Vilnius University
Faculty of Medicine

Report

Analyses of mice protein expression data using machine learning algorithms

Prepared by
Laima LUKOŠEVIČIŪTĖ
Student of Systems Biology
Autumn semester

Prepared for
Data Mining
Doc. Dr. E. PRANCKEVIČIENĖ
Lekt. V. TOMKUTĖ

2019

Contents

1 Introduction

Machine learning (ML) algorithms are used in a wide variety of applications, such as computer vision or self-driving cars, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task. ML algorithms perform a specific task without using explicit instructions. They rely on patterns and inference instead. Nowadays it is getting more and more popular to apply ML algorithms to analyse biological data.

One great example is an article "Self-Organizing Feature Maps Identify Proteins Critical to Learning in a Mouse Model of Down Syndrome" (Higuera, Gardiner & Cios 2015). In this paper the Self-Organizing Feature Maps approach identified reduced subsets of proteins predicted to make the most critical contributions to normal learning, to failed learning and rescued learning.

In this report I will try to train my ML models, using the same dataset, to predict whether the mice are of normal genotype or is trisomic.

2 Methods

2.1 Dataset

The dataset used here consists of the expression levels of 77 proteins/ protein modifications that produced detectable signals in the nuclear fraction of cortex. There were 38 control mice and 34 trisomic mice, for a total of 72 mice (7–10 mice in each of the eight groups. Measurements were made using reverse phase protein arrays (RPPA), a high throughput technique in which protein samples from individual mice are robotically spotted onto nitrocellulose-coated microscope slides. For experiments here, a single slide contained 20 spots per sample: three replicates of a five-point dilution series, plus replicate buffer controls, i.e., 15 measurements of each protein per sample. Therefore, for control mice, there were 38x15, or 570 measurements, and for trisomic mice, there were 34x15, or a total of 510 measurements, per protein. RPPA is highly sensitive and reproducible but technical artifacts can occur which require the elimination of data from individual spots. Similar to other high throughput techniques, it is not possible to repeat experiments for individual measurements and therefore the final dataset contains missing values, i.e., there were <15 measurements for a small number of samples/proteins measured.(Higuera et al. 2015)

2.2 Analyses

First of all, NA values in the dataset have been filled using Weighted Moving Average algorithm. Then data has been normalized, according to the formula below.

$$Norm(e_{ij}) = (e_{ij} - E_{j,min})/(E_{j,max}) - E_{j,min}))$$

In this report 4 ML algorithms have been chosen to compare. Two supervised learning algorithms: k-nearest neighbors algorithm (k-NN) and naïve Bayes classifier, and also two unsupervised learning algorithms: PCA and k-means.

With k-NN and naïve Bayes the algorithms were trained to classify dataset to two classes control and trisomic (Ts65Dn). Results were cross-validated by dividing dataset to 5 parts and using 4/5 of data to train and 1/5 to test the algorithms, with each part interchangeably.

PCA has been used to reduce the dimensionality of the data, and k-means has been applied to cluster the data into two clusters, hoping that mice of normal and trisomic genotype would differ significantly.

3 Results and discussion

3.1 K-Nearest Neighbors

From fig. 1 it is seen that the best value for k is 1, since it gives the mean accuracy of 1. The tendancy of the model is that the more neighbors it uses the accuracy gets worse. So k = 1 was applied and validation results of k-NN model can be seen in table 1.

