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
1. Introduction

2. Methodology
   - Methodology Overview
   - Data Preprocessing
   - Feature Selection
   - Classification
   - Verification of Results

3. Case Study and Experimental Results
   - Data Description
   - Experiments Assumptions
   - Experimental Results

4. Conclusions
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

1 Introduction

2 Methodology
   - Methodology Overview
   - Data Preprocessing
   - Feature Selection
   - Classification
   - Verification of Results

3 Case Study and Experimental Results
   - Data Description
   - Experiments Assumptions
   - Experimental Results

4 Conclusions
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.
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.
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

1. Introduction

2. Methodology
   - Methodology Overview
   - Data Preprocessing
   - Feature Selection
   - Classification
   - Verification of Results

3. Case Study and Experimental Results
   - Data Description
   - Experiments Assumptions
   - Experimental Results

4. Conclusions
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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. of samples</th>
<th>Initial no. of features</th>
<th>No. of features after pre-processing</th>
<th>No of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL/AML</td>
<td>72</td>
<td>7129</td>
<td>7070</td>
<td>2</td>
</tr>
<tr>
<td>CNS</td>
<td>60</td>
<td>7129</td>
<td>7070</td>
<td>2</td>
</tr>
<tr>
<td>Colon</td>
<td>62</td>
<td>2000</td>
<td>1988</td>
<td>2</td>
</tr>
<tr>
<td>Lung</td>
<td>181</td>
<td>12600</td>
<td>12533</td>
<td>2</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>96</td>
<td>4026</td>
<td>4026</td>
<td>11</td>
</tr>
<tr>
<td>GCM</td>
<td>192</td>
<td>16063</td>
<td>16004</td>
<td>14</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classif.</th>
<th>No of features</th>
<th>ACC</th>
<th>SENS</th>
<th>SPEC</th>
<th>FP rate</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL / AML</td>
<td>SMO</td>
<td>7070</td>
<td>100.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>CNS</td>
<td>SMO</td>
<td>7070</td>
<td>95.000</td>
<td>0.950</td>
<td>0.929</td>
<td>0.071</td>
<td>0.224</td>
</tr>
<tr>
<td>Colon</td>
<td>SMO</td>
<td>1988</td>
<td>93.548</td>
<td>0.935</td>
<td>0.924</td>
<td>0.076</td>
<td>0.254</td>
</tr>
<tr>
<td>Lung</td>
<td>LMT</td>
<td>12533</td>
<td>96.059</td>
<td>0.961</td>
<td>0.945</td>
<td>0.055</td>
<td>0.121</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>SMO</td>
<td>4026</td>
<td>94.792</td>
<td>0.948</td>
<td>0.987</td>
<td>0.013</td>
<td>0.266</td>
</tr>
<tr>
<td>GCM</td>
<td>SMO</td>
<td>16004</td>
<td>67.361</td>
<td>0.674</td>
<td>0.981</td>
<td>0.019</td>
<td>0.245</td>
</tr>
</tbody>
</table>
## Experimental Results

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classif.</th>
<th>No of features</th>
<th>ACC</th>
<th>SENS</th>
<th>SPEC</th>
<th>FP rate</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL/AML</td>
<td>SMO</td>
<td>34</td>
<td>100.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>Naive Bayes</td>
<td>152</td>
<td>100.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Colon</td>
<td>SMO</td>
<td>27</td>
<td>100.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Lung</td>
<td>SMO</td>
<td>143</td>
<td>98.522</td>
<td>0.985</td>
<td>0.978</td>
<td>0.022</td>
<td>0.317</td>
</tr>
<tr>
<td>CNS</td>
<td>SMO</td>
<td>38</td>
<td>98.333</td>
<td>0.983</td>
<td>0.991</td>
<td>0.009</td>
<td>0.129</td>
</tr>
<tr>
<td>GCM</td>
<td>SMO</td>
<td>42</td>
<td>81.250</td>
<td>0.813</td>
<td>0.989</td>
<td>0.011</td>
<td>0.243</td>
</tr>
</tbody>
</table>
## Experimental Results

