Prediction and classification of cancer

using artificial neural networks

Marcus Birgersson

Complex Systems Seminars, 2014/2015

Outline

Outline

Background and problem formulation

Cancer background

Relevant cancer types

- neuroblastoma (NB)
- rhabdomyosarcoma (RMS)
- non-Hodgkin lymphoma (NHL)
- Ewing family of tumors (EWS)

Collective name: The small, round blue cell tumors (SRBCTs)

Problem with diagnosis

- Similar apperence on routine histology
- accurate diagnosis of SRBCTs is essential since treatment vary widely depending on the diagnosis
- no single test can precisely distinguish these cancers

Testing difficulties

- The tests are slow
- Does not always prove a definite diagnosis
- Techniques for diagnosis:
  - Immunohistochemistry (Allows for protein expression, one gene at a time)
  - Cytogenesis
  - Interphase flourescence
  - Reverse transcription (do not always provide a definitive diagnosis)
  - **Gene expression profiling using cDNA microarrays**

Background and problem formulation

Cancer background

Why using gene expressions?

- Gene-expression profiling permits simultaneous analysis of multiple markers
- Has been used to categorize cancers into subgroups
- No method has so far been rigorously tested for it's ability to accurately distinguish cancers belonging to several diagnostic categories

Hypothesis:

Using an Artificial Neural Network for pattern recognition on these data

Outline

Creation of ANN model

Defining ANN model

ANN model

- An artificial neural network consisting of 10 inputs and 4 outputs where used.
- The ANN-model did not have any hidden layers
- A total of 3750 different ANN instances was created to the samples

Training network

- A 3-fold cross validation method was used to train and observe signs of over training
- The network was trained for 100 iterations
- The squared error was used as energy function to train the network

Creation of ANN model

Gene samples

Gene data samples

- A total of 88 samples was used
- 63 samples was used to train and validate the network and 25 for testing
- All four types of cancer was included in the training/validation set.
- In the test set there was also samples from healthy patients
- Each sample contained 6567 genes

