# Feature Selection and Classification Pairwise Combinations for High-dimensional Tumour Biomedical Datasets

A. Dziomdziora  A. Wosiak

Lodz University of Technology
Institute of Information Technology

Theoretical Foundations of Machine Learning, 2015

# Outline

# Introduction

- High-dimensional nature of biomedical data
  - hundreds or thousands of features,
  - a few samples.
- Dimensionality reduction appears to be crucial for the effective classification of tumour samples.
- Solution: dimensionality reduction
  - feature extraction,
  - feature selection.

# Main Objectives

- The goal of the research: to create a comparison of pairwise combinations of feature selection methods and classification techniques applied to the problem of binary and multi-class cancer classification.
- Contribution: to constitute an independent contribution to the relevant literature and try to find a successful way to perform efficient feature selection enhancing accurate classification of tumour specimens.
- Evaluation: six different either binary or multi-class cancer microarray gene expression datasets.

# Outline

# Methodology Foundation

- High-throughput technologies provide the opportunity to examine a large number of biological samples.
- High amounts of multivariate data corresponding to different biological aspects.
- Problem: there are only a few samples available - it increases the risk of overfitting the data and leads to unsatisfactory classification of new data points
- Solution: feature selection

# Methodology Overview

- Data preprocessing, which results in the initial dataset
- Feature selection, which enables the choice of the set of attributes crucial for the automated diagnosis
- Classification process based on the attributes derived from the previous step
- Verification by assessing appropriate comparison criteria

# Data Preprocessing

- Data preprocessing includes two main steps:
  - excluding housekeeping genes,
  - normalization.
- Housekeeping genes
  - take part in basic cell maintenance,
  - may provide serious redundancy and noise into the classification,
  - Affymetrix housekeeping genes identifiers are marked in datasets by the prefix "AFFX-".
- The values in the datasets are normalized - every gene expression value is characterized by mean of zero and unit variance.

# Feature Selection

- Feature selection:
  - improves the generalization performance concerning the model created using the entire set of features,
  - offers a substantially more robust generalization and a faster response with test data,
  - enables researchers to gain a deeper insight into the underlying processes that generated the data.

# Feature Selection

- Seven different approaches were implemented:
  - Correlation-based Feature Selection,
  - Chi-squared,
  - Information Gain,
  - Gain Ratio,
  - Symmetrical Uncertainty,
  - ReliefF,
  - SVM-RFE.
- All of these feature selection methods except for SVM-RFE belong to filter algorithms.

# Classification

- Six different approaches were implemented:
  - J48,
  - logistic model trees,
  - Bayes network,
  - Naïve Bayes,
  - k-nearest neighbours,
  - sequential minimal optimization algorithm for training support vector machines.

# Verification of Results

- Comparison criteria:
  - accuracy,
  - sensitivity,
  - specificity,
  - FP rate,
  - precision,
  - root mean square error,
  - number of features.

# Outline

# Datasets

- binary Colon Cancer Dataset,
- binary Lung Cancer Dataset,
- binary ALL/AML Dataset,
- multiclass Lymphoma Dataset,
- multiclass GCM Dataset,
- binary CNS Dataset.

# Datasets

- Colon Cancer Dataset
  - various patterns of gene expression levels obtained by clustering of tumour and normal colon tissues,
  - 40 tumour biopsies (negatives) and 22 normal biopsies (positives) extracted from colons of the same patients,
  - no missing values in the dataset.
- Lung Cancer Dataset
  - 181 tissue samples: 31 instances belonged to MPM (Malignant Pleural Mesothelioma) and 150 belong to ADCA (Adenocarcinoma) type of the human lung cancer,
  - 12533 genes for each sample,
  - no missing values in the dataset.

# Datasets

- ALL/AML Dataset
  - two acute cases of leukaemia: acute lymphoblastic leukaemia (ALL) and acute myeloblastic leukaemia (AML),
  - training dataset included 38 bone marrow samples (27 ALL and 11 AML), over 7129 probes from 6817 human genes,
  - testing data of 34 observations was provided, with 20 ALL and 14 AML,
  - no missing values in the dataset.
- Lymphoma Dataset
  - distinct types of diffuse large B-cell lymphoma identified by gene expression profiling,
  - 96 observations with 11 classes,
  - 4026 attributes and 19667 missing values in the dataset - missing values were filled in using a filter on the basis of the mean value of each attribute.

# Datasets

- CNS Dataset
  - heterogeneous group of embryonal tumours of the central nervous system (CNS),
  - 60 samples,7129 features in total,
  - two classes: 21 survivors (1) and 39 failures (0),
  - no missing values.
- GCM Dataset
  - Global Cancer Map is a multiclass cancer diagnosis dataset,
  - 190 human tumour examples of 15 types,
  - 16063 attributes in total,
  - 144 samples of training data and 46 samples of testing data,
  - no missing values in the dataset.