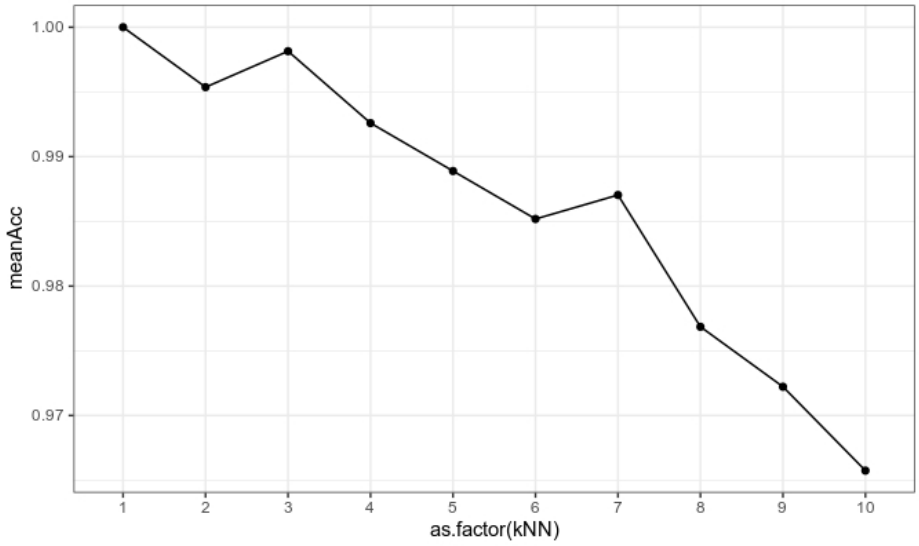

Figure 1: The accuracy of k-nearest neighbors algorithm depending on k.

	foldID	kNN	validationAccuracy
Accuracy	1	1	1
Accuracy1	2	1	1
Accuracy2	3	1	1
Accuracy3	4	1	1
Accuracy4	5	1	1

Table 1: The validation results for accuracy using k-NN model.

The model seems to work perfectly of the given data set. It would be interesting to test this model of another dataset and see how the results change.

3.2 Naïve Bayes

In case of Naïve Bayes classifier the best results that have been achieved in terms of accuracy is 81%. The cross-validation results can be seen in table 2. Confusion matrix and other statistics for the most accurate classifier can be seen in fig. 2

	foldID	validationAccuracy
Accuracy	1	0.7916667
Accuracy1	2	0.7824074
Accuracy2	3	0.7500000
Accuracy3	4	0.7314815
Accuracy4	5	0.8101852

Table 2: The validation results for accuracy using Naïve Bayes classifier.

The model seems to work quit well on the given data set. It would be interesting to test this model of another dataset and see how the results change.

```
[1] "foldID 5"
Confusion Matrix and Statistics

          Reference
Prediction Control Ts65Dn
   Control      75     15
   Ts65Dn       26    100

               Accuracy : 0.8102
                 95% CI : (0.7514, 0.8602)
    No Information Rate : 0.5324
    P-Value [Acc > NIR] : <2e-16
```

Figure 2: The confusion matrix and other statistics of best accuracy Naïve Bayes classifier.

3.3 Principal Component Analyses

Results of PCA can be seen in fig. 3. It is seen that first 4 PC would explain the most variation, the increse of variation in other components is quit small. Since we cannot plot 4D graph, the 3D representation can be seen in fig. 6. From fig. 6 it is seen that there is no distinct distribution between the points of mice with normal and trisomic genotypes.