Creation of ANN model

Filtering samples

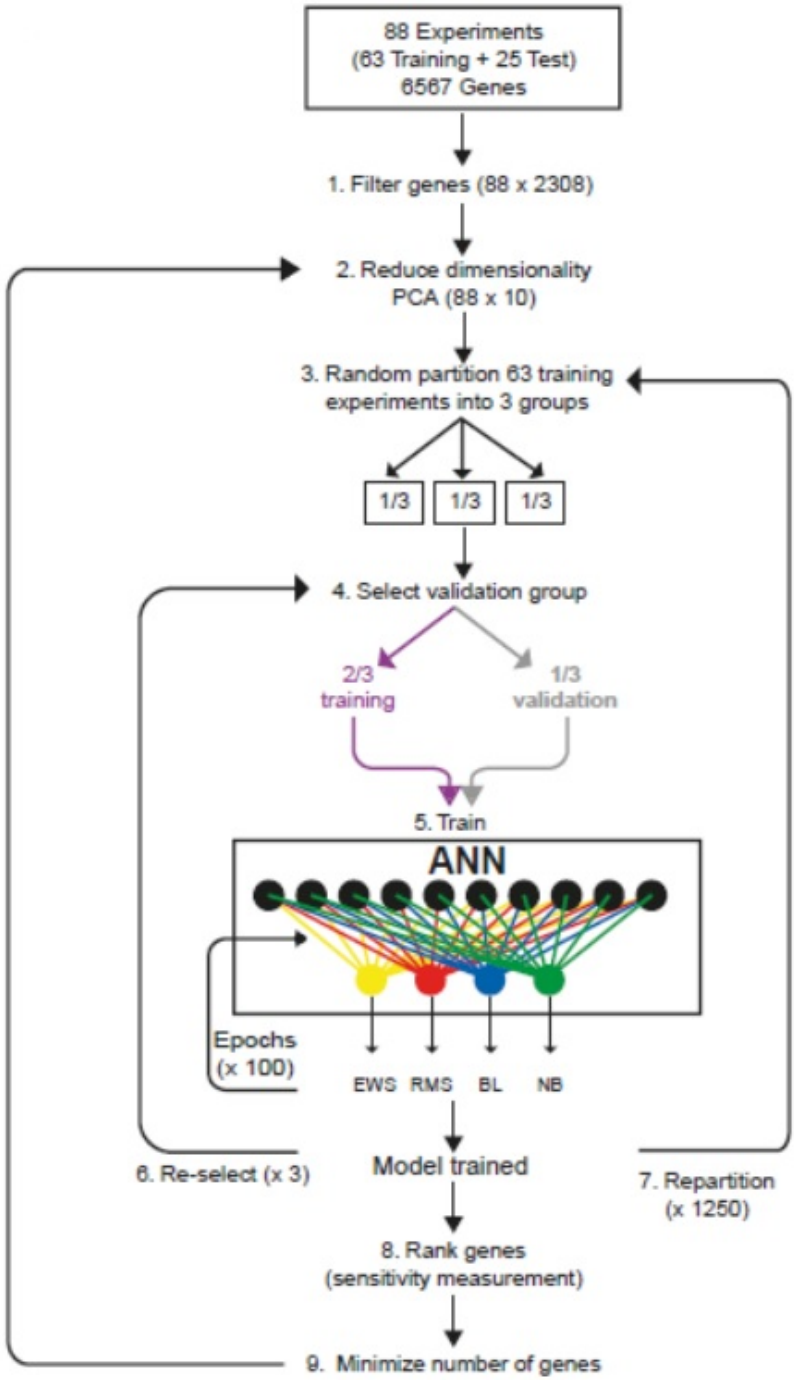
Filtering data

- two independent microarray experiments to test the reproducibility of the experiments and these were subsequently treated as separate samples.
- Filtering for a minimal level of expression reduced the number of genes to 2308
- Principal component analysis (PCA) further reduced the dimensionality, and we found that using the 10 dominant PCA components per sample as inputs and four outputs (EWS, RMS, NB or BL) produced well-calibrated ANN models.
- These 10 dominant components contained 63% of the variance in the data matrix.

Creation of ANN model

Training the network

1. Quality filtering, 6567 $\rightarrow$ 2308
2. Principal component analysis (PCA) 2308 $\rightarrow$ 10
3. Random partition into 3 groups
4. 3-fold cross validation (2/3 for training, 1/3 validation)
5. Training neural network for 100 iterations
6. Reselect training/validation samples (4.) 3 times
7. Repartition random samples (3.) 1250 times
8. Rank genes
9. Minimize number of genes (2.)



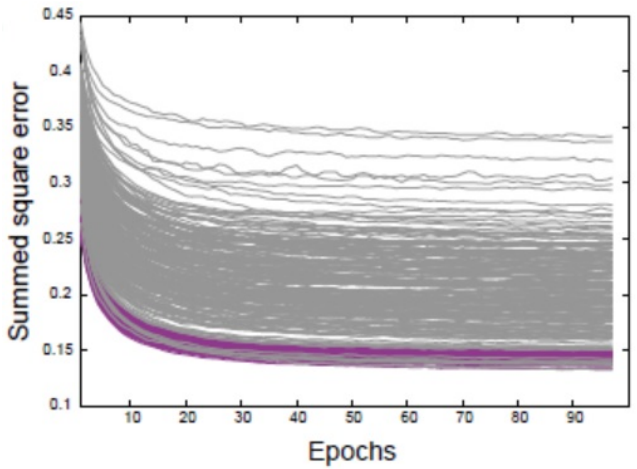
Outline

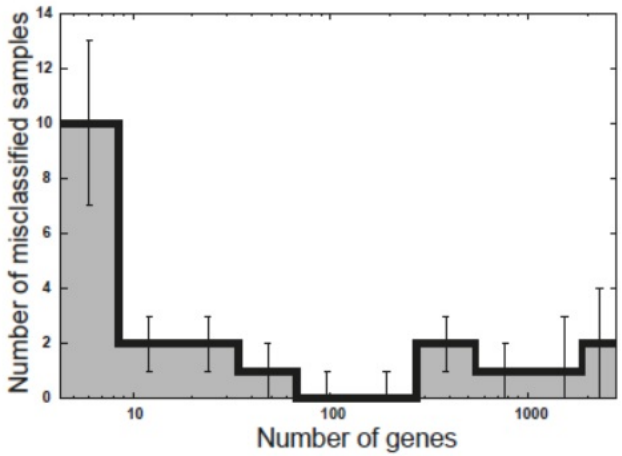
Calibrated ANN model

Validation of network

Result of training

- Created 3750 instances of neural networks
- No sign of over training
- Classification error rate minimized to 0% at 96 genes
- The 10 dominated PCA components for these 96 genes contained 79% of the variance
- Re-calibrated the neural networks using 96 genes and correctly classified all training samples





Classification and rejection

Classify cancer, reject if healthy

Classification

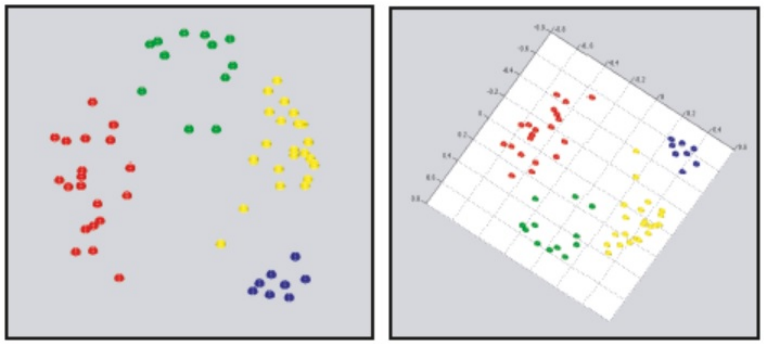
- A committee vote: Average of all predicted outputs
- Classified as cancer i if i gets the highest committee vote