### Best classification results for Chi-squared feature selection

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classif.</th>
<th>No of</th>
<th>Comparison criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>method</td>
<td>features</td>
<td>ACC</td>
</tr>
<tr>
<td>ALL/AML</td>
<td>SMO</td>
<td>150</td>
<td>100.000</td>
</tr>
<tr>
<td>Lung</td>
<td>SMO</td>
<td>150</td>
<td>97.044</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>NaiveBayes</td>
<td>150</td>
<td>93.939</td>
</tr>
<tr>
<td>Colon</td>
<td>SMO</td>
<td>150</td>
<td>93.548</td>
</tr>
<tr>
<td>CNS</td>
<td>SMO</td>
<td>150</td>
<td>95.000</td>
</tr>
<tr>
<td>GCM</td>
<td>SMO</td>
<td>150</td>
<td>62.500</td>
</tr>
</tbody>
</table>
Experimental Results

Best classification results for InfoGain feature selection

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classif.</th>
<th>No of features</th>
<th>ACC</th>
<th>SENS</th>
<th>SPEC</th>
<th>FP rate</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL/AML</td>
<td>SMO</td>
<td>150</td>
<td>100.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>CNS</td>
<td>SMO</td>
<td>150</td>
<td>95.000</td>
<td>0.950</td>
<td>0.929</td>
<td>0.071</td>
<td>0.224</td>
</tr>
<tr>
<td>Colon</td>
<td>SMO</td>
<td>150</td>
<td>93.548</td>
<td>0.935</td>
<td>0.924</td>
<td>0.076</td>
<td>0.254</td>
</tr>
<tr>
<td>Lung</td>
<td>SMO</td>
<td>150</td>
<td>97.044</td>
<td>0.970</td>
<td>0.946</td>
<td>0.054</td>
<td>0.318</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>SMO</td>
<td>150</td>
<td>93.750</td>
<td>0.938</td>
<td>0.984</td>
<td>0.016</td>
<td>0.266</td>
</tr>
<tr>
<td>GCM</td>
<td>SMO</td>
<td>150</td>
<td>59.722</td>
<td>0.597</td>
<td>0.976</td>
<td>0.024</td>
<td>0.246</td>
</tr>
</tbody>
</table>
# Experimental Results

## Best classification results for Gain Ratio feature selection

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classif.</th>
<th>No of features</th>
<th>ACC</th>
<th>SENS</th>
<th>SPEC</th>
<th>FP rate</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL/AML</td>
<td>SMO</td>
<td>150</td>
<td>100.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>CNS</td>
<td>SMO</td>
<td>150</td>
<td>95.000</td>
<td>0.950</td>
<td>0.929</td>
<td>0.071</td>
<td>0.224</td>
</tr>
<tr>
<td>Colon</td>
<td>SMO</td>
<td>150</td>
<td>93.548</td>
<td>0.935</td>
<td>0.924</td>
<td>0.076</td>
<td>0.254</td>
</tr>
<tr>
<td>Lung</td>
<td>SMO</td>
<td>150</td>
<td>95.074</td>
<td>0.951</td>
<td>0.913</td>
<td>0.087</td>
<td>0.319</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>LMT</td>
<td>150</td>
<td>78.125</td>
<td>0.781</td>
<td>0.950</td>
<td>0.050</td>
<td>0.183</td>
</tr>
<tr>
<td>GCM</td>
<td>IBk</td>
<td>150</td>
<td>56.944</td>
<td>0.569</td>
<td>0.975</td>
<td>0.025</td>
<td>0.194</td>
</tr>
</tbody>
</table>
# Experimental Results

