Saturday, 12th July 2014

Classification and Feature Selection across the ICGC Cancer Projects in the CAMDA 2014 Challenge

Jari Björne, Antti Airola, Tapio Pahikkala and Tapio Salakoski

Department of Information Technology, University of Turku

Turku Centre for Computer Science (TUCS)

Joukahaisenkatu 3-5, 20520 Turku, Finland

firstname.lastname@utu.fi

Introduction

- ICGC (the International Cancer Genome Consortium) coordinates over 40 cancer type projects across 10 nations
- In the CAMDA 2014 challenge, **three less-studied cancer types** from the ICGC dataset are proposed for analysis
  - Head and Neck Squamous Cell Carcinoma (HNSC-US)
  - Lung Adenocarcinoma (LUAD-US)
  - Kidney Renal Clear Cell Carcinoma (KIRC-US)
- **Two questions** are proposed
  - What are disease causal changes?
  - Can personalized medicine and rational drug treatment plans be derived from the data?

Our Approach

- We approach the task through state-of-the-art machine learning methods
- Our work proceeds in three consecutive steps
  1) **Classify** cancer samples into subgroups
  2) Use **feature selection** to uncover the genetic basis behind the classification
  3) Use **external knowledge** (from NCI Cancer Gene Index) to assess the biological relevance of the models

TUCS

Resources

Resources

1) Data from the ICGC Data Portal (https://dcc.icgc.org/)
2) Machine learning
   - RLS from RLScore (http://users.utu.fi/aatapa/RLScore/)
   - SVM from scikit-learn (http://scikit-learn.org)
   - Extra Trees Classifier from scikit-learn
3) NCI Cancer Gene Index for verifying results (https://wiki.nci.nih.gov/x/8i1yAQ)

ICGC Datasets (1/2)

- We downloaded all 42 cancer projects from the ICGC Data Portal (https://dcc.icgc.org/)
- The data comes in several detailed TSV-tables → However, conceptually the data can be thought of as
  - Patient sample / patient status pairs (*classes*)
    - both cancer tissue and normal control samples
  - Patient sample / biochemical feature pairs (*features*)
- We stored the data in an SQLite database (≈ 20 Gb)
  - Fast random access
  - Easy to construct complex views into the data

ICGC Datasets (2/2)

- All datatypes were used except for methylation
  - SSM (Simple somatic mutation)
  - CNSM (Copy number somatic mutation)
  - EXP (Expression microarray)
  - PEXP (Protein expression)
  - miRNA (micro-RNA)
- Methylation is a promising source but was just too big to handle through SQLite
  - Several gigabytes of methylation data per cancer project
  - SQLite compression might help but is expensive

NCI Cancer Gene Index

- CGI provides pairs of
  - **genes** associated with **cancer**
  - **drugs** known to affect these **genes**
- CGI is manually verified
  - Associations are automatically text-mined from MEDLINE abstracts
  - All text-mined associations are manually verified
  - Connects 6,955 unique human genes to nearly 12,000 cancer terms and 2,180 drug terms from the NCI Thesaurus
- We convert the original CGI XML files into an SQLite database (≈ 300 Mb)

Machine Learning

- We use methods for **classification** and **feature selection**
- Three systems are tested
  - SVM (Support Vector Machine, scikit-learn)
  - RLS (Regularized Least Squares, RLSCore)
  - ETC (Extra Trees Classifier, scikit-learn)
- SVM and RLS
  - Representative examples for methods based on the Tikhonov regularization framework
  - Generally good classification performance
  - RLS can perform greedy forward feature selection
- The Extra Trees Classifier
  - An ensemble of extremely randomized trees
  - Classification performance less well known
  - Provides embedded feature selection

TUCS

Classification Tasks

Classification Tasks

- We define two classification tasks
  - **Cancer:** Separate normal controls from cancer tissue
  - **Remission**: Whether a sample comes from a tumour that will go into remission or from one that will progress until the death of the donor.
- Experimental goals
  - An automatically learned model for extraction of relevant genes through feature analysis
    - Cancer-task for cancer-relevant genes
    - Remission-task for remission-relevant genes
  - A model for disease prognostic predictions (remission task)

Machine Learning Examples

- Examples for classification tasks are built from the ICGC data
- A random 30% of the donors of each cancer type are left as a hidden set for final performance estimation
- The remaining 70% are used for parameter optimization
  - SVM: 5-fold cross-validation
  - RLS: Leave-one-out cross-validation (through RLScore)
  - Extra Trees Classifier: Use 10,000 trees
    - Still fits into memory
    - Larger ensembles stabilize the feature selection as shown in *J. Paul, M. Verleysen, and P. Dupont. The stability of feature selection and class prediction from ensemble tree classifiers. In ESANN 2012, 2012.*
- For each donor, only one example (cancer or control) is used per experiment to avoid dependencies in cross-validation and evaluation

