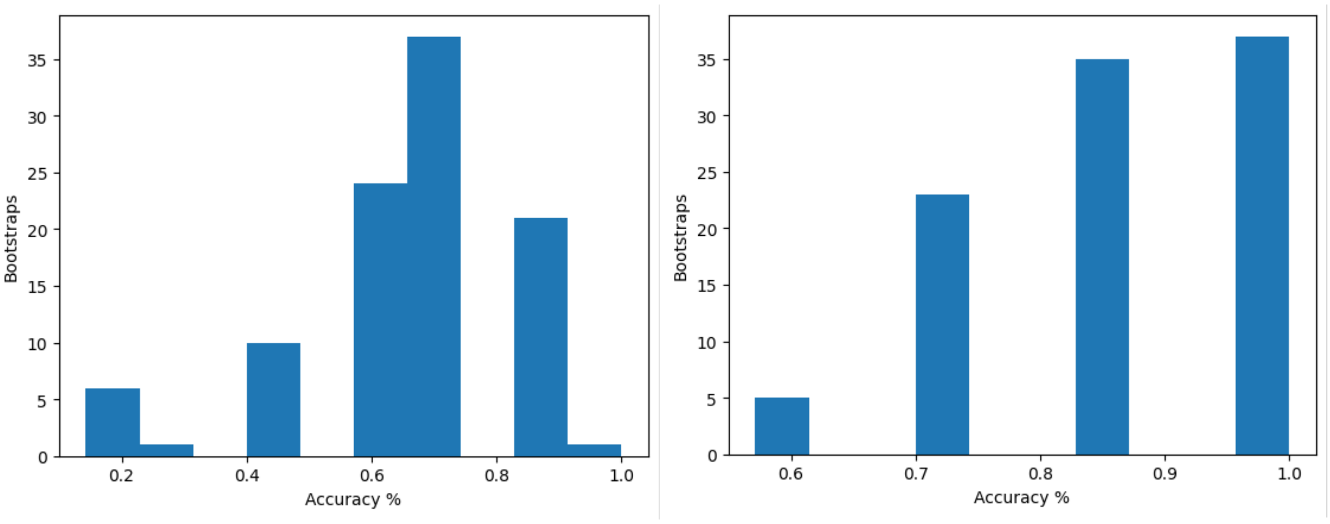

Figure 1- Accuracy distribution across 100 bootstraps for SVM (left) and RF (right) models in faeces sample set.