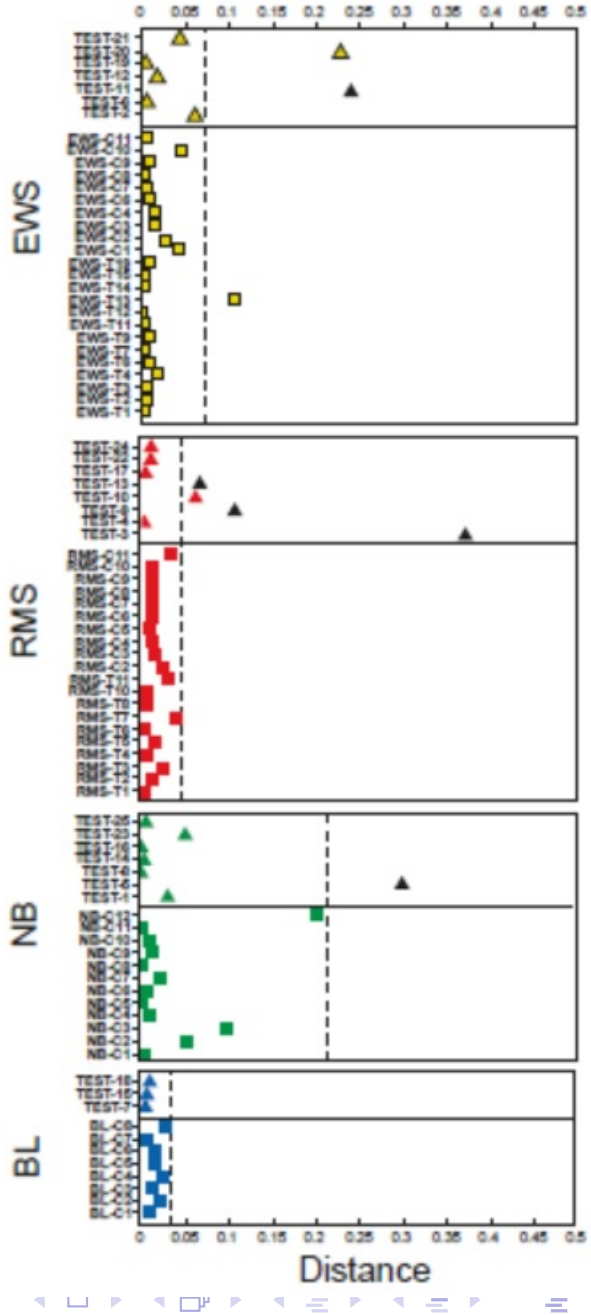
Rejection

- Since the neural network only allows 4 outputs in the form of a classification of a cancer, a rejection procedure is necessary to classify healthy samples.
- For this a squared euclidean distance was computed for each cancer type , between the committee vote for a sample and the "ideal" output for that cancer type.
- By using the ANN models for each validation sample, a probability distribution was created for each cancer type.
- Samples outside the 95th percentile was then rejected as cancer

Clustering

Method for rejection





- In the Figure above, a two dimensional projection of the clustering for each cancer type is shown.
- In the Figure to the right, the classification and their distance to the optimal output is shown.

Summary and Result

Training and validation

- A neural network without hidden layers was calibrated
- All 63 training samples was correctly classified and no sign of over training was shown
- Genes were ranked according to their significance for classification.
- This yielded 96 genes.
- All 3750 ANN instances were used to classify the additional test samples
- Using these 96 genes on the 25 test samples
  1. all 20 SRBCT:s was correctly classified (although, not always confidently enough)
  2. the 5 non-SRBCT:s was correctly rejected.

Discussion

Comments

- The ANN model that was used was a very simple model and could easily be extended to a more advanced one

Discussion

Comments

- The ANN model that was used was a very simple model and could easily be extended to a more advanced one
- Despite the simple model, the results was good, even though it is unclear how it would work with larger number of samples

Discussion

Comments

- The ANN model that was used was a very simple model and could easily be extended to a more advanced one
- Despite the simple model, the results was good, even though it is unclear how it would work with larger number of samples
- The article was published in 2001 and much could have happen in medicine since then

Discussion

Comments

- The ANN model that was used was a very simple model and could easily be extended to a more advanced one
- Despite the simple model, the results was good, even though it is unclear how it would work with larger number of samples
- The article was published in 2001 and much could have happen in medicine since then
- The article does show that these kind of methods might be a very good tool to use in diagnostic purposes