# Datasets

| Dataset name | No. of samples | Initial no. of features | No. of features after pre-processing | No of classes |
|---|---|---|---|---|
| ALL/AML | 72 | 7129 | 7070 | 2 |
| CNS | 60 | 7129 | 7070 | 2 |
| Colon | 62 | 2000 | 1988 | 2 |
| Lung | 181 | 12600 | 12533 | 2 |
| Lymphoma | 96 | 4026 | 4026 | 11 |
| GCM | 192 | 16063 | 16004 | 14 |

# Description of Experiments

- The experiments were based on the Weka data mining tool.
- 10-fold cross-validation was used in order to assess the accuracy of the J48, LMT, IBk and SMO.
- The 66% split option was used in the case of Naïve Bayes and Bayes Network classifiers.
- The original division into test set and training set was maintained wherever possible.

# Experimental Results

## The results of classification performed using all the features

| Dataset | Classif. method | No of features | Comparison criteria ACC | SENS | SPEC | FP rate | RMSE |
|---|---|---|---|---|---|---|---|
| ALL / AML | SMO | 7070 | 100.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| CNS | SMO | 7070 | 95.000 | 0.950 | 0.929 | 0.071 | 0.224 |
| Colon | SMO | 1988 | 93.548 | 0.935 | 0.924 | 0.076 | 0.254 |
| Lung | LMT | 12533 | 96.059 | 0.961 | 0.945 | 0.055 | 0.121 |
| Lymphoma | SMO | 4026 | 94.792 | 0.948 | 0.987 | 0.013 | 0.266 |
| GCM | SMO | 16004 | 67.361 | 0.674 | 0.981 | 0.019 | 0.245 |

# Experimental Results

## Best classification results for Information Gain/CFS feature selection

| Dataset | Classif. method | No of features | Comparison criteria ACC | SENS | SPEC | FP rate | RMSE |
|---|---|---|---|---|---|---|---|
| ALL/AML | SMO | 34 | 100.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| Lymphoma | Naive Bayes | 152 | 100.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| Colon | SMO | 27 | 100.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| Lung | SMO | 143 | 98.522 | 0.985 | 0.978 | 0.022 | 0.317 |
| CNS | SMO | 38 | 98.333 | 0.983 | 0.991 | 0.009 | 0.129 |
| GCM | SMO | 42 | 81.250 | 0.813 | 0.989 | 0.011 | 0.243 |

# Experimental Results

## Best classification results for Chi-squared feature selection

| Dataset | Classif. method | No of features | Comparison criteria ACC | SENS | SPEC | FP rate | RMSE |
|---|---|---|---|---|---|---|---|
| ALL/AML | SMO | 150 | 100.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| Lung | SMO | 150 | 97.044 | 0.970 | 0.946 | 0.054 | 0.318 |
| Lymphoma | NaiveBayes | 150 | 93.939 | 0.939 | 0.981 | 0.019 | 0.062 |
| Colon | SMO | 150 | 93.548 | 0.935 | 0.924 | 0.076 | 0.254 |
| CNS | SMO | 150 | 95.000 | 0.950 | 0.929 | 0.071 | 0.224 |
| GCM | SMO | 150 | 62.500 | 0.625 | 0.977 | 0.023 | 0.246 |

# Experimental Results

## Best classification results for InfoGain feature selection

| Dataset | Classif. method | No of features | Comparison criteria ACC | SENS | SPEC | FP rate | RMSE |
|---|---|---|---|---|---|---|---|
| ALL/AML | SMO | 150 | 100.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| CNS | SMO | 150 | 95.000 | 0.950 | 0.929 | 0.071 | 0.224 |
| Colon | SMO | 150 | 93.548 | 0.935 | 0.924 | 0.076 | 0.254 |
| Lung | SMO | 150 | 97.044 | 0.970 | 0.946 | 0.054 | 0.318 |
| Lymphoma | SMO | 150 | 93.750 | 0.938 | 0.984 | 0.016 | 0.266 |
| GCM | SMO | 150 | 59.722 | 0.597 | 0.976 | 0.024 | 0.246 |

