Performance of classifiers for gene expression data classification on PAN-CAN data subset

Chantal van Duin s1004516, Nayeong kim s1006313

January 2019

1 Abstract

Tumour classification using data mining techniques based on gene expression data has lead to a new interest and potential for data mining algorithms. This research will take a closer look into the performance of the data mining classifiers : decision tree, k-nearest neighbour classifier, neural network and support vector machines. The selected classifiers will be compared with each other which one performs better on a subset of the PAN-CAN data set, a data set containing the gene expression for 5 tumour types classified by the TCGA organization. Using feature selection of decision tree classifier, 3 potential informative genes have been found. The results of the research remains inconclusive and more research with other data sets must be conducted however the potential for the classification for cancer using data mining seems to be apperant.

2 Introduction

Modern technology such as Micro-array gene expression technology has given rise to new ways in how DNA can be examined and studied. With the use of Micro-array technology the profile of gene expression in DNA can be observed and documented which has lead to new insights of how gene expression behaves under various conditions of the human body. Particularly of interest is how these techniques can be used to study and treat cancer patients considering the fact that tumours are the result of malfunction of gene expression.
Cancer is a highly complex disease consisting of various sub-types and development stages. Cancer is paired with various changes on epigenetic level such as epigentic regulation, RNA transcription, RNA translation and histone acetylation. Discerning at which stages of differing cancer types these changes develop will improve the understanding of the causes of cancer and how it can be treated. If cancer could be diagnosed at an earlier stage, the fatality rate of cancer may decline. Cancer at this moment in time is the second most leading cause of death world wide and approximately 42 million people have a form of cancer [25]. The possibility of a method capable of correctly classifying cancer could lead to solving a few of complications that arise from cancer diagnosis and its treatment. Using classification for cancer if successful could make the diagnosing of tumours more easy and be possible at an earlier stage. In addition, when tumours are noted to be similar in genetic profile or development, the treatment of one tumour may be extended to another. Thus the potential of the use of using technology such as Micro-array technology in combination with analysis and data mining techniques for predicting cancer types has been a focal point of recent research Some data mining or machine learning techniques have designed for the task of classification. They have the

potential of classifying patients into high and low risk groups based on their genetic profile [26]. In addition, using feature selection the data mining or machine learning algorithms may highlight which factors play a leading role in the development of cancer. In turn, DNA profiles resulting from these techniques form an excessive amount of data, which is one of the problems that the field of data mining is attempting to solve. The complications of the classification of cancer may provide insights in how to improve, adjust or further develop the data mining or machine learning techniques. This research will look into to which degree Data Mining classification techniques can be used to predict a tumour type based on the gene expression profile of a person and which existing classifier performs best for a chosen data set.

2.1 Previous research

From conducted research, a few Data Mining techniques have been proven to have potential and resulting complications have been formulated. An overview of insights from earlier research is given below :

- Interaction classifiers or machine learning techniques. An important component of human disease is how complex interactions between genes are more influential than the effect of one gene[1]. The classifier technique should take this aspect into account. Classifiers that allow for interaction effects have been proven to be more effective than classifiers that do not [2][3]. Examples of interaction classifiers techniques and machine learning techniques are tree classifiers, Artificial Neural Networks, Support Vector Networks, k-Nearest Neighbour, logistic regression, cellular automate, random forest, Multifactor Dimensionality Reduction and weighting voting methods [3][4].

- High dimensionality. One of the biggest complications is how the Micro-array technique result in data sets with a high dimensionality. On average, each gene in the gene profile represents one attribute that needs to be taken into account. Few techniques are equipped to handle this degree of dimensions efficiently [5][6]. This is especially made more difficult as the majority of data bases consist of having a relatively low samples sizes [7].
  As a result, feature selection is often considered as a prepossessing method to reduce the amount of dimensions. Feature selection forms one of the biggest challenges for gene classification [8][13][14]. Multiple techniques have been considered and developed to select the genes with the most relevancy / informative genes :
  - Backward Elimination Hilbert-Schmidt Independence Criterion [9][10].
  - Extreme Value Distribution based gene selection [10][11]
  - Singular Value Decomposition Entropy gene selection [10][12]
  - Mutual Information [15]

- Inconsistency. Results from research show a lack of consistency, even in the Micro-array analysis results for the same data sets [16][17]. Further deviations between results of research can be attributed to having different methods to build classification models [17][18] and the outcomes may depend on the specific Micro-array data used [19].

2.2 Chosen data set

For this research the data set that was chosen is a part of the data collection 'RNA-Seq (HiSeq) PANCAN data set' [20]. The PANCAN data set has been constructed for a project the Pan-Cancer

intiative, which is was developed by Cancer Genome Atlas Research Network (TCGA) [21]. TCGA is dedicated to storing, profiling and analyzing information about on various cancer types [22]. Its database contains data of hundreds tumor types and its sub-types on RNA, DNA, protein and epigenetic level [23]. The aim of the organization is to acquire knowledge about the difference and common ground of cancer types, the reason and way tumours develop and how it can be treated.
The Pan-Can project was created to analyze the connection between the first twelve tumour types profiled by the TCGA data base in order to extend treatment based on similar genomic profiles. The chosen data set is a sub-set of the data set of the Pan-Can project. It is consisted of the RNA sequence data for 5 tumour types instead of the 12 of the PANCAN data set, namely :

- Kidney renal clear cell carcinoma (KIRC)
- Colon adenocarcinoma (COAD)
- Lung adenocarcinoma (LUAD)
- Prostate adenocarcinoma (PRAD)
- Breast invasive carcinoma (BRCA)

The chosen data set stores data of the degree of gene expression for 20531 genes of 801 patients that has one of the 5 tumours depicted above. The data does not contain any missing data and the genes are not saved under their name but under a number. The original names of the genes can be found in the original PANCAN data set [24]. The gene attribute variables are continuous while the tumour label variable is categorical. The data set does not contain any information about gene expressions for people without cancer.

