Spring 2019

Microarray Data Analysis

Introduction to Bioinformatics – Project 3

Bini Elsa Paul
DEPARTMENT OF COMPUTER SCIENCE | UNIVERSITY OF AKRON

Abstract

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is applied to human acute leukemia as a test case[1]. *This project involves analysis of a set of real microarray data. A training dataset with 38 bone marrow samples and a testing dataset with 35 bone marrow and Peripheral Blood samples are used in this project. After the preprocessing, T Test and p-values are used for feature selection. KNN algorithm is used for classification of the samples into two classes, ALL or AML. The model is tested with the testing dataset and got an accuracy of 94 %. As a conclusion, we can say that the KNN can be used for class prediction of acute leukemia.*

1.0 INTRODUCTION

Summary of the research paper: Molecular Classification of Cancer

The challenge of cancer treatment has been to target specific therapies to pathogenetically distinct tumor types, to maximize efficacy and minimize toxicity. Improvements in cancer classification have thus been central to advances in cancer treatment. Cancer classification has been based primarily on morphological appearance of the tumor, but this has serious limitations. Tumors with similar histopathological appearance can follow significantly different clinical courses and show different responses to therapy[1].

We divided cancer classification into two challenges: class discovery and class prediction. Class discovery refers to defining previously unrecognized tumor subtypes. Class prediction refers to the assignment of particular tumor samples to already defined classes, which could reflect current states or future outcomes[1]. In this project we are focusing on class prediction.

We chose acute leukemia as a test case. We will try to classify acute leukemia into those arising from lymphoid precursors (acute lymphoblastic leukemia, ALL) or from myeloid precursors (acute myeloid leukemia, AML). Although the distinction between AML and ALL has been well established, no single test is currently sufficient to establish the diagnosis. Our initial leukemia data set consisted of 38 bone marrow samples (27 ALL, 11 AML) obtained from acute leukemia patients at the time of diagnosis. RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix and containing probes for 6817 human genes. For each gene, we obtained a quantitative expression level[1].

The first issue was to explore whether there were genes whose expression pattern was strongly correlated with the class distinction to be predicted. The 6817 genes were sorted by their degree of correlation. To establish whether the observed correlations were stronger than would be expected by chance, we developed a method called "neighborhood analysis". Briefly, one defines an "idealized expression pattern" corresponding to a gene that is uniformly high in one class and uniformly low in the other. For the 38 acute leukemia samples, neighborhood analysis showed that roughly 1100 genes were more highly correlated with the AML-ALL class distinction than would be expected by chance. This suggested that classification could indeed be based on expression data[1].

The second issue was how to use a collection of known samples to create a "class predictor" capable of assigning a new sample to one of two classes. We developed a procedure that uses a fixed subset of "informative genes" (chosen based on their correlation with the class distinction) and makes a prediction on the basis of the expression level of these genes in a new sample[1].
The third issue was how to test the validity of class predictors. We used a two-step procedure. The accuracy of the predictors was first tested by cross-validation on the initial data set. Then builds a final predictor based on the initial data set and assesses its accuracy on an independent set of samples[1].
We applied this approach to the 38 acute leukemia samples. The set of informative genes to be used in the predictor was chosen to be the 50 genes most closely correlated with AML-ALL distinction in the known samples. The parameters of the predictor were determined by the expression levels of these 50 genes in the known samples. The predictor was then used to classify new samples, by applying it to the expression levels of these genes in the sample. The 50-gene predictors derived in cross- validation tests assigned 36 of the 38 samples as either AML or ALL and the remaining two as uncertain. All 36 predictions agreed with the patients' clinical diagnosis. We then created a 50-gene predictor on the basis of all 38 samples and applied it to an independent collection of 34 leukemia samples. The specimens consisted of 24 bone marrow and 10 peripheral blood samples. In total, the predictor made strong predictions for 29 of the 34 samples, and the accuracy was 100%. The success was notable because the collection included a much broader range of samples, including samples from peripheral blood rather than bone marrow. Together, these data suggest that genes useful for cancer class prediction may also provide insight into cancer pathogenesis and pharmacology[1].

