Individual Project - Interim Report

Daren Sin (ds2912) | CID: 00732331

January 20, 2017

1 Introduction

1.1 Schizophrenia, etiology and genes

Schizophrenia is a complex mental disorder that displays an array of symptoms. It is commonly perceived that schizophrenia is a hereditary disease that can be passed down within the family, but some individuals diagnosed with schizophrenia do not have a family member with the disorder [1]. Thus, it is postulated that the heritability of schizophrenia might not be as high as what is commonly believed [2].

Furthermore, there is a strong indication that environmental factors - such as tobacco smoke and viruses - and genetic factors have an influence on the development of psychiatric disorders in an individual [1, 3]. This results in a hypothesis that the epigenetics - "any process that alters gene activity without changing the DNA sequence" - of an individual might have a role to play in the development of schizophrenia [4]. However, exactly how these two factors play a part is still unclear [5]. What we should also note about the current research on psychiatric disorders is that studies on such disorders do not receive as much attention as other illnesses such as cancer [6]. Thus, any insight generated from this project would be beneficial to helping us understand psychiatric disorders better.

Overall, this project aims to predict Schizophrenia cases on the basis of epigenetics and epivariations.

1.2 Using machine learning to predict Schizophrenia cases

Using data from a recent study on epigenetics and schizophrenia (see Section 2.4), the project aims to use machine learning to elucidate any statistical regularity in the data, in hope that any insight into the data can help geneticists and psychiatrists understand the etiology of schizophrenia - and indeed, other psychiatric disorders - better.

As a starting point, simple classifiers can be used on the data, to determine the classification accuracy of the data. Later on, we would then use more complex classifiers and techniques which might produce better results.

What make this endeavour interesting are the potential problems that we might face while drawing inferences from the data.

Previous work on using machine learning on biological data (see Section 2.1) has always been plagued with the “Curse of dimensionality”, where the number of biological samples is far lesser than the number of features (or dimensions) of the data. In our case, we have 847 samples (individuals) with 420374 features, resulting in about 2 gigabytes of data.

Here, we face a potential problem of a similar nature - the data has high dimensions, but not all the genes involved in the study would directly play a part in the classification of the disorder; some genes may only contribute a little to the outcome of the classification. In this case, we would need to perform feature/dimension reduction to only select features that have significant contribution to the classification outcome.

Moreover, linear classifiers may not be able to capture the complexity of the data, as it is hypothesised that subsets of genes - rather than single genes - contribute to the genesis of the disorder [3].

These potential problems make the project interesting, as we cannot simply use ordinary machine learning techniques to manipulate the data - we have to adapt our algorithms and classifiers to suit the complexity and context of the dataset and problem.

2 Background

2.1 Machine learning and cancer classification

There is a significant amount of literature on gene expression data and cancer classification. These works primarily aim to uncover biological or medical insights using biological data obtained from microarrays, which are tools to measure the gene expression of thousands of genes simultaneously [7]. For example, using neural networks, gene expression data can be used to distinguish between tumour types, which helps in solving cancer diagnosis problems [8, 9].

What is similar about this project and previous work on cancer classification is that the data for both cases are plagued with high dimensionality (“Curse of dimensionality”). For example, in cancer studies, microarrays are used, with a large number of genes (features) but a small number of samples (observations) [9]. Furthermore, only a (small) subset of the features are relevant for the studies, as not all genes are relevant for determining the type of cancer a patient has. This is known as biological noise [10]. As such, a feature/dimensionality reduction on the data has to be performed to select only the relevant genes/features for the classification problem. In other words, the solution for our situation (and also for cancer classification) would ideally be sparse, as we seek to identify the features are the most relevant to the classification.

However, what is fundamentally different about studies on psychiatric disorders and cancer, is that the latter is observable, such that we can know for sure that an individual has cancer using medical tools. However, it is not obvious that an individual has a psychiatric disorder, as its symptoms might not be as obvious or observable.

2.2 Existing work on psychiatric disorders

2.3 Review of machine learning classifiers

This section reviews machine learning classifiers that might be relevant for this project.