2.3 Research Problem

This report will look into the problem of using data mining classifiers for the prediction of tumour type based on the gene expression. It will take a look into how a few selected classifiers perform, how well it generalizes to the overall public, what issues arise from the usage of these selected classifiers and the choices made during the process of comparing the classifiers. Based on previously conducted research, the data mining classifiers that were chosen to be considered in this research are Decision Tree Classifier, k-Nearest Neighbour classifier, Artificial Neural Network and Support Vector Machine. The motivation of why these classifiers were chosen was that they have a lot of potential to work relatively well for the classification of tumours based on the results of earlier research. Further their characteristics allows for the usage of classification based on the type of data contained in the chosen data set and highlights a few interesting fundamental differences between the classifiers. The reason why for example rule-based classifiers such as association rule classifier was not used, is that the predictor variables - the selected genes - of the chosen data set is made of quantitative data while association rule classifier requires qualitative data. Another interest point of this research is feature selection of genes. This research will not extensively research this aspect but it will be briefly addressed as the importance of feature selection has been highly emphasized in related research. Techniques such as the Decision Tree Classifier have naturally feature selection build in the technique however k-Nearest Neighbour classifier does not for example. This research will take a shallow look into how well a model with the genes selected by the feature selection part of the constructed Decision Tree Classifier performs and predict the data.

3 Methods and Approach

3.1 General approach performance classifiers

The appliance Jupyter's notebook from Anaconda will be used throughout the testing to conduct the analysis. To see how the analysis were done in detail and the results, see the added file of the Jupyter notebook added to the submission of this report. The general approach for this research for comparing the selected type of classifiers, is to first find the optimal hyper-parameters for each classifier type by testing the accuracy for one classifier for multiple hyper-parameter settings, for example the optimal tree depth for Decision Tree Classifier. To ensure that the performance of the classifiers are compared on the same standards and to test the accuracy of the model based on independent test data sets, the classifiers are trained on the same train set and tested with the same test set using k-fold cross-validation. Leave-one-out cross-validation was deemed to computationally expensive for the size of the chosen data set. It was chosen to do 10-fold cross-validation as it is most commonly used in data mining. For each classifier when applicable a graph will be constructed visualizing the classification error rate for various settings of the hyper-parameters of the classifier. A graph will be constructed to showcase the classification error for each optimized hyper-parameter tuned classifier in order to compare the performance of the classifiers.

3.2 Classifiers

3.2.1 Decision Tree Classifier

The decision tree classifier was particularly of interest for this research as the technique automatically does feature selection and works well with high dimensional data, which is a fundamental characteristic of the chosen data set. For a detailed description of how and why the feature selection was used in this research, see the section 3.3 Feature selection and prepossessing techniques and 3.3.2 gene feature selection of this report. The decision tree classifier has two hyper-parameter that needs to be fine-tuned, the minimal samples split and the maximum tree depth. For choosing the optima minimal samples split value, a various of possible minimal samples split values are chosen : 100, 200, 300 and the accuracy of the decision tree classifier created with those values is graphed. This resulting graph will give a rough idea of the value that would be best, in order to choose the best minimal sample split value more specific values are chosen and graphed in the same way. Using the found optimal value for minimal sample split, the optimal maximal tree depth is fine-tuned by using 10-fold cross validation and the maximal tree depth is graphed as classification error on possible tree depth values. The tree depth with the lowest classification error is selected as optimal maximum tree depth.

3.2.2 K-Nearest Neighbour Classifier

K-Nearest neighbour Classifier is that calculates k distances of data to be classified and all the distances of given data, and classifies them as the most frequent class among them. First, it is needed to find the most optimal number of neighbours for the data given when the remaining conditions are the same except for the classifier setting. In this process, data is divided into train set and test set using 10-fold cross-validation. The number of neighbours is set as a variable, learn the train set using the classifier, and then use the test set to calculate the error rate with the actual data. At this time, the number of neighbours(= a) having the lowest error rate is considered to be optimal n_neighbours. Next, a new classifier is defined using the best number of neighbours found. That is, n_neighbours = a is the most accurate classifier with 10-fold cross-validation condition when using

KNN classifier. However, the KNN classifier has the drawback of high computational cost because it must always have learning data and it needs to calculate the distance between the data to be categorized and all the learning data.

3.2.3 Artificial Neural Network

For the artificial neural network, the multiple-layer perceptron classifier was chosen to be used. Since the number of genes / attributes is 20531, it was chosen to have the hidden units also be 20531 to represent each gene in order to ensure that the classifiers considers each gene in the data set. Other hyper-parameters of the multiple-layer perceptron classifier such as maximal number of iterations and learning rate will not be fine-tuned for the reason that it is too computational expensive for the size and objective of this research. Since the code will not terminate if the original data is used as it is too computationally expensive, the compressed dimensional data of PCA projected onto the first 300 PAs will be used. The average classification error of the neural network of 5 learning processes will be computed for the training and test set using 10-kfold cross validation technique.

3.2.4 Support Vector Machine

The support vector machine is an advanced algorithm belonging to supervised learning which is used for Support Vector Machine classification. SVM is a binary classification algorithm that finds decision boundaries that maximize margins. It is much more powerful than Logistic regression or Neural network when dealing with complex nonlinear problems. The margin in the SVM is the distance between the decision boundary and the support vector nearest to this boundary. To compare the performance of this classifier with other classifiers, it is necessary to find a classifier that returns the highest accuracy when the same condition, 10-fold cross-validation, occurs. In this study, class data has all string types as cancer types, so we check the performance of the classifier when gamma is not 'scale'. When we classify the data using the linear kernel and the poly kernel with gamma = 'auto', we can confirm that the accuracy of the two classifiers is exactly the same for each learning data. Further testing with fine-tuning the hyper-parameters will not be done as the expertise required to do so is out of our current knowledge.