Machine Learning Features

- All features are derived from the ICGC data
- Five feature groups are used
  - **SSM** (Simple somatic mutation) binary features
    - Gene affected 1/0 (e.g. KRT13 = 1)
    - Point-mutation on gene 1/0 (e.g. KRT13_R272C = 1)
  - **CNSM** (Copy number somatic mutation) binary feature
    - Chromosomal position + (gene name + mutation type)
  - Normalized expression levels from
    - **EXP** (gene expression microarray)
    - **PEXP** (protein expression)
    - **miRNA** (micro-RNA)
- EXP is the most commonly and consistently available ICGC data type

Classification Results

Classification Performance

- Performance measured with EXP (gene expression) features
- ***Cancer or normal*** classification
  - Best AUCs in 0.99-1.0 range for the CAMDA cancer types
  - Almost perfect performance can be reached easily → Radical metabolic changes due to cancer?
  - Small number of analyzed normal control samples in the ICGC data limits class size
- ***Remission or progression*** classification
  - A more difficult task with best AUCs in 0.74-0.85 range for the CAMDA cancer types
  - Different feature groups can increase performance (SVM)

Project	All	CNSM	EXP	MIRNA	PEXP	SSM
HNSC-US	0.64	0.74	0.63	0.74	0.69	-
KIRC-US	0.83	0.83	0.83	0.86	0.85	0.83
LUAD-US	0.79	0.76	0.79	0.79	0.76	-

Classification Performance

Table 1: Classification task performance and example counts. In AUC_X R=RLS, S=SVM and E=Extra Trees Classifier.

project	cancer	normal	AUC_R	AUC_S	AUC_E	remission	progression	AUC_R	AUC_S	AUC_E
BLCA-US	169	16	-	0.94	0.98	114	32	-	0.76	0.78
BRCA-US	853	109	-	0.98	0.98	729	50	-	0.93	0.93
CESC-US	-	-	-	-	-	41	9	-	0.85	0.90
CLLE-ES	-	-	-	-	-	46	24	-	0.48	0.52
COAD-US	361	41	-	1.00	1.00	179	26	-	0.82	0.84
GBM-US	-	-	-	-	-	26	388	-	0.93	0.96
HNSC-US	314	39	0.99	0.97	0.98	208	80	0.60	0.63	0.74
KIRC-US	426	72	0.95	0.99	0.99	333	92	0.71	0.83	0.85
KIRP-US	99	28	-	0.93	1.00	96	7	-	1.00	0.95
LGG-US	-	-	-	-	-	103	46	-	0.74	0.81
LIHC-US	77	46	-	0.91	0.95	73	26	-	0.71	0.71
LUAD-US	389	55	1.00	0.98	0.98	238	60	0.57	0.79	0.80
LUSC-US	359	44	-	0.95	1.00	228	43	-	0.79	0.81
OV-US	-	-	-	-	-	139	260	-	0.55	0.60
PAAD-US	-	-	-	-	-	11	11	-	0.33	0.78
PRAD-US	136	38	-	0.94	0.96	-	-	-	-	-
READ-US	137	8	-	1.00	0.97	-	-	-	-	-
RECA-EU	46	45	-	1.00	-	-	-	-	-	-
SKCM-US	-	-	-	-	-	137	115	-	0.63	0.63
THCA-US	426	53	-	0.93	0.97	-	-	-	-	-
UCEC-US	450	22	-	1.00	1.00	365	30	-	0.88	0.93