2.3.1 Decision Trees

In our context, the task is to classify the data according to whether a sample (individual) has a psychiatric disorder - in particular, schizophrenia - or not. In other words, the classification task is binary. An intuitive solution is to use decision trees; problems with discrete output values can be solved using decision trees [11].

A decision tree algorithm is capable of sorting the instances - in our context, samples with different features - down the tree until the algorithm reaches a leaf node, during which a classification is given to the node. At each level of the tree, the intermediate node is split according to some attribute. One variant of the decision tree algorithm is the ID3 algorithm [12]. The ID3 algorithm makes use of a statistical quantitative measure, the information gain, to determine the attribute to classify the samples with. Using definitions from [11], let S be the set of all the samples that we want to classify at a particular node. The samples can also be separated into two groups, those with a positive classification and those with a negative classification. Define the entropy of S as:

$$Entropy(S) \equiv -p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$$

where $p_{(+)}$ and $p_{(-)}$ represents the proportion of samples with positive and negative classification respectively.

Then, the information gain with respect to the set S and an attribute (feature) A is defined as:

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where $Values(A)$ is the set of all possible values of attribute A, and $|S_v|$ is the number of elements in the set S with value v for its attribute A.

We then classify the samples in the node according to the attribute with the highest information gain. Intuitively, we want to choose the attribute that can give the most distinct separation between the positive and negative classification (instead of choosing an attribute that, say, splits the samples into half according to their classification).

Although the decision tree algorithm is said to be robust to errors [11] and the resulting decision tree can be easily interpreted by humans, it might be difficult to classify samples according to features that are highly correlated. This also happens to be a potential characteristic of our dataset, and we would expect some of the features to be correlated.

2.3.2 Random forest

The random forest method [13], a form of "ensemble learning", is an extension of the decision tree algorithm described above, and it has been used in areas such as multi-class object detection in images [14]. Overall, a random forest algorithm can be outlined as such:

- Split the dataset into distinct subsets.
- Using each subset, train a decision tree using a relevant algorithm, such as the ID3, as outlined above.
- Suppose we have an unseen sample $\boldsymbol{x}$. Put the $\boldsymbol{x}$ through each tree, and obtain the resulting classification for each tree.
- Based on a "majority vote" system, determine the final classification of $\boldsymbol{x}$; that is, choose the classification that is the most popular among the decision trees.

Even though random forests have been shown to outperform decision trees [15], the limitations of decision trees as described above would still be inherent in the random forest method. Besides, Random Forest requires more parameters in general. For example, we would need to determine the number of trees to be trained, and the optimal number of trees to be trained is not obvious, as we would still need to perform numerical experiments to ascertain this optimal number.

Since the number of trees grow with the number of features that directly affect our classification (predictors) [16], and we do not know beforehand what these features are, we might potentially have to train a lot of trees. This might take up a lot of memory and time, and might also be computationally expensive.

So, overall, the decision tree and random forest methods might not be the best methods for our context, even though they are considered to be popular machine learning techniques [15].

2.3.3 Support vector machines

2.3.4 Lasso and Elastic net

In Section 1.2, we discussed how, in this project, not only are we seeking low classification errors, we also have to select features/variables in the data that are relevant in producing accurate predictions. An obvious, but naive, solution is to consider all the features in different combinations. But this solution is evidently computationally expensive, much less with data as large as the one we consider in this project.

One method to overcome this problem is by Lasso regression [17], which is a regularised least squares scheme that imposes an l_1-norm penalty on an error function that it tries to minimise. More importantly, in the context of big data and especially this project, the Lasso is an appealing solution because it produces a sparse solution, by shrinking the coefficients of insignificant features to 0. However, Zou and Hastie [18] examined the limitations of the Lasso method, especially in the context of microarray data. In particular, Lasso has some limitations in variable selection, if a subset of features have high correlation with one another. This is precisely a characteristic of genes, as genes often interact with one another.