3.3 Feature selection and prepossessing technique

As mentioned before, feature selection plays an important role in gene expression classification. When using classification, feature selection can be a needed prepossessing technique in order tot run the classifier or feature selection is already an integrated part of it.

3.3.1 Principal Component Analysis

Since the chosen data set has a high dimensionality, classifiers such as k-Nearest Neighbour Classifier and multiple-layer perceptron may not compute an output as it is too computationally time consuming because of the high dimensions. Multiple ways to reduce the dimensions by selecting a smaller numbers of genes that represent the most fundamental genes / informative genes, has been proposed in earlier research, see high dimensionality part of the section 2.1 previous research. As we are not experts on gene expression nor on advanced feature analysis techniques, Principal Component Analysis was chosen to reduce the amount of dimensions. In order to be able to calculate the variance explained by a certain number of principal components and use the data of the projection of the data on a certain number of principal components, the principal component analysis will be calculated manually. To graph the explained variance by a certain number of principal components,

the principal component analysis module was used from Anaconda sklearn module. The number of dimensions will be reduced to the number of components that explain 90 % of the variance of the data, which can be deduced the computed graph.

3.3.2 Gene Feature Selection

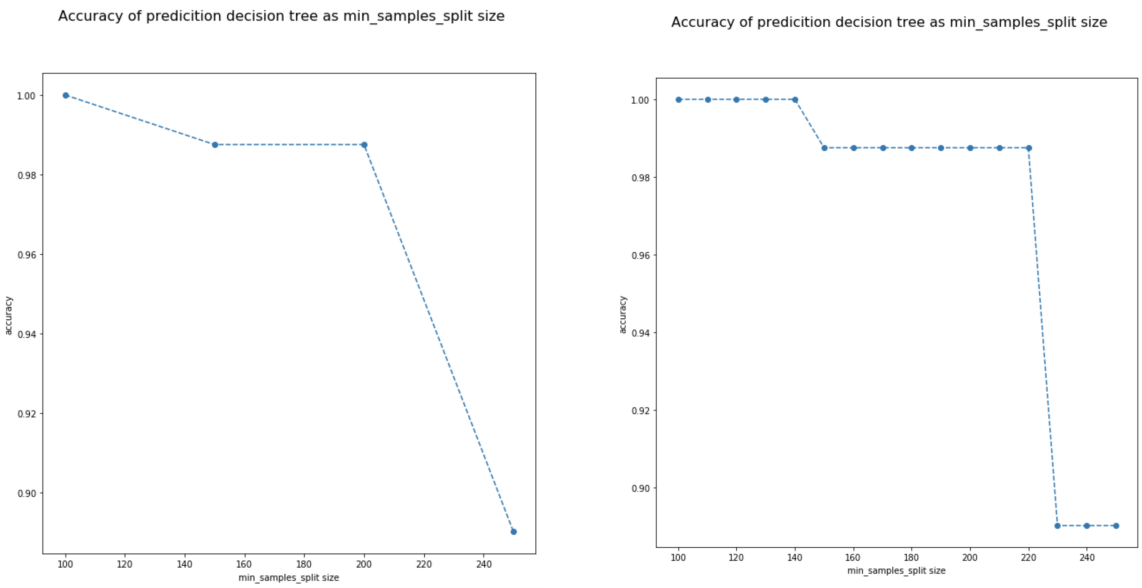
The decision tree classifier automatically does feature selection when computing the classifier. By printing the tree of the decision tree classifier, the genes that are the most informative can be found. For choosing the most important on genes, a tree classifier for the original data will be created and its tree will be printed and the genes that are most frequently present in the data will be chosen. To test whether these genes are indeed informative genes for predicting the sort tumour, from the original data set a subset is created containing only the most important genes. Using the created subset of original data with the most informative genes, the classification for decision tree classifier and for k-Nearest Neighbour classifier is tested using 10-kfold cross validation and graphed to see how well the classifier can predicted the tumour type on the genes selected by the decision tree classifier as informative genes.

4 Results

4.1 Find best classifiers of the data given

1. Decision Tree classifier (DCT)

   (a) Results to find the best min_sample_split number : from the graph showing the accuracy as a function of splitting the sample from 100 to 250 in low and high detail, it can be seen from Figure 1 that the accuracy was highest at minimal sample size value 100. The decision tree with minimal sample size value 100 also contains each tumour type that is concluded in the data set, see Figure 2.



(a) Low detail

(b) High detail

Figure 1: Graph to find optimal minimal splitting sample value for decision tree classifier

```
                        |->5  BRCA
                        |
                    |->4 then if gene_15898 =< 2.67: go to 5, else go to 6
                    |   |
                    |   |->6  LUAD
                    |
                |->3 then if gene_3523 =< 6.32: go to 4, else go to 7
                |   |
                |   |->7  KIRC
                |
            |->2 then if gene_9175 =< 7.83: go to 3, else go to 8
            |   |
            |   |->8  PRAD
            |
        |->1 then if gene_12983 =< 9.06: go to 2, else go to 9
        |   |
        |   |   |->10  LUAD
        |   |   |
        |   |->9 else if gene_1413 =< 8.61: go to 10, else go to 11
        |       |
        |       |->11  COAD
        |
if gene_18746 =< 10.73: go to 1, else go to 12
        |
        |   |->13  PRAD
        |   |
        |->12 else if gene_11491 =< 4.79: go to 13, else go to 14
            |
            |->14  BRCA
<------------------------>
Tree Depth:  5
```

Figure 2: Decision tree with minimal splitting sample value 100.

(b) Using on the above results, the classifier with min_sample_split = 100 is used to find the optimal max_depth when K-fold cross-validation occurs for k = 10. The error rate is calculated by performing cross-validation on each classifier of the depths from 2 to 11. As

a result, it is shown that the lowest error rate(=0%) occurred in train data set when depth >= 5, see Figure 3. Thus, min_sample_split = 100 and max_depth = 5 are the optimal hyper-parameters to be used when creating the decision tree classifier when comparing classifiers.