2.0 MATERIALS and METHOD

There are two datasets containing the initial (training, 38 samples) and independent (test, 34 samples) datasets used in the above paper. These datasets contain measurements corresponding to ALL and AML samples from Bone Marrow and Peripheral Blood[2].
T test and P-values are used for feature selection and KNN classification is used for class prediction. The T test compares two averages (means) and tells you if they are different from each other. The T test also tells how significant the differences are; In other words it tells if those differences could have happened by chance. A p-value is the probability that the results from the sample data occurred by chance. P-values are from 0% to 100%. They are usually written as a decimal. For example, a p value of 5% is 0.05. Low p-values are good; they indicate the data did not occur by chance. For example, a p-value of .01 means there is only a 1% probability that the results from an experiment happened by chance. In most cases, a p-value of 0.05 (5%) is accepted to mean the data is valid[3].
KNN is a non-parametric and lazy learning algorithm. Non-parametric means there is no assumption for underlying data distribution. In other words, the model structure determined from the dataset. This will be very helpful in practice where most of the real world datasets do not follow mathematical theoretical assumptions. Lazy algorithm means it does not need any training data points for model generation. In KNN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor. K is generally an odd number. Suppose P1 is the point, for which label needs to predict. First, you find the k closest point to P1 and then classify points by

majority vote of its k neighbors. Each object votes for their class and the class with the most votes is taken as the prediction. For finding closest similar points, you find the distance between points using distance measures such as Euclidean distance, Hamming distance, Manhattan distance and Minkowski distance[4]. KNN has the following basic steps:

- Calculate distance
- Find closest neighbors
- Vote for labels

The below figure shows how KNN works.

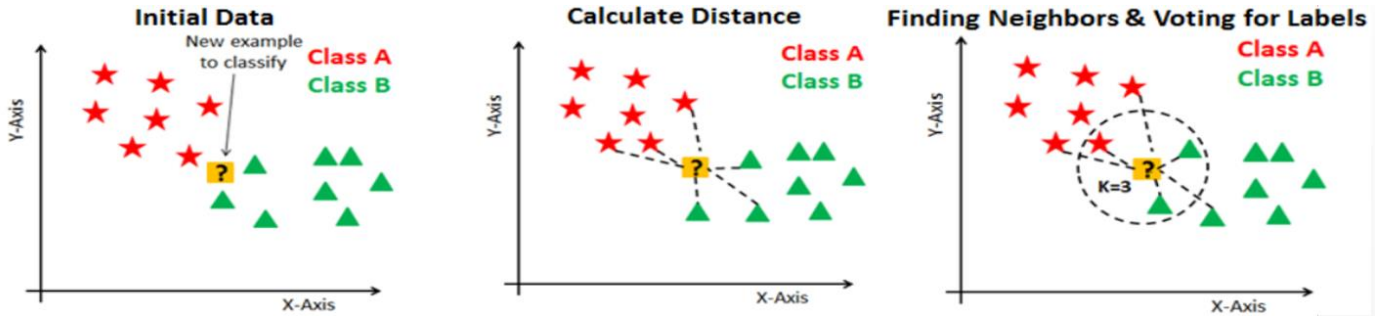

Figure 1: KNN classification for new data

3.0 IMPLEMENTATION

The steps used in this project are

- Preprocessing
  - → Eliminated the "endogenous control" genes (housekeeping genes)
  - → Eliminated the genes with all As (Absent) across the experiments
  - → Replaced all the expression values below some threshold cut-off value (used 20) to that threshold value
  - → Eliminate the genes with less than two fold change across the experiments
- Feature Selection
  - → T test (arr1, arr2, tail=3, type=3) and get the P-Values for the genes using Excel.
  - → Arranged the dataset in ascending order of p-values
  - → Selected top 50 genes are the feature to the classifier
- Class Prediction
  - → Sklearn KNeighbours Classifier
  - → Picked K as 5

We have used python to implement this project. The datasets are saves as csv files for easiness. After the feature selection, we have changed the columns and rows (took a transpose), so that our dataset will be compatible to the datasets used for KNN. We also did the same preprocessing and feature selection in testing data set. Both the training and testing datasets are arranged in ascending order of the gene expression in order to make the features same in both datasets. We tried with K = 3 and K = 5 and found out 5NearestNeighbours suits better for our project.

4.0 RESULTS and DISCUSSION

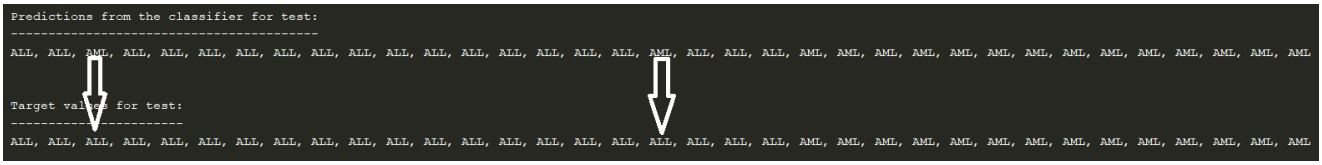

Figure 2: Output of the Classifier

Confusion Matrix

```
[[19  2]
 [ 0 14]]
```

Classification Result

	precision	recall	f1-score	support
ALL	1.00	0.90	0.95	21
AML	0.88	1.00	0.93	14
micro avg	0.94	0.94	0.94	35
macro avg	0.94	0.95	0.94	35
weighted avg	0.95	0.94	0.94	35

Figure 3: Confusion Matrix and Classification Result

As Figure 2 shows, we got only 2 misclassified test instances. ALL is misclassified as AML. Figure 3 shows the confusion matrix and classification result. As the figure 4 shows, K =5 gives a better accuracy.