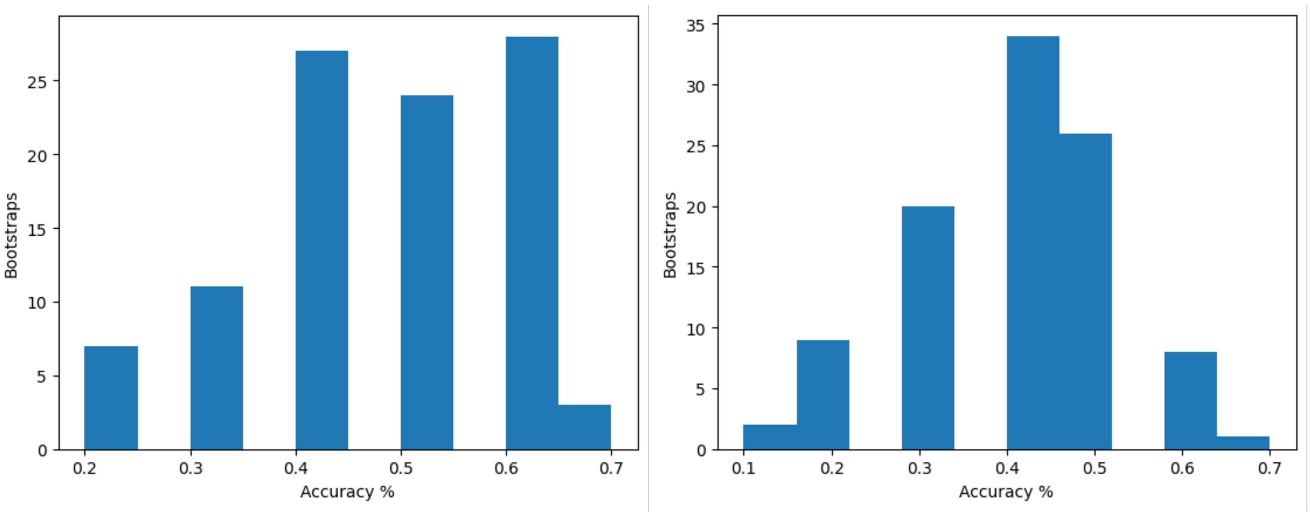

Figure 2- Accuracy distribution across 100 bootstraps for SVM (left) and RF (right) models in blood sample set.

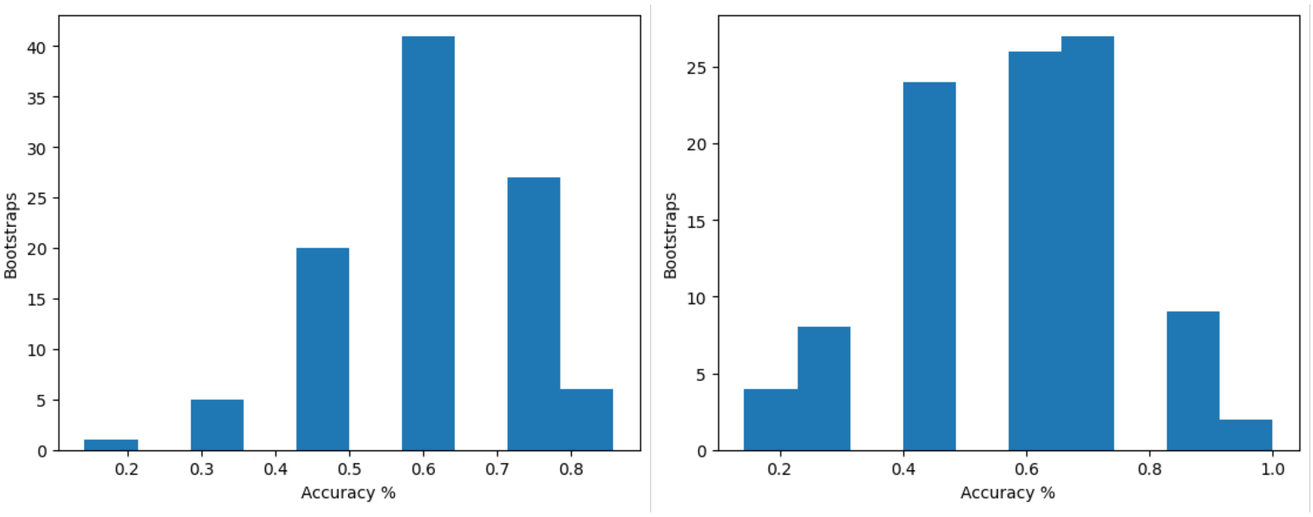

Figure 3- Accuracy distribution across 100 bootstraps for SVM (left) and RF (right) models in urine sample set.