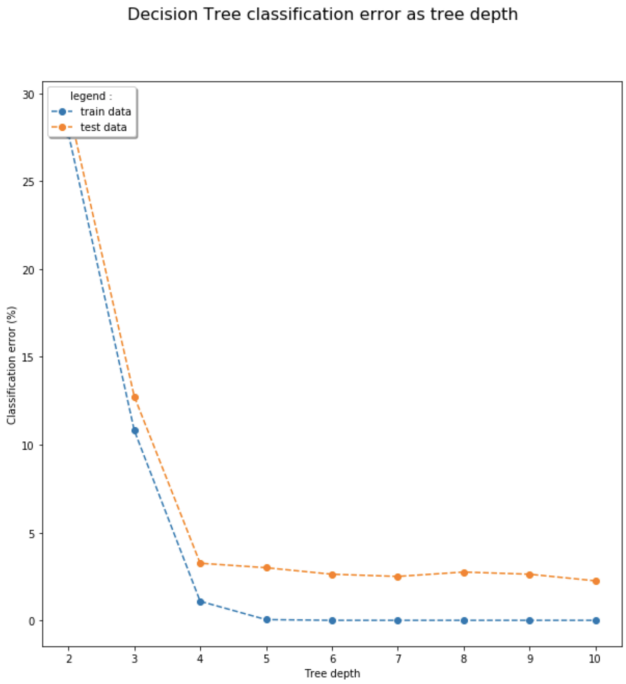


Figure 3: Graph to find optimal maximal tree depth for decision tree classifier with minimal splitting value 100

2. K-Nearest Neighbour classifier (KNN)

   (a) The optimal number of neighbors was found using the condition of 10-fold cross-validation. The number of neighbors of 5 resulted in the lowest error rates in both the train set and the test set, see Figure 4. Therefore, n_neighbors = 5 will be used as hyper-parameter for k-Nearest Neighbour classifier when comparing classifiers.

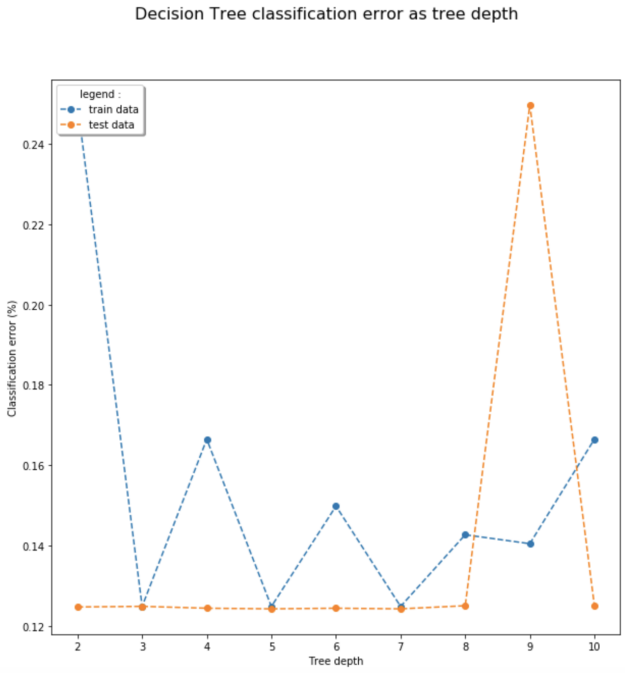


Figure 4: Graph to find optimal number of neighbours for KNN classifier

(b) For data that has been reduced to PCA, n_neighbors = 5 appears to also have highest accuracy. Using the data computed by using PCA reduced the computation time significantly and results in having the same accuracy as using KNN on the original data, see Figure 5.

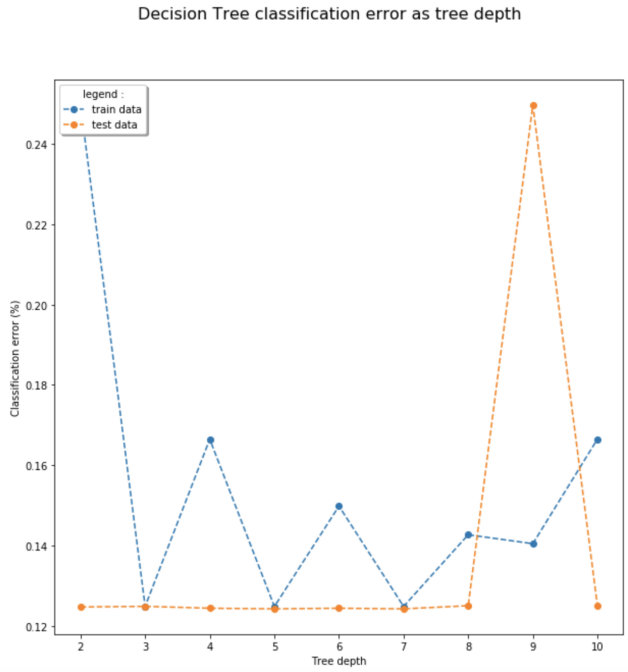


Figure 5: Graph to find optimal number of neighbours for KNN classifier using PCA projected data

3. Artificial Neural Network

   (a) For the artificial neural network, the multi-layer perceptron classifier is used. Since the number of genes / attributes is 20531, it was chosen to have the hidden units also be 20531 to represent each gene. The original code was not used because the data is too computationally expensive to compute when using multi-layer peceptron. Instead, the compressed dimensional data of PCA projected onto the first 300 PAs is be used but even then the computation time is long. All of the average classification error rates of the neural network of 5 learning processes using 10-kfold cross validation and the PCA projected data, are lower than 2%, see Figure 6.

