CATS report

Carlos, Annemijn, Jacobus and Lianne

April 20, 2020

Abstract

Breast cancer is the most common cancer type in women worldwide. However, the diversity among breast tumor types complicates its diagnosis and treatment. Some techniques to identify breast tumor subtypes, like expression profiling microarrays or array comparative genomic hybridization (array-CGH) generate big amounts of information. Machine Learning algorithms have been developed to analyse this information, find patterns within the data, and classify tumor subtypes. Different feature selection methods exist to reduce dimensionality and noise, for a more efficient learning process. Even though previous studies using microarray data investigate the most efficient feature selection method, it is not known whether they can also be applied to array-CGH breast cancer data. In this work, we test performance of three selected feature selection methods, ReliefF, SVM-RFE and Information Gain based on their ability to classify breast cancer subtypes from array-CGH samples.

1 Introduction

Breast cancer is one of the three most prevalent types of cancer in the world, and the most common malignancy in women. Early breast cancer is considered curable, depending on the type of tumor and the presence of metastasis. Therefore, early detection and characterization of the tumor type are essential for an improved prognosis and treatment [6] However, the big molecular, phenotypic, and functional diversity within and among tumors complicate the diagnosis.[4] Breast cancer can be divided into three different molecular subtypes: luminal, HER2-enriched and basal-like. In tumor samples, the subtypes can be identified based on three parameters: presence of oestrogen receptor (ER), presence of progesterone receptor (PgR) and HER2 status. Luminal type is ER or PgR positive and HER2 negative. HER-2+ type is negative for the hormone receptors (HR-) but is HER2 positive. Basal-like tumors are both receptor and HER2 negative (which is why they are also called Triple-Negative) and are associated with the poorest prognosis among the three subtypes.[6]

Several tools are available to identify breast tumor subtype based on gene copy number or expression profiling using microarrays. [11] One of these tools is called array comparative genomic hybridization (array-CGH) analysis. Array-CGH is a technique used to detect chromosomal DNA copy number alterations (CNAs) in samples that suffer from genomic instability. It measures the relative copy number of whole genomes compared to a reference DNA, including those with complex karyotypes.[9] Thanks to the usage of sequence tags, the genes or chromosomal regions with a CNA can be accurately located within a genome. It is therefore widely used for identification, characterization and profiling of tumors. Specifically, CGH-arrays performed in breast cancer samples discovered sets of CNAs, including gains and losses of whole or parts of chromosomes, deletions or gene amplifications, representative for each type of breast cancer.[1]

CGH-arrays and microarray experiments generate large amounts of information. Machine Learning (ML) algorithms have proven to be exceptionally useful for the analysis of complex, multivariate data; discovery of patterns and relationships within this data and further classification into groups or classes. ML methods usually require a pre-processing step (dealing with outliers, missing data, noise. . .) followed by identification of relevant features within the data. This step is called feature selection and aims at reducing the dimensionality of the data for a more efficient learning process. Then, information about the data can be gathered into a model following a learning process. The model can then be used to perform classifications or predictions about new data. [5]

ML algorithms perform better with lower dimensionality. A proper pre-processing of the data followed by an efficient feature selection method are essential for a well-functioning ML method, eliminating irrelevant features and noise. Feature selection is performed before building the classifier, and therefore has wide impact in the whole classification process. There are three different approaches for feature selection, called embedded, filter and wrapper methods. [5] Filters extract features without learning from the data. Wrappers use ML to test the most useful features. Embedded methods merge the feature selection process with the classifier construction. [3]

However, there is a lack of agreement in the field about which feature selection method performs best. Each method depends on the dataset used, some require thousands of variables and sometimes manual annotation and biological interpretation are required. The number of features selected also has a great impact in the model, leading to a common ML problem called “overfitting” when too many features are selected. [5] Various algorithms have been developed to investigate microarray data and help in the classification of tumors from patients based on their microarray profiles. In Hira and Gillies review [3], the feature selection methods ReliefF (wrapper), SVM- Recursive Feature Elimination (SVM-RFE) (wrapper) and Information Gain (filter) were shown to give the highest accuracy in classifying new microarray samples among all the feature selection methods tested when combined with a Support Vector Machine classifier (SVM). However, the most optimal combination of feature selection method and classifier for array-CGH data remains elusive. Considering that each ML method depends on the type of data it is applied to, it is possible that the methods used by Hira and Gillies on microarray data are also applicable to CGH data, or that any other combination of them yields better results.[3]