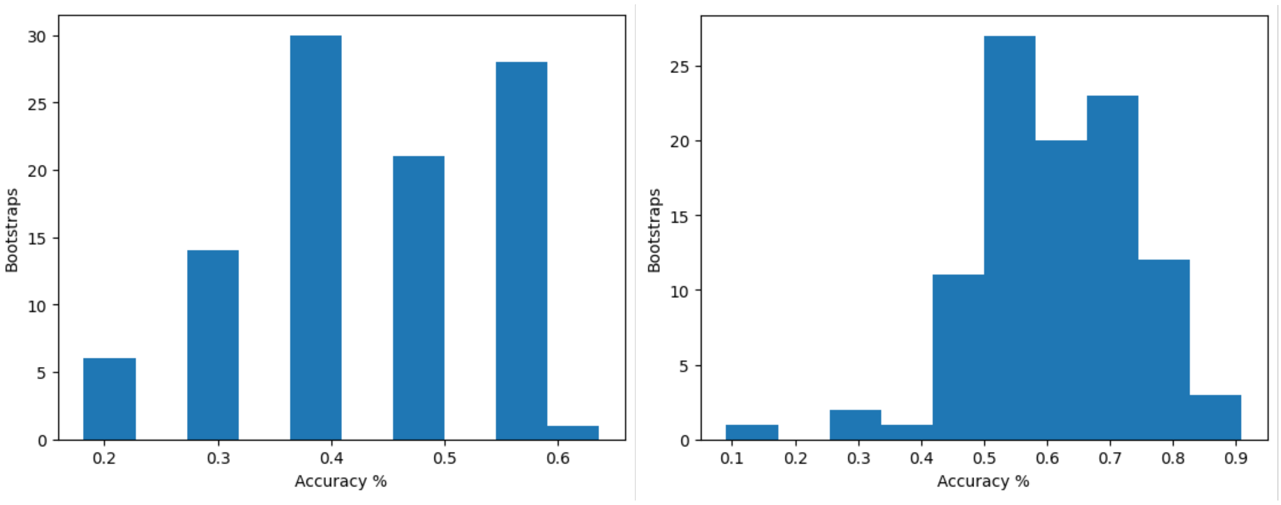

Figure 4- Accuracy distribution across 100 bootstraps for SVM (left) and RF (right) models in breath sample set.

Figures 1-4 are a visual representation of the accuracy distribution for the SVM and RF classifiers. The distribution is across 100 bootstraps and the accuracy scores given in table 1 is the average of the accuracy scores of 100 bootstraps.

In table 1 a statistical difference between SVM and RF classifiers was seen in ‘faeces’ (t-test, p-value<0.01) and ‘breath’ (t-test, p-value<0.01) sample types only, with RF being the more accurate classifier in both sample types. There was no significant difference between SVM and RF in sample types ‘urine’ (student’s t-test, p-value=0.67) and ‘blood’ (student’s t-test, p-value=0.08).

SVM is a great method for classification and regression although it did not perform as well as RF in this experiment. SVM works well with smaller data sets, with few outliers. However, when target classes overlap and data is not linearly separable, prediction accuracy drops and requires some modifications such as kernel tricks to perform well. For these reasons accuracy may have been low for the SVM model in this experiment. RF classifier makes predictions using the decision of multiple individual decision trees and merges then together for an accurate prediction. Unlike SVM, RF works better with a larger number of variables in the dataset, going into the thousands and the number of variables in our data set was ~ 4300. For this reason, RF may have outperformed SVM because the dataset was more suited to a RF classifier rather than an SVM classifier.

Conclusion:

In conclusion, a faecal sample is the best sample type to look at when comparing metabolomes in Crohn’s disease patient’s vs healthy patients and is therefore the best sample type to train a machine learning model to accurately diagnose Crohn’s disease. The RF classifier was also more suited for this dataset and was significantly more accurate in predicting disease compared to the SVM classifier in faeces and breath sample types (with urine and blood sample types showing no significant difference between classifier accuracy). The chosen model (RF for faeces sample type) also showed to perform significantly better than random chance.

References:

De Preter, V. *et al.* (2014) “Faecal metabolite profiling identifies medium-chain fatty acids as discriminating compounds in IBD,” *Gut*, 64(3), pp. 447–458. Available at: https://doi.org/10.1136/gutjnl-2013-306423.

Scoville, E.A. *et al.* (2017) “Alterations in lipid, amino acid, and energy metabolism distinguish crohn’s disease from ulcerative colitis and control subjects by serum metabolomic profiling,” *Metabolomics*, 14(1). Available at: https://doi.org/10.1007/s11306-017-1311-y.