Feature Selection and Analysis

Feature Selection Techniques

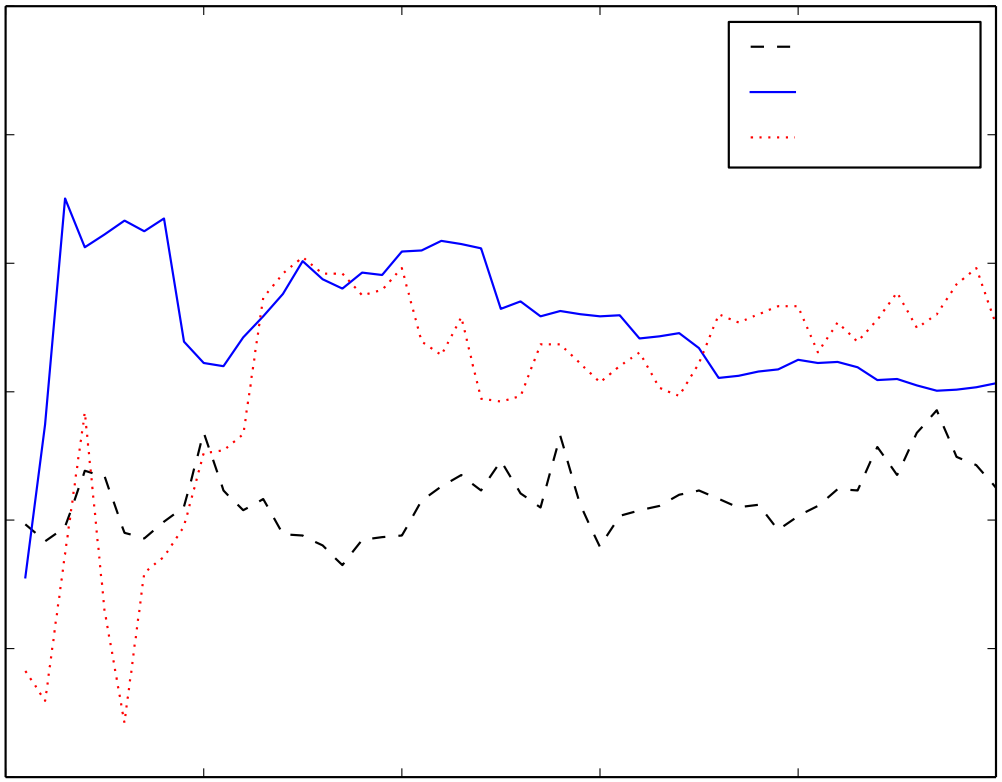
- **Filter-based selection methods** used outside the classifier tell which individual features are the best in isolation
  - → Useful for ***explaining*** e.g. the causes of a disease
- To obtain best performance with the lowest amount of features **greedy forward selection** methods can be efficient
  - → Useful for producing a ***computationally efficient model***
- **Embedded methods** that are part of e.g. ensemble classifiers can improve performance and stability of the model
  - → Useful for producing a ***model with good predictive performance***
- We use the latter two methods to select and analyze EXP-features (genes)

Greedy RLS feature selection (1/2)

- Greedy Selection
  - Start from an empty set
  - Select the feature that best correlates with the classes
  - Continue by selecting at each round the feature that most improves performance together with the already selected features
- Built-in to the RLScore package
- Results show how many features are required to reach optimal performance

Greedy RLS feature selection (2/2)

- Very few EXP features (genes) are required for optimal performance
- Performance is very unstable, likely due to small class sizes

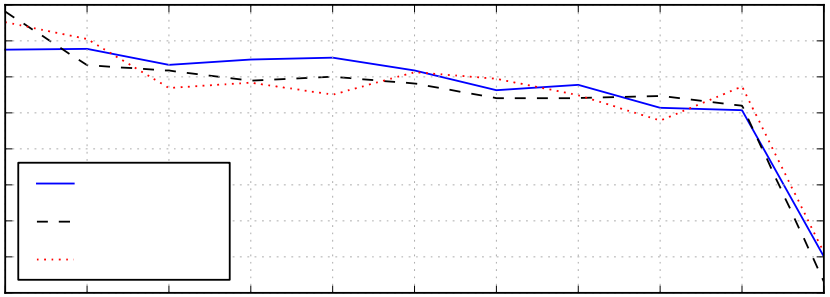


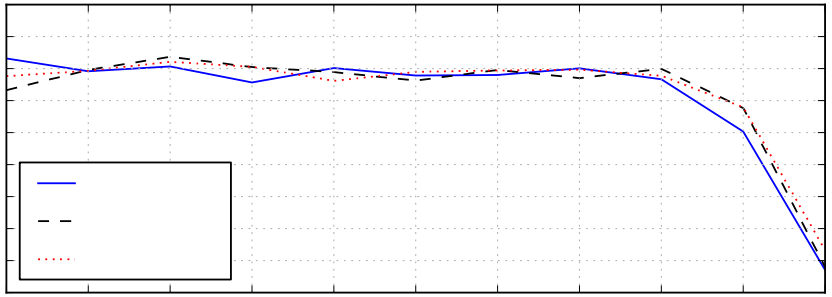
ETC Feature Analysis

- The Extra Trees Classifier (ETC) provides embedded feature selection, a common feature of ensemble methods
- Feature selection with the Extra Trees Classifier is stabilized by increasing the ensemble size
- Selected features are analyzed by comparing to the manually curated NCI Cancer Gene Index

NCI Cancer Gene Index Analysis (1/2)