In this project, high resolution CGH-array data from 100 breast cancer samples was provided. The array contained 244000 probes and measured the quantity of chromosomal DNA per genomic region. Data was pre-processed to show whether each region is a loss (-1), normal (0), gain (+1) or amplification (+2) of the DNA. Another file was provided indicating the associated clinical outcome of each sample, belonging to one of the three tumor subgroups (HER2+, HR+, TN). The aim of this study was to compare the performance of three feature selection methods, named ReliefF, SVM-RFE and Information Gain together with SVM classifier. To do so, different models were created using one of the three feature selection methods. The models were cross-validated using a common scheme to avoid overfitting and their performance to classify breast cancer subtypes from new CGH data was measured using accuracy as parameter. Knowing which feature selection-classification method combination performs best could help in the optimization of array-CGH data analysis, therefore improving breast cancer classification and its treatment.

2 Method

The entire workflow of the research is illustrated in figure 1. The initial input is a data-set with 100 tumor samples with aCGH information. This data-set is referred to as X. The workflow of the research is divided in three steps (three grey blocks in figure 1). In the left grey block, we determine (A) which Features Selection Method in combination with a number of features has the best classification performance. (B,C) To select the best features to use for prediction, the best FSM and number of features combination is applied on the whole data-set X. This will lead to the selection of features FX. In the right upper block, (D) the performance of FX to use in prediction, is validated with a 5-Fold Cross Validation. Lastly, in the right bottom block, (E) the FX features are used on a not seen before data-set T, to predict the breast cancer type per sample, based on aCGH data.

2.1 Array based comparative genomic hybridization (aCGH) and tumor samples (data-set X)

We analyzed genomic DNA of 100 breast cancer samples. These samples where provided with the cancer subtype: HER2+, HR+ or TN. Also the samples where analysed with on a high-resolution array CGH