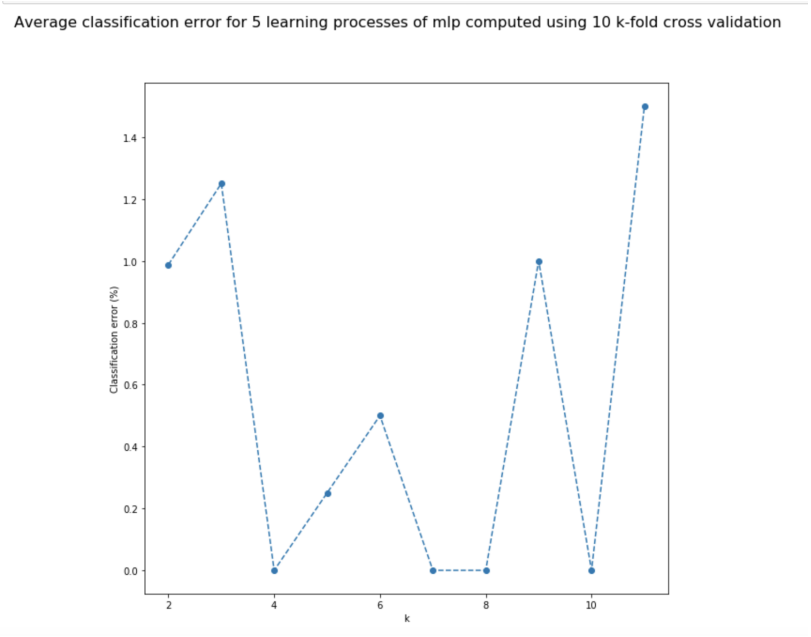


Figure 6: Graph for average classification error for MPC classifier of 5 learning rates computed 10 times

4. Support Vector Machine (SVM)

   (a) The classification error rates of the linear kernel and the polynomial auto kernel are exactly the same. Both support vector machines have an accuracy of 100 % 9 out of 10 times, see Figure 7. There is no difference between the performance of the linear and the polynomial kernel support vector machine, either one of them can be used to compare with the other classifiers.

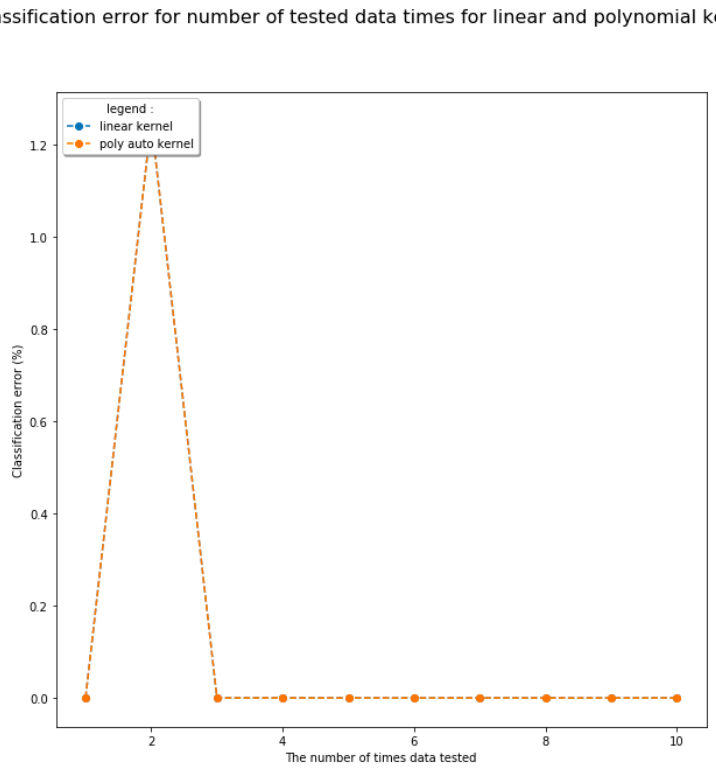


Figure 7: SVM Classification error for linear and polynomial kernel type

4.2 Compare classifiers

To get the most accurate classification of a given data, it is necessary to compare the classifier under the same conditions. From the above founded results the optimal hyper-parameters were found to provide most optimal model for each classifier when k = 10. These parameters are used to define best classifier in the condition given, after which the average classification error rate is calculated of running each test set created by the cross validation five times. A graph with classification error rate for all the classifications is used to determine which classifier predicts the various tumour types the most accurate (the lowest classification error rate). Most ideal would be to create train sets and test sets through cross validation and then test all of the classification criteria at once. This method is reliable because it uses exactly the same data set when it comes to classification. However, using multiple classifiers at once to train data set and calculating the accuracy of all them requires too much computation memory for most computers, which makes the computation time too long and may crash the computer. That is the reason, why the multiple-layered preceptor classifier was left out of the graph below, the high computation time of the MDP classifier was just too high in combination with the computation time of the other classifiers.

