A data mining strategy for the search and classification of gene expression data in cancer

G. Cuder[1], C. Baumgartner[2]

[1]Know-Center GmbH, Graz, Österreich
[2]Institut für Health Care Engineering, Technische Universität Graz, Österreich

gcuder@know-center.at

Abstract— *Cancer is one of the most uprising diseases in our modern society and is defined by an uncontrolled growth of tissue. This growth is caused by mutation on the cellular level. In this thesis, a data-mining workflow was developed to find these responsible genes among thousands of irrelevant ones in three microarray datasets of different cancer types by applying machine learning methods such as classification and gene selection. In this work, four state-of-the-art selection algorithms are compared with a more sophisticated method, termed Stacked-Feature Ranking (SFR), further increasing the discriminatory ability in gene selection.*

Keywords— *cancer, classification, feature selection, microarrays, cross validation*

Introduction

Cancer is one of the most uprising diseases in our modern society due to certain epidemiologic factors like high sugar intake combined with little exercise, smoking and alcohol abuse [1]. Since cancer is mainly developed by genetic mutations of cells enabling them to proliferate uncontrollably [2], medical researchers have gained a lot of knowledge by applying machine learning methods to microarray datasets containing cancerous tissue samples and the expression levels of thousands of corresponding genes. However, only a comparably small subset of these genes carries information on the underlying disease. In recent years, researches have been focusing of finding this subset among all irrelevant genes in the dataset using machine learning approaches such as *classification* and *feature selection (FS)* [3]. In classification, a statistical model is trained on a dataset, containing samples with cancer (case) as well as healthy ones (control), each of them described by a qualitative target Y (*cancer* or *healthy*) and a vector of features $\boldsymbol{X}$ (expression levels of genes) [4]. In the training procedure, the model identifies and learns pattern in the data to be able to classify those samples into the right category (*cancer* or *healthy*). A successfully trained model can then be used to classify unseen samples. One of the most crucial influences that determines the success of the training procedure is the use of informative features (genes) [5]. Therefore, researches are eager to develop robust feature selection algorithms to find and use only relevant genes for cancer classification. In this thesis, a workflow is proposed to find these genes by applying several state-of-the-art FS approaches and novel approach called *Stacked-Feature-Ranking (SFR)* [6] on three microarray datasets of three different cancer types. Furthermore, a Random Forest decision tree model is used to evaluate the gene subsets found by the FS algorithms [7].

Methods

In general, three publicly accessible microarray datasets containing samples of different cancers were used which characteristics are described in table 1.

Table 1: Characteristics of the used datasets

Dataset	Cancer	Control	Features
Lung [8]	97	90	22,215
Prostate [9]	264	160	20,254
Breast [10]	205	205	20,180

In this workflow, four different FS filters are applied, namely information gain, hypothesis testing, [11], relief [12] and minimum description length [13]. A filter algorithm ranks features according to an importance measure in a decreasing manner with the most important feature ranked first. Although filter algorithms are very popular in cancer research due to fast computation, they are often very unstable. In order to overcome this issue, SFR is used to combine the four individual rankings by applying the principle of stacking using a RF decision tree model. The principle of SFR is described in figure 1.