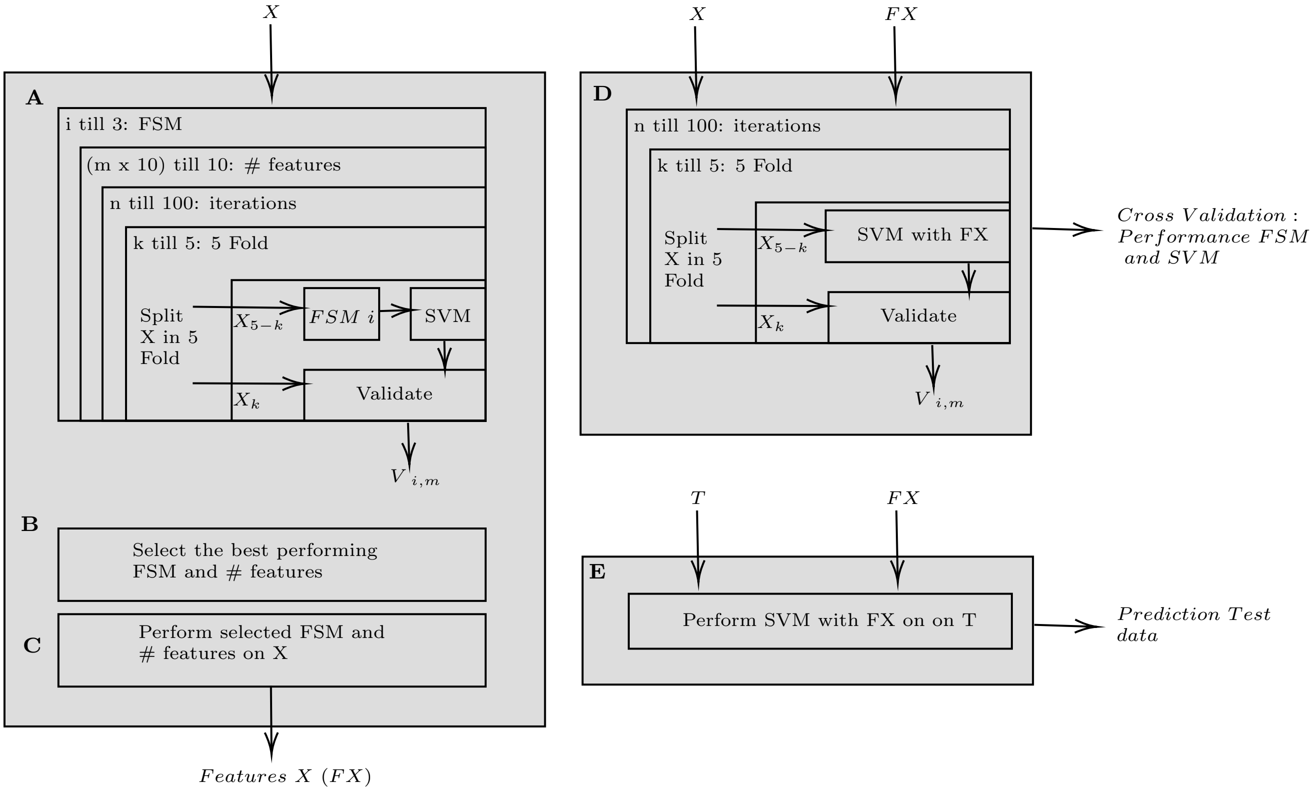


Figure 1: The workflow of the research is divided in 3 steps (3 grey blocks). A) Use the training data-set X to calculate Accuracy percentage (V) for for the 3 different Features Selection Methods (FSM) and 10 different numbers of features to select. B) From A, the best performing FSM in combination with the best number of features is selected. C) The selected FSM and number of features are used to get a feature selection on the training data set X. D) Do Cross Validation to check the performance of the selected FSM and number of features. E) Get the prediction on the test data.

platform with 244,000 probes per array that measures the quantity of chromosomal DNA. For each of these regions and each sample it is determined whether that region is a gain, an amplification, a loss or normal. The data was split in to a training and test set

2.2 Three Feature Selection Methods (FSM)

Three types of feature selection were performed.

Information Gain (InfoGain) InfoGain is a supervised filter selection method, which calculates the dependencies of each feature to the target feature. To do so, the conditional probability is calculated per feature, after which the most dependent N features are chosen for classification [3].

ReliefF ReliefF is a supervised filter selection algorithm that constructs a vector of weights per feature. This vector is created iteratively using a scoring method that relies on a nearest neighbor approach. All instances are placed in a space with respect to their feature values. Of one randomly picked instance, the distance to the nearest instance with the same class (the nearest hit) and the distance to the k nearest instances of all other subtypes (nearest misses) are calculated. The weights vector is updated by decreasing the score of each feature by the value vector of the hit and increasing the score of each feature by the value vector of the nearest misses. In this way, features with varying downstream effects will get higher scores. The highest N-scoring features are then selected for classification [8].

Recursive Feature Elimination (SVM-RFE) Recursive feature elimination is a supervised wrapper selection method that operates by evaluating the contribution of each feature to the outcome of a classifier, which in this case is the SVM classifier. Then, it will recursively eliminate features until a specified number of features is reached.

2.3 Classification with Support Vector Machines (SVM)

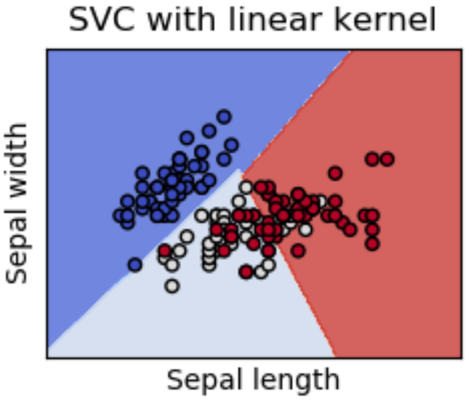


Figure 2: An example of SVC classefier on Sepal lenght and width with a linear kernal. Image from [7]

A supervised Support Vector Machine algorithm was chosen to classify cancer subtype for each sample, based on aCGH data. An SVM will optimize the distance between the separotry lines and the closest data points. See for example figure 2. Although this paper focuses on the performance of the feature selection methods, a classifier with a good performance was needed to get results with a high accuracy. It is shown that SVM algorithms perform well on classifying tumors from aCGH data [10].