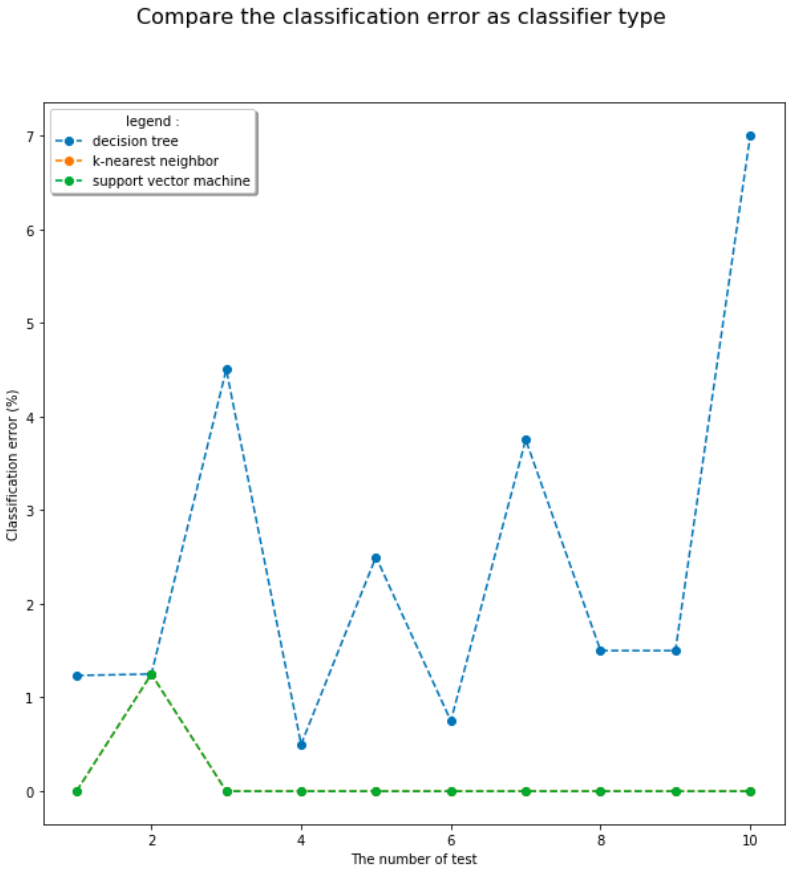


Figure 8: Graph to classification error as classifier type

The K-nearest neighbour classification(KNN) and the support vector machine classification(SVM) showed similar error rates, and the decision tree classification(DCT) has slightly higher error rates than the other two error rates. However, the error rate difference between DCT and KNN is not noticeably large. DCT have the advantage that no prepossessing technique has to be used and that it does feature selection automatically. KNN and SVM show similar results, but the computational time for SVM is lower and no prepossessing technique is required to be used. Because of the high computational time needed for KNN when using the original data and of the fact that PCA or another prepossessing technique that reduces the dimensions is on average is needed, KNN is less practical in real life situations. Furthermore, it can not be computed which gene are actually informative, this is something that DCT can do.

4.3 Data Analysis

1. Principle Component Analysis (PCA)

   (a) From the graph in Figure 9, it can be seen that around 90% of the variance can be explained from the first 300 Principal Components. To examine this further, a closer

look is taken at the graphs of Figure 10 : the explained variance of the first 10 PC and the explained variance of the first 300 PC.

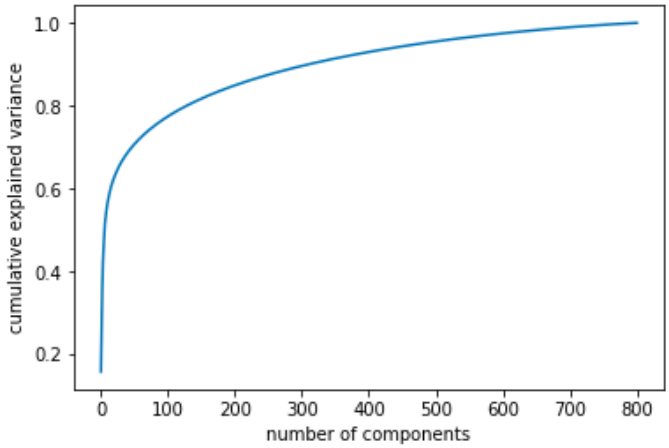
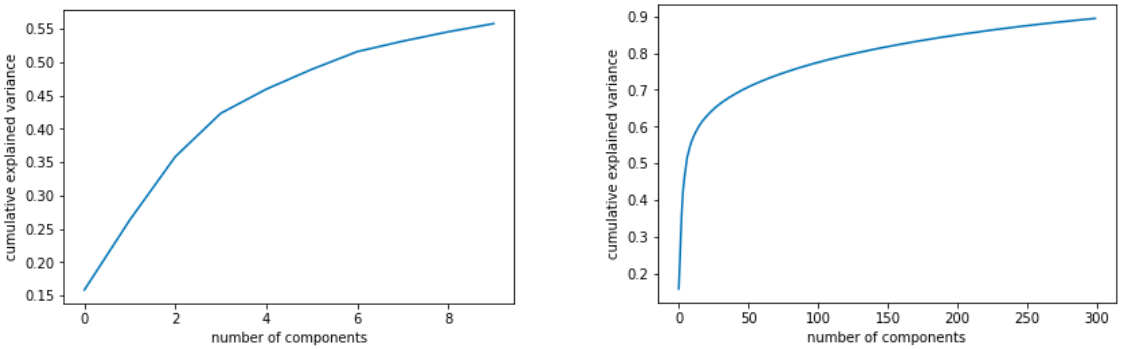


Figure 9: The variance of the total as the number of components



(a) 10 PCs

(b) 300 PCs

Figure 10: Graph to explained variance of the first 10 PC and of the first 300 PC.

From these graphs, it can be seen that first 10 principal components explain around 55 % of the variance in the data and the first 300 around 90 %. The number of principal components explaining around 90 % of the variance is usually used to project the PCA data on and to be used in analysis requiring a lower dimensionality.