As a result, Zou and Hastie proposed the *elastic net*, which imposes a linear combination (weighted) of the l_1-norm and the square of the l_2-norm. This method performs feature selection, presents a sparse solution and takes into account variables with high correlation, where groups of correlated variables are not known in advance [19]. Furthermore, Zou and Hastie showed that the elastic net method outperforms Lasso. As such, elastic net can be used on our data set.

Besides, we can also utilise the elastic net library in `scikit-learn` implemented in Python. This allows us to experiment with elastic net easily, to see if it would be suitable for our dataset.

2.4 Review of genetic data

This project makes use of data from a recent genetic-epigenetic analysis of schizophrenia, conducted in 2016 [20].

3 Project plan

3.1 Work that has been done

The first step that was done for this project was to find ways to understand the data. At the time of writing this report, the Python library `pandas` is used to manipulate the csv data set. `pandas` is a high-performance data analysis tool, which is suitable to be used on large datasets.

On a typical laptop, attempting to read the csv file row by row would result in a memory error in Python. The `pandas` library allows the file to be read chunk by chunk. This method then enables us to find out exactly how many rows and columns there are in the csv file. Moreover, we found that the rows represent the sites in the DNA that can be methylated (features), while the columns represent the samples (individuals).

Furthermore, in the National Center for Biotechnology Information (NCBI) database, a "series matrix" file explains what the data in the csv file represents. In particular, the columns (individuals) can be divided into the "control" (individuals with no schizophrenia) and the "cases" (individuals known to have schizophrenia). This is labelled by `disease_status=1` and `disease_status=2` respectively. This is helpful for classification, as we know beforehand what the label of each column (individual) is. This is necessary for supervised learning algorithms to be applied.

3.2 Performance of classifiers

Similar to previous work on cancer classification, we would need to employ different types of supervised learning algorithms on the dataset. Although several experiments have shown that support vector machines (SVMs) are superior compared to other methods [21], we should still consider other classifiers, such as decision trees and random forests. We then need to determine the classification accuracy of each classifier, and select the best classifier for our data.

Furthermore, classifiers such as the SVM requires several parameters. We would then need to run experiments to find out what the optimal values for the parameters are. These optimal values should give us the best classification accuracy.

3.3 Timeline, planning of project

The significant milestones relevant for this project are stated in the table below.

Date	Significant event
27 Jan	Submission of interim report
17 Feb	Project review deadline
25 Mar	Start of Easter break Math exam revision
30 Apr	End of Easter break Math Exam
15 May	Project health check-up
31 May	Latest date to start final report
21 Jun	Final report due

3.4 Possible extensions

First, more complex algorithms, such as the neural network, which requires more time to train due to the high dimensionality of the data, can be used. We can then, similarly, compare the classification accuracy of the neural network, and decide if we should adopt the classifier that was deemed the best, or the neural network. We also need to take into account the time that the network takes to be trained.

Second, we can create a toolbox or script that consists of an analysis pipeline for biomedical researchers to analyse similar epigenetic data. In other words, we can automate the process of manipulating the data and training the classification algorithm, such that future epigenetic data can be similarly analysed. We can then obtain biological and/or medical insight from the data much quicker.

4 Evaluation plan

4.1 An investigation into the use of epigenetic data

Essentially, this project is about investigating whether the use of epigenetic data is relevant in helping us to identify individuals with psychiatric disorders. In other words, the project might conclude that epigenetic data is not able to help us to predict individuals that potentially have schizophrenia. Nevertheless, this is also a beneficial development to the biological community.

References

[1] National Institute of Mental Health. Schizophrenia, 2016.

[2] O. J. Bienvenu, D. S. Davydow, and K. S. Kendler. Psychiatric ‘diseases’ versus behavioral disorders and degree of genetic influence. *Psychological medicine; Psychol.Med.*, 41(1):33–40, 2011. ID: TN_cambridgeS003329171000084X.

[3] Schizophrenia.com. Heredity and the genetics of schizophrenia, 2004.

[4] Bob Weinhold. Epigenetics: The science of change. *Environmental health perspectives*, 114(3):A160–A167, 2006. ID: TN_pubmed_central1392256.