Best classification results for Symmetrical uncertainty feature selection

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classif.</th>
<th>No of features</th>
<th>ACC</th>
<th>SENS</th>
<th>SPEC</th>
<th>FP rate</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL/AML</td>
<td>SMO</td>
<td>150</td>
<td>100.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Colon</td>
<td>SMO</td>
<td>150</td>
<td>93.548</td>
<td>0.935</td>
<td>0.924</td>
<td>0.076</td>
<td>0.254</td>
</tr>
<tr>
<td>Lung</td>
<td>Naive Bayes</td>
<td>150</td>
<td>98.551</td>
<td>0.986</td>
<td>0.967</td>
<td>0.033</td>
<td>0.076</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>SMO</td>
<td>150</td>
<td>93.750</td>
<td>0.938</td>
<td>0.986</td>
<td>0.014</td>
<td>0.266</td>
</tr>
<tr>
<td>CNS</td>
<td>SMO</td>
<td>150</td>
<td>95.000</td>
<td>0.950</td>
<td>0.929</td>
<td>0.071</td>
<td>0.224</td>
</tr>
<tr>
<td>GCM</td>
<td>LMT</td>
<td>150</td>
<td>58.333</td>
<td>0.583</td>
<td>0.973</td>
<td>0.027</td>
<td>0.213</td>
</tr>
</tbody>
</table>
## Experimental Results

### Best classification results for ReliefF feature selection

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classif.</th>
<th>No of features</th>
<th>ACC</th>
<th>SENS</th>
<th>SPEC</th>
<th>FP rate</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL/AML</td>
<td>Naive Bayes</td>
<td>150</td>
<td>100.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>Naive Bayes</td>
<td>150</td>
<td>100.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Colon</td>
<td>SMO</td>
<td>150</td>
<td>88.709</td>
<td>0.887</td>
<td>0.887</td>
<td>0.123</td>
<td>0.336</td>
</tr>
<tr>
<td>Lung</td>
<td>Naive Bayes</td>
<td>150</td>
<td>98.551</td>
<td>0.986</td>
<td>0.967</td>
<td>0.033</td>
<td>0.076</td>
</tr>
<tr>
<td>CNS</td>
<td>SMO</td>
<td>150</td>
<td>86.677</td>
<td>0.867</td>
<td>0.818</td>
<td>0.182</td>
<td>0.365</td>
</tr>
<tr>
<td>GCM</td>
<td>LMT</td>
<td>150</td>
<td>58.333</td>
<td>0.583</td>
<td>0.620</td>
<td>0.026</td>
<td>0.206</td>
</tr>
</tbody>
</table>
# Experimental Results

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classif.</th>
<th>No of features</th>
<th>Comparison criteria</th>
<th>Method</th>
<th>ACC</th>
<th>SENS</th>
<th>SPEC</th>
<th>FP rate</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL/AML</td>
<td>Naive Bayes</td>
<td>150</td>
<td></td>
<td>ACC</td>
<td>100.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Colon</td>
<td>SMO</td>
<td>150</td>
<td></td>
<td>SENS</td>
<td>0.952</td>
<td>0.932</td>
<td>0.068</td>
<td>0.220</td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>SMO</td>
<td>150</td>
<td></td>
<td>SPEC</td>
<td>0.985</td>
<td>0.978</td>
<td>0.022</td>
<td>0.317</td>
<td></td>
</tr>
<tr>
<td>Lymphoma</td>
<td>Naive Bayes</td>
<td>150</td>
<td></td>
<td>FP rate</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>CNS</td>
<td>SMO</td>
<td>150</td>
<td></td>
<td>RMSE</td>
<td>0.917</td>
<td>0.889</td>
<td>0.111</td>
<td>0.289</td>
<td></td>
</tr>
<tr>
<td>GCM</td>
<td>SMO</td>
<td>150</td>
<td></td>
<td></td>
<td>73.611</td>
<td>0.736</td>
<td>0.984</td>
<td>0.016</td>
<td>0.244</td>
</tr>
</tbody>
</table>
Experimental Results

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

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

1. Introduction

2. Methodology
   - Methodology Overview
   - Data Preprocessing
   - Feature Selection
   - Classification
   - Verification of Results

3. Case Study and Experimental Results
   - Data Description
   - Experiments Assumptions
   - Experimental Results

4. Conclusions
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.