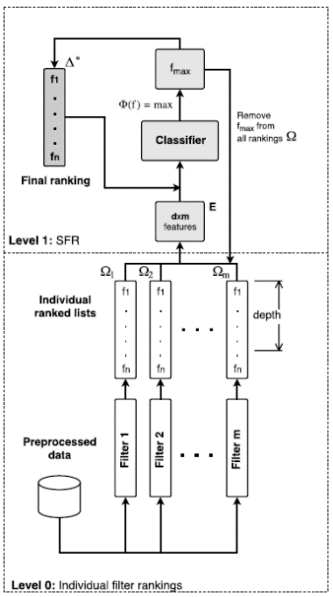

Figure 1: Stacked Feature Ranking algorithm

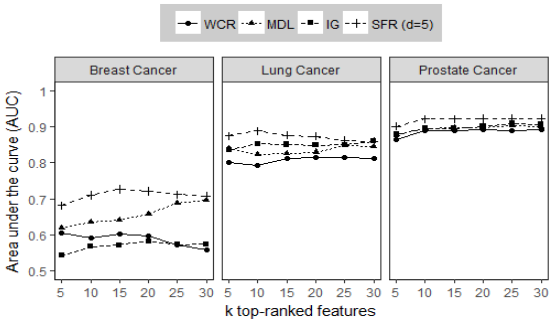

Figure 2: Comparison of FS filters and SFR algorithm

The application of SFR is rather simple, thus the user has only to define a depth parameter d. The algorithm takes the feature rankings produced by the filters as input and takes the d top-ranked features of the level-0 rankings as input. Each of them is then evaluated by the RF model. The feature with the highest discriminatory ability (most relevant gene to classify samples accordingly) is then placed as first feature in the final ranking and removed from all level-0 rankings. Then the process is repeated until all features in the level-0 rankings are removed. The discriminatory ability is described by computing the area under the curve (AUC) which originates in the receiver operating characteristic (ROC) analysis [6].
Classification and FS are validated using 10-fold cross validation, a well-known way of statistical validation in gene expression analysis [3]. This is done by splitting the samples into 10 folds and repeatedly training the model on 9 folds and evaluating the model on the left-out fold, till all folds were used for evaluation once. The result is then represented by the mean AUC of all evaluation folds.

Results

Figure 2 shows the classification results obtained by the described FS approaches on the three datasets. It is clearly shown that SFR results in better classification accuracy than each filter ranking individually. The performance increase of SFR over the other FS algorithms is highest in the breast cancer dataset and lowest in the prostate cancer dataset.

Discussion

In this work, four state-of-the-art FS algorithms were compared to SFR, a novel approach which aggregates the output of these algorithms to a more robust ranking of genes. Hence, this algorithm had never been applied to gene expression datasets before, the performance increase over standard approaches is still very high. Due to the overall stable classification results, further investigation of the gene rankings produced by the SFR algorithm might discover unknown insights in cancer research.

Literature

[1] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, “Global Cancer Statistics: 2011,” *CA. Cancer J. Clin.*, vol. 61, no. 2, pp. 69–90, 2011.

[2] G. Cooper, *The Cell: A Molecular Approach*, 2nd editio. Sunderland (MA): Sinauer Associates, 2000.

[3] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer, 2009.

[5] F. K. Ahmad, N. M. Norwawi, S. Deris, and N. H. Othman, “A review of feature selection techniques via gene expression profiles,” *2008 Int. Symp. Inf. Technol.*, pp. 1–7, 2008.

[6] M. Netzer *et al.*, “A new ensemble-based algorithm for identifying breath gas marker candidates in liver disease using ion molecule reaction mass spectrometry,” *Bioinformatics*, vol. 25, no. 7, pp. 941–947, 2009.

[7] R. Díaz-Uriarte and S. Alvarez de Andrés, “Gene selection and classification of microarray data using random forest.,” *BMC Bioinformatics*, vol. 7, p. 3, 2006.

[8] A. Spira *et al.*, “Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer,” *Nat Med*, vol. 13, no. 3, pp. 361–366, Mar. 2007.

[9] K. L. Penney *et al.*, “Association of prostate cancer risk variants with gene expression in normal and tumor tissue,” *Cancer Epidemiol. Biomarkers Prev.*, vol. 24, no. 1, pp. 255–260, 2015.

[10] A. C. Godfrey *et al.*, “Serum microRNA expression as an early marker for breast cancer risk in prospectively collected samples from the Sister Study cohort.,” *Breast Cancer Res.*, vol. 15, no. 3, p. R42, 2013.

[11] Y. Wang *et al.*, “Gene selection from microarray data for cancer classification - A machine learning approach,” *Comput. Biol. Chem.*, vol. 29, no. 1, pp. 37–46, 2005.

[12] M. Robnik-Šikonja and I. Kononenko, “Theoretical and Empirical Analysis of ReliefF and RReliefF,” *Mach. Learn.*, vol. 53, no. 1–2, pp. 23–69, 2003.

[13] I. Kononenko, “On Biases in Estimating Multi-Valued Attributes,” *Proc. 14th Int. Jt. Conf. Artif. Intell.*, pp. 1034–1040, 1995.