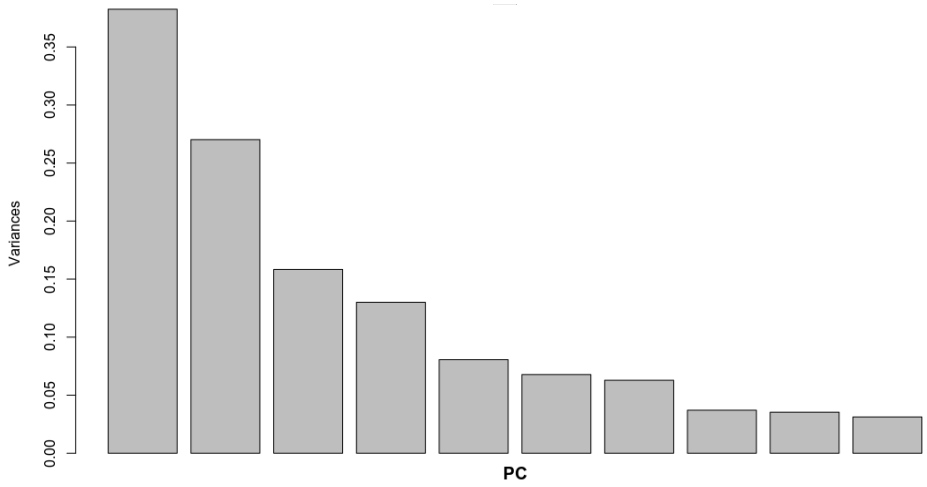

Figure 3: The variance explanation of each PC.

PCA does not work well, it is not possible to separate the points. In order to improve situation other dimensionality reduction algorithm could be implied.

3.4 K-means

The confusion table of k-means based model can be seen in **??**. Since the results are very bad, practically the model is guessing, then it was decided to do k-means on 4 main PC, since they explain the biggest amount of variation in the data. The confusion matrix can be seen in **??**. The results are still very bad. It would be advisable to try different clustering algorithm.

4 Conclution

To conclude the models of supervised learning worked better. Comparing k-NN and Naïve Bayes classifier, k-NN was more acurate then Naïve Bayes. Still, these models depends on the data, and the testing on different data set can actually prove how good they are.

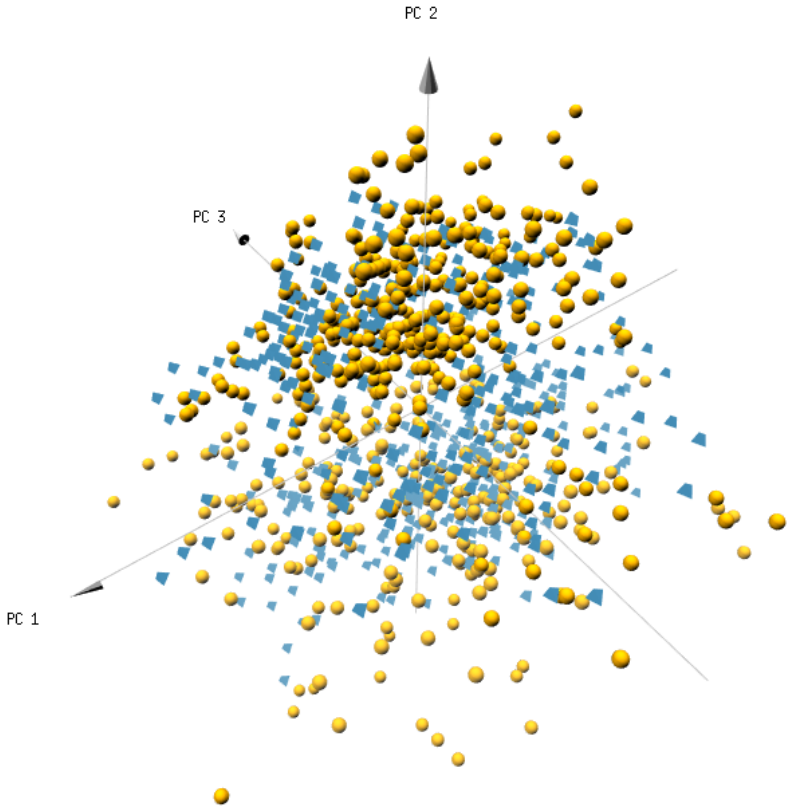

Figure 4: The 3D plot of 3 main PC that explains a lot of varience in the data.

	Control	Ts65Dn
1	329	255
2	241	255

Figure 5: The confusion matrix for k-means model.

	Control	Ts65Dn
1	322	259
2	248	251

Figure 6: The confusion matrix for k-means model using PC as variables.

References

Higuera, C., Gardiner, K. J. & Cios, K. J. (2015), 'Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome', *PloS one* **10**(6), e0129126.