2. Feature Selection

   (a) Check feature selection

```
                  |->5  BRCA
                  |
               |->4 then if gene_15898 =< 2.67: go to 5, else go to 6
               |  |
               |  |->6  LUAD
               |
            |->3 then if gene_3523 =< 6.32: go to 4, else go to 7
            |  |
            |  |->7  KIRC
            |
         |->2 then if gene_11910 =< 10.69: go to 3, else go to 8
         |  |
         |  |->8  PRAD
         |
      |->1 then if gene_12983 =< 9.06: go to 2, else go to 9
      |  |
      |  |  |->10  COAD
      |  |  |
      |  |->9 else if gene_5595 =< 7.61: go to 10, else go to 11
      |     |
      |     |->11  LUAD
      |
if gene_18746 =< 10.73: go to 1, else go to 12
      |
      |  |->13  PRAD
      |  |
      |->12 else if gene_4861 =< 0.90: go to 13, else go to 14
         |
         |->14  BRCA
<----------------------->
Tree Depth:  5
                  |->5  BRCA
                  |
               |->4 then if gene_15896 =< 1.93: go to 5, else go to 6
               |  |
               |  |->6  LUAD
               |
            |->3 then if gene_3523 =< 6.32: go to 4, else go to 7
            |  |
            |  |->7  KIRC
            |
         |->2 then if gene_203 =< 9.75: go to 3, else go to 8
         |  |
         |  |->8  PRAD
         |
      |->1 then if gene_12983 =< 9.06: go to 2, else go to 9
      |  |
      |  |  |->10  LUAD
      |  |  |
      |  |->9 else if gene_15782 =< 9.11: go to 10, else go to 11
      |     |
      |     |->11  COAD
      |
if gene_18746 =< 10.73: go to 1, else go to 12
      |
      |  |->13  PRAD
      |  |
      |->12 else if gene_14604 =< 3.07: go to 13, else go to 14
         |
         |->14  BRCA
<----------------------->
Tree Depth:  5
```

Figure 11: Two Samples among 10 trees printed

In the 10 trees, gene_18746, gene_12983 and gene_3523 always appear as classification criteria to decide which tumour type it is in the decision tree classifier. Therefore, these genes can be seen as informative genes in cancer classification.

(b) Check model

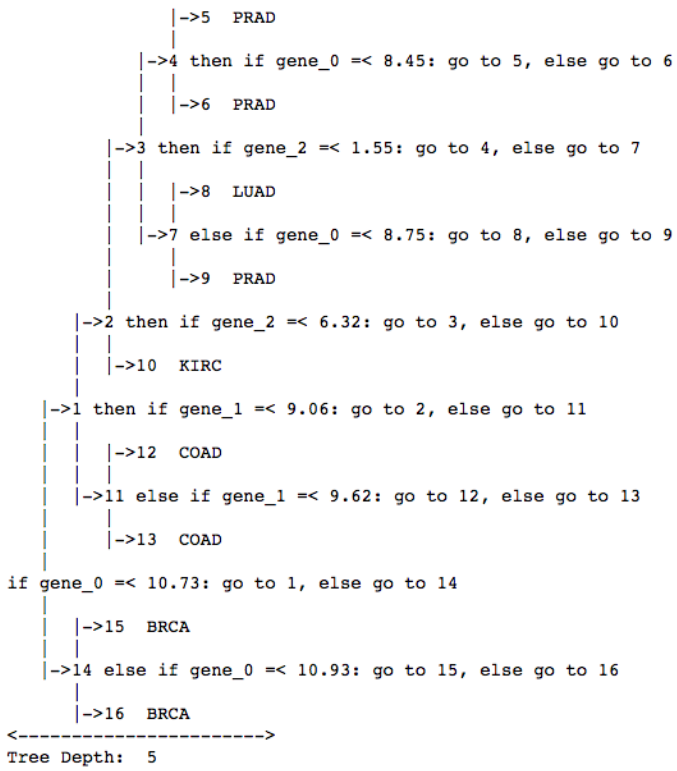


Figure 12: Feature-selection model tree

Using the previously found feature, a new set with only the corresponding gene was created and used in the creation of a decision tree classifier. Using only the selected three genes are enough to classify all types of cancers, which can be seen in Figure 12.

(c) Predict model

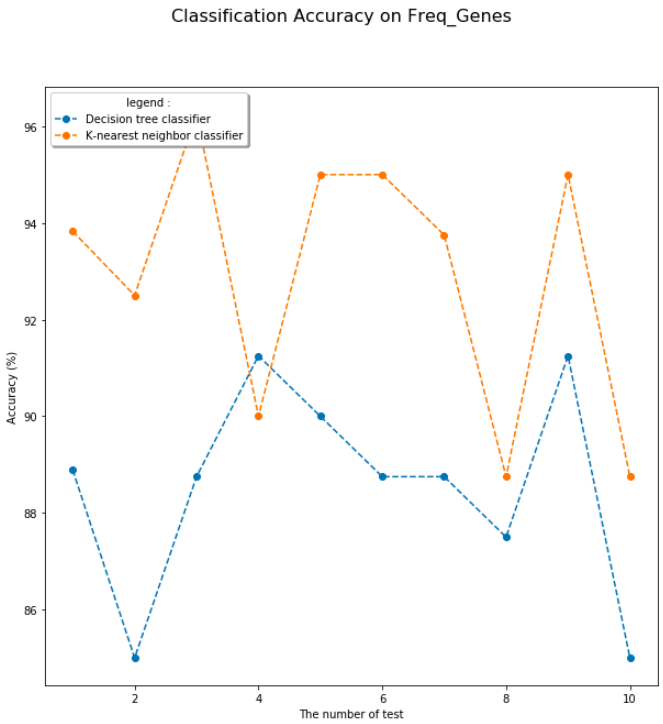


Figure 13: Classification Accuracy on Freq$_{G}enes$

The newly defined feature subset is used to create the decision tree classifier and a k-nearest neighbour classifier to test the accuracy of the two classifier using 10-fold cross-validation when only using the selected informative genes as predictors. The accuracy for both classifiers is very different depending on which test data it used. However, it is shown that with the exception of the 4th test, KNN classifier works much better than DCT classifier and that the classifiers can explain with an accuracy from 84 to 96 % the tumour types using the information of the selected informative genes alone.