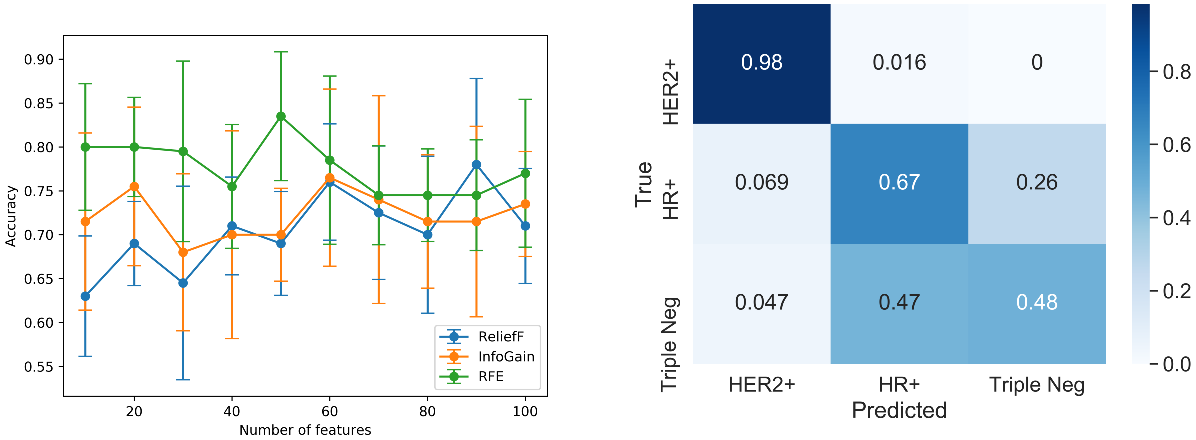
We used the **sklearn.svm.SVC** package in Python to train the algorithm. The kernel was set to linear and the random_state to 0.

2.4 k-Fold Cross Validation

With the best performing FSM and number of features, a selection of features is selected from data-set X. This selection of features (FX) is validated with Cross Validation (see figure 1 step D). The ratio between the train set and validation set that was used in this report was 75:25. This ratio was chosen as an attempt to keep the train set relatively large, while also keeping the validation set large enough to provide an accurate indication of the performance of the algorithm. A validation set that was too small seemed to largely increase the standard deviation of the accuracy scores during cross validation. Each cross validation routine was performed two times with a different random split to train set and validation set, to reduce the probability that the accuracy scores were influenced by random favourable or unfavorable selections of entries in the test set and the train set. As an example of such a random selection, it could be that by chance a test set contains all entries with the HER2+ classification. This would negatively influence the performance of the classifier.

2.5 Parameter Optimization

To investigate the performance of the selected feature selectors, benchmarking was performed with a range of varying parameters. Parameter optimization was done on the following parameters: the different feature selection methods, the number of features, the maximum number of iterations (*max iter*) of the SVC and the number of nearest neighbours of the ReliefF algorithm (k). The most extensively analysed parameter was the number of features, with searched values 10, 20 ... 100. The values for *max iter* that were checked were 800 and 1000. The values for k that were checked were 7,8 and 9.



(a) Accuracy scores

(b) Confusion matrix

Figure 3: Accuracy scores and confusion matrix of the best performing feature selection method.

3 Results

After executing parameter optimization, parameters were chosen for the three feature selection methods (see table 1).

Table 1: Parameters chosen using optimization for three different feature selection methods.

Parameter	ReliefF	InfoGain	SVM-RFE
N Features	90	60	50
max iter	800	800	800
k	9	-	-