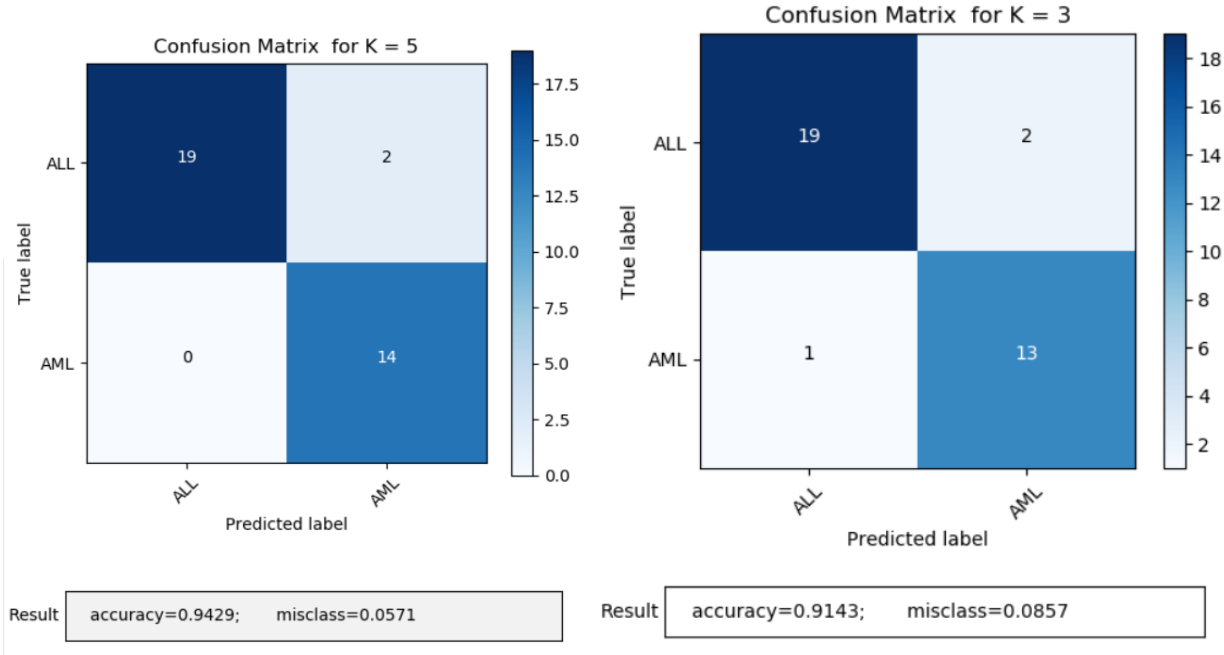

Figure 4: Output for K = 3 and K = 5

Discussion

- K = 3
  - → AML is predicted as ALL (second last row)
  - → Higher values of K predicts this sample correctly
- K = 7 and 9
  - → ALL is predicted as AML
  - → 14th sample
- K = 11
  - → 4 errors but since AML has only 11 instance for training, higher values of K does not work

For K = 3 from the confusion matrix above, it can be seen that an instance of AML is predicted wrong, but higher values of K predicts this instance correctly. This shows that this instance is very similar to ALL.

Similarly for K = 7 and 9, an instance of ALL is predicted as wrong. This shows that even though it is close to other ALL classes (smaller values of K predicts this instance correctly), it can be considered towards the margin of different classes.

5.0 CONCLUSION

We have successfully implemented the class prediction of acute leukemia into ALL or AML from a set of real micro array data. The classifier is 94 % accurate.

6.0 REFERENCES

[1] T. R. Golub *et al.*, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[2] "Cancer Program Legacy Publication Resources." [Online]. Available: http://portals.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43. [Accessed: 04-May-2019].

[3] "T Test (Student's T-Test): Definition and Examples," *Statistics How To*. [Online]. Available: https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/. [Accessed: 04-May-2019].

[4] "KNN Classification using Scikit-learn," *DataCamp Community*, 02-Aug-2018. [Online]. Available: https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn. [Accessed: 04-May-2019].