[5] Florence Thibaut. Why schizophrenia genetics needs epigenetics: a review. *Psychiatria Danubina*, 24(1):25, 2012. ID: TN_medline22447081.

[6] Heidi Ledford. If depression were cancer. *Nature*, (515):182–184, 2014.

[7] W. P. Kuo, E. Y Kim, J. Trimarchi, T. K Jenssen, S. A. Vinterbo, and L. Ohno-Machado. A primer on gene expression and microarrays for machine learning researchers. *Journal of Biomedical Informatics*, 37(4):293–303, 08 2004.

[8] A. Bharathi and A. M. Natarajan. Microarray gene expression cancer diagnosis using machine learning algorithms. In *3rd IEEE International Conference on Signal and Image Processing, ICSIP 2010, December 15, 2010 - December 17*, pages 275–280, Chennai, India, 2010 2010. Bannari Amman Institute of Technology, Tamilnadu (State), India, IEEE Computer Society. Compilation and indexing terms, Copyright 2016 Elsevier Inc.; T3: Proceedings of the 2010 International Conference on Signal and Image Processing, ICSIP 2010.

[9] Chang Kyoo Yoo and Krist V. Gernaey. Classification and diagnostic output prediction of cancer using gene expression profiling and supervised machine learning algorithms. *Journal of Chemical Engineering of Japan*, 41(9):898–914, 2008. Compilation and indexing terms, Copyright 2016 Elsevier Inc.

[10] Y. Lu and J. Han. Cancer classification using gene expression data. *Information Systems*, 28(4):243–68, 06 2003.

[11] Tom M. (Tom Michael) Mitchell 1951. *Machine learning.* International 1997 edition, 1997. Includes bibliographical references and index.; ID: 44IMP_ALMA_DS2143719110001591.

[12] J.R. Quinlan. Induction of decision trees, 1986. ID: RS_60168743381inductionofdecisiontrees.

[13] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ID: TN_springer_jour1010933404324.

[14] Juergen Gall, Nima Razavi, and Luc Van Gool. An introduction to random forests for multi-class object detection. In *15th International Workshop on Theoretical*

Foundations of Computer Vision, June 26, 2011 - July 1, volume 7474 LNCS, pages 243–263, Dagstuhl Castle, Germany, 2011 2012. Computer Vision Laboratory, ETH Zurich, SwitzerlandMax Planck Institute for Intelligent Systems, GermanyE-SAT/IBBT, Katholieke Universiteit Leuven, Belgium, Springer Verlag. Compilation and indexing terms, Copyright 2016 Elsevier Inc.; T3: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).

[15] Songul Cinaroglu. Comparison of performance of decision tree algorithms and random forest: An application on oecd countries health expenditures. *International Journal of Computer Applications*, 138(1):37–41, March 2016.

[16] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, December, 2002.

[17] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, pages 267–288, 1996.

[18] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. ID: TN_wj10.1111/j.1467-9868.2005.00503.x.

[19] C. De Mol, E. De Vito, and L. Rosasco. Elastic-net regularization in learning theory. *Journal of Complexity*, 25(2):201–30, 04 2009.

[20] Eilis Hannon, Emma Dempster, Joana Viana, Joe Burrage, Adam R. Smith, Ruby Macdonald, David St Clair, Colette Mustard, Gerome Breen, Sebastian Therman, Jaakko Kaprio, Timothea Toulopoulou, Hilleke E. Hulshoff Pol, Marc M. Bohlken, Rene S. Kahn, Igor Nenadic, Christina M. Hultman, Robin M. Murray, David A. Collier, Nick Bass, Hugh Gurling, Andrew McQuillin, Leonard Schalkwyk, and Jonathan Mill. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential dna methylation. *Genome biology*, 17(1):176, 2016. ID: Hannon2016.

[21] Michael P. S. Brown, David Lin, Terrence S. Furey, David Haussler, Charles Walsh Sugnet, Manuel Ares Jr., William Noble Grundy, and Nello Cristianini. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, 97(1):262–267, 2000. ID: TN_scopus2-s2.0-0034602774.