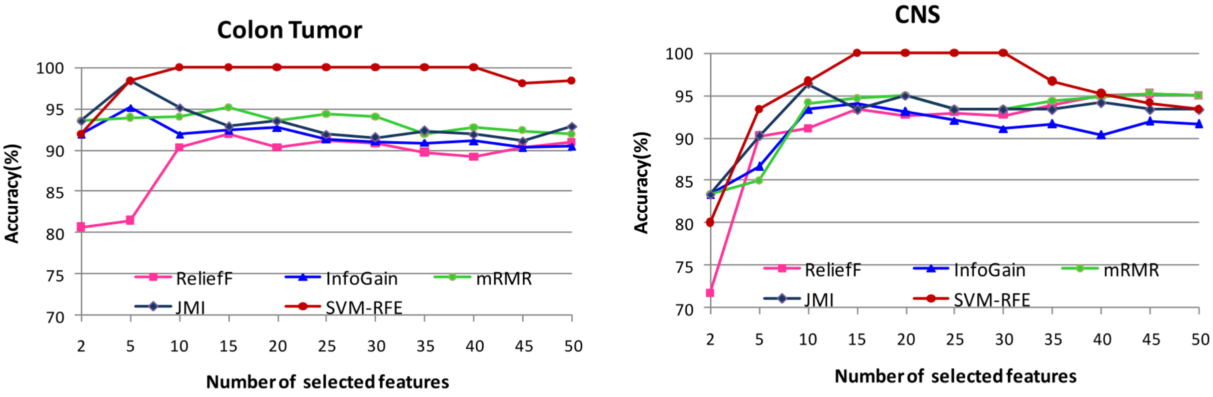
The results of the performance of the feature selection methods for the optimized SVC classifier can be found in figure 3a. Since RFE with 50 features seemed perform best, this feature selection method was chosen to build the classifier. Figure 3b shows the confusion matrix for the RFE method with 50 features. Figure 4 shows the accuracy of a number of feature selection methods for the SVC classifier from the 2018 paper of Cai et al. [2].

4 Discussion

As we can see in figure 3a, the RFE method seems to perform best when 50 of the features are selected. However, as the standard deviation is rather large, it is not possible to draw any definitive conclusions from this. Futhermore, since

Figure 3b shows that the RFE method with 50 features performs best on the classification of HER2+ subtype cancer and performs the worst when classifying Tril Neg.

In figure 4 we can see that for the colon tumor and CNS data sets that were used in the research of Cai *et al.*, the feature selection methods performed much better than the results that were found during cross validation on the breast cancer data set that was used in this report. Although a possible explanation for this could be the limited size of the dataset used, the dataset used by Cai *et al.* was also small with only 60 and 62 samples for the CNS and Colon tumor, respectively. Another reason could be that they investigated a different type of cancer with a different type of data (gene expression data), which could mean that the found accuracies can not be compared in an absolute manner. However, figure 3b does show that the relative differences between the three features shown in figure 4 are similar, since SVM-RFE seems to perform best



(a) Colon Tumor results

(b) CNS results

Figure 4: Figure of Cai et al. (2018) from data-set containing gene expression. Colon: 2000 features, 62 samples and 2 classes. CNS: 7129 features, 60 samples and also 2 classes

in all experiments, and InfoGain and ReliefF came in second and third, respectively. The differences between InfoGain and ReliefF are not extreme, something that is also observed in figure 4.

In this research, an SVM classifier with a linear kernel function was used with different feature selection methods to classify breast cancer subtypes based on CGH-array data. In future research, it might be interesting to analyse and compare the performance of the feature selection methods when different classifiers are used.

5 Conclusion

References

[1] D. G. Albertson. Profiling breast cancer by array cgh. *Breast Cancer Research and Treatment*, 78(3):289–298, 2003.

[2] J. Cai, J. Luo, S. Wang, and S. Yang. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018.

[3] Z. M. Hira and D. F. Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015, 2015.

[4] S. Koren and M. Bentires-Alj. Breast tumor heterogeneity: Source of fitness, hurdle for therapy. *Molecular Cell*, 60(4):537–546, 2015.

[5] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[6] H. N. and G. M. Breast cancer. *Lancet*, 389: 1134–50, 2017.

[7] scikit-learn developers. Plot different SVM classifiers in the iris dataset. https://scikit-learn.org/stable/auto$_{e}xamples/svm/plot_{i}ris_{s}vc.html/.[Online; accessed15 - April - 2020]$.

[8] M. M. L. C. W. O. R. M. J. Urbanowicz, R.J. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 85:189–203, 2018.

[9] E. H. van Beers and P. M. Nederlof. Array-cgh and breast cancer. *Breast Cancer Research*, 8(3), 2006.

[10] M. A. Van de Wiel, F. Picard, W. N. Van Wieringen, and B. Ylstra. Preprocessing and downstream analysis of microarray dna copy number profiles. *Briefings in bioinformatics*, 12(1):10–21, 2011.

[11] L. J. vant Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, and A. T. e. a. Witteveen. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(31 JANUARY 2002):530–535, 2002.