# Experimental Results

## Best classification results for Gain Ratio feature selection

| Dataset | Classif. method | No of features | Comparison criteria ACC | SENS | SPEC | FP rate | RMSE |
|---|---|---|---|---|---|---|---|
| ALL/AML | SMO | 150 | 100.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| CNS | SMO | 150 | 95.000 | 0.950 | 0.929 | 0.071 | 0.224 |
| Colon | SMO | 150 | 93.548 | 0.935 | 0.924 | 0.076 | 0.254 |
| Lung | SMO | 150 | 95.074 | 0.951 | 0.913 | 0.087 | 0.319 |
| Lymphoma | LMT | 150 | 78.125 | 0.781 | 0.950 | 0.050 | 0.183 |
| GCM | IBk | 150 | 56.944 | 0.569 | 0.975 | 0.025 | 0.194 |

# Experimental Results

## Best classification results for Symmertical uncertainty feature selection

| Dataset | Classif. method | No of features | Comparison criteria ACC | SENS | SPEC | FP rate | RMSE |
|---|---|---|---|---|---|---|---|
| ALL/AML | SMO | 150 | 100.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| Colon | SMO | 150 | 93.548 | 0.935 | 0.924 | 0.076 | 0.254 |
| Lung | Naive Bayes | 150 | 98.551 | 0.986 | 0.967 | 0.033 | 0.076 |
| Lymphoma | SMO | 150 | 93.750 | 0.938 | 0.986 | 0.014 | 0.266 |
| CNS | SMO | 150 | 95.000 | 0.950 | 0.929 | 0.071 | 0.224 |
| GCM | LMT | 150 | 58.333 | 0.583 | 0.973 | 0.027 | 0.213 |

# Experimental Results

## Best classification results for ReliefF feature selection

| Dataset | Classif. method | No of features | Comparison criteria ACC | SENS | SPEC | FP rate | RMSE |
|---|---|---|---|---|---|---|---|
| ALL/AML | Naive Bayes | 150 | 100.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| Lymphoma | Naive Bayes | 150 | 100.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| Colon | SMO | 150 | 88.709 | 0.887 | 0.887 | 0.123 | 0.336 |
| Lung | Naive Bayes | 150 | 98.551 | 0.986 | 0.967 | 0.033 | 0.076 |
| CNS | SMO | 150 | 86.677 | 0.867 | 0.818 | 0.182 | 0.365 |
| GCM | LMT | 150 | 58.333 | 0.583 | 0.620 | 0.026 | 0.206 |

# Experimental Results

## Best classification results for Information Gain/SVM-RFE feature selection

| Dataset | Classif. method | No of features | Comparison criteria | | | | |
|---|---|---|---|---|---|---|---|
| | | | ACC | SENS | SPEC | FP rate | RMSE |
| ALL/AML | Naive Bayes | 150 | 100.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| Colon | SMO | 150 | 95.161 | 0.952 | 0.932 | 0.068 | 0.220 |
| Lung | SMO | 150 | 98.522 | 0.985 | 0.978 | 0.022 | 0.317 |
| Lymphoma | Naive Bayes | 150 | 100.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| CNS | SMO | 150 | 91.667 | 0.917 | 0.889 | 0.111 | 0.289 |
| GCM | SMO | 150 | 73.611 | 0.736 | 0.984 | 0.016 | 0.244 |

# Experimental Results

## Comparison of no. of features and accuracy with and without FS

| Dataset | No of features without FS | No of features after FS | Features reduction [%] | ACC without FS | ACC after FS | ACC diff. [%] |
|---|---|---|---|---|---|---|
| ALL/AML | 7070 | 34 | 99.52 | 100.00 | 100.00 | 0.00 |
| CNS | 7070 | 150 | 97.88 | 95.00 | 95.00 | 0.00 |
| Colon | 1988 | 150 | 92.45 | 93.55 | 95.16 | +1.61 |
| Lung | 12533 | 150 | 98.80 | 96.06 | 98.55 | +0.967 |
| Lymphoma | 4026 | 150 | 96.27 | 94.79 | 93.94 | -0.85 |
| GCM | 16004 | 42 | 99.74 | 67.36 | 81.25 | +13.89 |

# Outline

# Conclusions and Future Work

- The classification of high-dimensional biomedical datasets is regarded as a challenging task.
- The enormous dimensionality of the microarray expression data is a serious concern during gene selection.
- Multi-class classification issues are more difficult than the binary ones - researches are conducted and often succeed in new approaches.

# Conclusions and Future Work

- It was demonstrated that the hybrid strategies (classification algorithms and feature selection methods) resulted in more satisfactory outcomes.
- The SMO classifier outperforms other classification methods in the majority of cases.
- The SVM-RFE algorithm combined with SMO classification was considered as the most beneficial choice for constructing the learning model.

# Conclusions and Future Work

- Future works:
  - to involve other algorithms and strategies,
  - other combinations of various classifiers and attribute selectors should be investigated in depth,
  - the results of our research can be further implemented in practice for Lodz Medical University Hospital No 4.

Thank you for your attention.