5 Discussion

From the results, it can be seen that the selected have a high accuracy for predicting the tumour types based on the gene expression information and that the difference between the accuracy values are quite small. This was an unexpected result as we had expected to see a lower accuracy level from the results from earlier conducted research. We suspect that this high level of accuracy comes from the data set used and not necessarily because the classifiers performs so well. The data set only contains 5 types of tumours and no information of gene expression for a person that does not have a tumour. It would be interesting to conduct a further experiment, where this information is included as well to see whether the classifiers still has such a high accuracy levels. As we are not experts on gene expression or cancer tumour characteristics, we can not theorize whether the reason for the found high accuracy level of the classifiers has to do with the tumour types itself rather than the data set. From our results we do see the advantage of having feature selection already included in the classifier algorithm. For further research, it would be good to expand on more feature selection

techniques for the various selected classifiers to see whether this will influence their performance, from other research it is suspected that it does. Another thing that could be more tested in is the support vector machine algorithms. From our results it became apparent that the computation was low in comparison to the multiple-level preceptor algorithm and that the classification error was very low. This is in line with other conducted researches, it would be interesting to expand the research more using various hyper-parameters of the support vector machine. Overall, it is hard to get a concrete conclusion from which classifiers works the best for this kind of classification problem. It is recommended to repeat the experiment with another data set to see whether the performance of the classifiers is similar. Another thing is that more knowledge is required to better analyze the results and to get results with a higher validity. For example, the reason why no split test or ROC graph was conducted to compare the classifiers, is that the creation and evaluation of a multi-class ROC is beyond our knowledge at the moment. The genes found to be informative genes using feature selection of the decision tree classifier were genes given the dummy names 3522, 12983 and 18745. Because of complications with opening the file containing the actual names for the genes as the file is too huge to display for the programs on our computers, the names of the genes will not be reported here. If interested it can be found in the link noted in list of the references [24].

6 Conclusion

When trying to classifying different types of tumours using gene expression data, there are various classifiers available. The performance between classifiers Decision Tree classifier, K-Nearest Neighbour classifier, Artificial Neural Network and Support Vector Machine was compared. There did not seem to be a big difference in accuracy for predicting the classification of the various tumour types. This is suspected to have to do with the fact that the data set is too incomplete or biased. The data has a high dimensionality, resulting in the computation time for classifiers such multiple-level perceptron classifier and k-nearest neighbour to be very long. This can make classifiers multiple-level perceptron and k-nearest neighbour be less applicable in real life situations. How features selection plays a role in the performance of the classifiers still needs to be researched further however it can be said that the decision tree classifier is better in this regard than the other classifiers used in this research. A few informative genes has been identified through classification by using the decision tree classifier, the accuracy of the classification of the models using only these genes is moderately high. Further research needs to be conducted to see whether these genes do play a factor in the expression of the tumour they classify for. In addition, the feature-selection model may sometimes bring the results with a large difference in accuracy depending on which test set used. The most important take from this research is that the potential of application of data mining classification techniques for the classification of tumour types is there but that much more research with bigger data sets must be conducted.

7 Jupyter notebook used for testing and analysis

See file of jupyter notebook handed in with submissinon of report.

8 References

1 The ubiquitous nature of epistasis in determining susceptibility to common diseases. Hum Hered, Moore JH.

2 An Integrative Multi-Network and Multi-Classifier Approach to Predict Genetic Interactions

3 Machine Learning for Detecting Gene-Gene Interactions: A Review Brett A. McKinney, David M. Reif, Marylyn D. Ritchie, Jason H. Moore Gaurav Pandey, Bin Zhang, Aaron N. Chang, Chad L. Myers, Jun Zhu, Vipin Kumar, Eric E. Schadt

4 Evaluation of Gene Expression Classification Studies: Factors Associated with Classification Performance. Putri W. Novianti, Kit C. B. Roes, Marinus J. C. Eijkemans

5 Cancer classification using gene expression data. Ying Yu, Jiawei Han

6 Feature selection of gene expression data for Cancer classification using double RBF-kernels. Shenghui Liu, Chunrui Xu, Yusen Zhang

7 Performance of error estimators for classification. E.R. Dougherty, C. Sima, B. Hanczar, U.M. Braga-Neto.

8 Comparison of discrimination methods for the classification of tumors using gene expression data. Dudoit S. et al.

9 Supervised feature selection via dependence estimation. L. Song, A. Smola, A. Gretton, K.M. Borgwardt, J. Bedo.

10 Gene expression based cancer classification. Sara Tarek, Reda Abd Elwahab, Mahmoud Shoman.

11 Extreme value distribution based gene selection criteria for discriminant microarray data analysis using logistic regression. W. Li, F. Sun, I. Grosse

12 Novel unsupervised feature filtering of biological data. R. Varshavsky, A. Gottlieb, M. Linial, D. Horn.

13 Single-layer artificial neural networks for gene expression analysis. Narayanana A, Keedwell EC, Gamalielsson J, Tatineni S.

14 An extensive evaluation of recent classification tools applied to microarray data. Computation Statistics and Data Analysis. Lee JW, Lee JB, Park M, Song SH.

15 Radial basis function networks for fast contingency ranking. Devaraj D, Yegnanarayana B, Ramar K.

16 Microarrays: retracing steps. Coombes KR, Wang J, Baggerly KA.

17 Evaluation of Gene Expression Classification Studies: Factors Associated with Classification Performance. Putri W. Novianti, Kit C. B. Roes, Marinus J. C. Eijkemans.

18 The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, et al.

19 Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Ntzani EE1, Ioannidis JP.

20 $http://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq$

21 *https* : *//www.cell.com/pb − assets/consortium/pancanceratlas/pancani*3*/index.html*

22 Chang, Kyle, Chad J Creighton, Caleb Davis, Lawrence Donehower, Jennifer Drummond, David Wheeler, Adrian Ally, et al. "The Cancer Genome Atlas Pan-Cancer analysis project." Nature Genetics 45, no. 10 (September 26, 2013): 1113-1120.

23 *https* : *//portal.gdc.cancer.gov*

24 *https* : *//www.synapse.org/*!*Synapse* : *syn*4301332

25 *https* : *//ourworldindata.org/cancer*

26 Machine learning applications in cancer prognosis and prediction Author links open overlay panel. Konstantina Kouroua, Themis P. Exarchosab, Konstantinos P. Exarchosa, Michalis V.Karamouzisc, Dimitrios I. Fotiadisab.