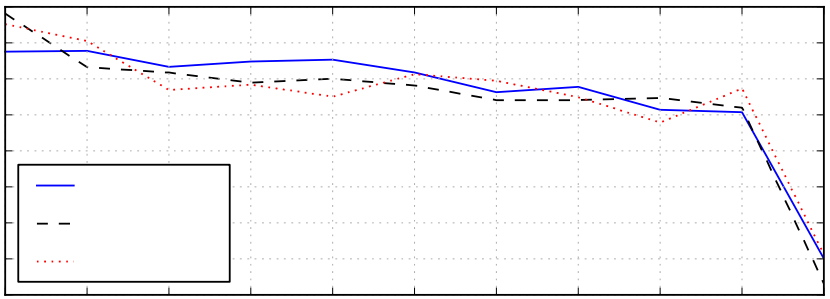
- To evaluate the **biological significance** of selected gene features we checked how many were known in the manually curated NCI Cancer Gene Index
- Known cancer genes are **slightly more common** in the first deciles of the selected features of the *cancer or control* task
- Known cancer genes were **much less common** among features not selected at all

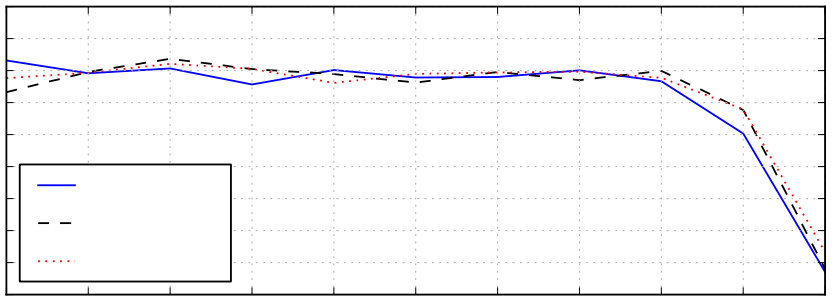




NCI Cancer Gene Index Analysis (2/2)

- For the ***cancer or control*** classification, the results hint at a real biological basis for the model (as opposed to e.g. an artifact arising from sample handling)
- For the ***remission or progression*** classification known cancer genes are more balanced across the feature selection deciles, as is expected in a within-cancer classification.





Top Selected Features (1/2)

- We next look at the top individual gene features
  - Gene features are selected with the Extra Trees Classifier
  - For genes that match an NCI Cancer Gene Index entry we can retrieve:
    - annotated cancer terms
    - cancer drugs known to have an effect
- Cancer-type relevant annotations can be found in top HNSC-US and LUAD-US genes when searching with keywords:
  - HNSC-US ("head and neck", "head", "neck", "squamous")
  - LUAD-US ("lung", "adenocarcinoma")
  - KIRC-US ("kidney", "renal", "clear cell")
- Extra Trees feature selection is stable only on a large scale → too many conclusions should not be drawn from individual features

Top Selected Features (2/2)

Table 2: Top 5 features for Extra Trees Classifier cancer/normal classification, with counts and examples for matching NCI Cancer Gene Index cancer and drug terms.

P	gene	n(c)	cancer term, most relevant or most common	n(d)	drug term, most common
KIRC-US	KNG1	170	9l gliosarcoma [C3796]	460	acetazolamide [C28809]
	KCNJ1	0	-	0	-
	HS6ST2	0	-	0	-
	GGT6	0	-	0	-
	KCNJ10	1	brain tumours [C2907]	0	-
HNSC-US	KRT13	232	head and neck squamous cell carcinoma [C34447]	44	5-f uorouracil [C505]
	CAB39L	0	-	0	-
	FAM3D	0	-	0	-
	BARX2	14	epithelial ovarian cancer [C4908]	7	platinum [C376]
	KRT4	46	oral squamous-cell carcinoma [C4833]	4	liarozole [C1433]
LUAD-US	SFTPC	73	adenocarcinoma of the lung [C3512]	58	all-trans-retinoic acid [C900]
	TNNC1	0	-	0	-
	LGI3	0	-	0	-
	STX11	8	acute myelogenous leukaemia [C3171]	0	-
	CAT	10	non-small cell lung carcinoma [C2926]	10	antioxidant [C275]

TUCS

Conclusions

Conclusions (1/2)

- We developed models for ***cancer or control*** and ***remission or progression*** classification
  - Good prognostic performance on datasets such as KIRC-US
- Feature selection was used to analyse the classification process and to extract the gene sets it relies on (**CAMDA Question 1**)
  - The automatically learned models were evaluated by comparing to the NCI Cancer Gene Index
- Gene–drug associations from the NCI Cancer Gene Index may provide paths towards utilization of learned gene–cancer associations in personalized medicine (**CAMDA Question 2**)

Conclusions (2/2)

- The feature selection approaches used have still a lot of noise
  - Stabilizing these systems on the limited size biological datasets is a clear direction for future work
- During this work we have developed efficient methods for fast handling of the large datasets involved
- All our experimental code is provided for the community under an open source license at https://github.com/jbjorne/CAMDA2014

TUCS

Thank You!