A study on feature selection based on AICc and its application to microarray data

Małgorzata Wesołowska

June 5, 2009

Abstract

The following paper contains a report of the final project written at the Faculty of Informatics (Facultat d'Informàtica de Barcelona) on Technical University of Catalonia (Universitat Politècnica de Catalunya) under the supervision of professor Luis Belanche. The project provides R language procedures, collected in a library, to select the best features, among the given feature set within the data, to describe a classifier. The principal criterion in such a selection process would be Akaike's information criterion - $\mathcal{AIC}$.

1 Objectives of the project

Feature selection is a key issue in machine learning and its importance is beyond doubt. This is especially so in learning tasks that are characterized by a very high dimensionality and a low number of learning examples. The purpose of this thesis is to investigate the potential of Akaike's information criterion $\mathcal{AIC}$ and it's derivatives such as $\mathcal{AIC}$c for feature selection in such learning tasks, exemplified by microarray data, as well as its cooperation with resampling techniques like the bootstrap.

The principal objective of the project is to join four issues:

1. Feature selection,
2. Akaike's information criterion,
3. Microarray Data,
4. Bootstrap resampling techniques.

Reducing the dimension of data brings profits from the various points of view. It's obvious, that the more parameters used, the better can be the fit of the model to the given data. It's worth to mention, that when there are many explanatory variables used to predict a classifier, it's possible to develop a model with many variables being significant (according to the high values of t-statistics), even if they are independent of the classifier. The problem of having many variables, known as as a "curse of dimensionality", is that addition of extra variables (dimensions) causes an exponential increase of data and operations necessary to execute the approximation. Mainly that's why it's important, to

keep the number of candidate features small while trying to achieve a large sample size.

To perform feature selection we can use **wrapper** or **filter** methods. In the wrapper methodology of feature selection we can use the classifier's accuracy as the performance measure. We build models with an aim to achieve the highest predictive accuracy possible and then select the best one according to its predictive accuracy. We take features used by this classifier as the **optimal features**. In a large data set feature selection process which put attention on the classifier's accuracy may be quite difficult or even impossible to execute. To decrease the size of the feature set from thousands to hundreds or tens, in a sensible period of time, we need simple methods selecting a subset of features with an aim to achieve the highest relevance to the classifier on the basis of the given data. These methods are known as filter methods.

According to the machine learning approach in classification tasks the model is being trained on the part of the data called **training set** and then should be validated on the remained data, called **validating set**. Training the model may result in overfitting the training data, which provide the model excellently fitted to the training set, but presenting poor predictability. Other inconvenience resulting in the division of the data into training and validating set is reducing the number of observation being taken to train the model. This can be really troublesome when deal with microarray data. That's why following $\mathcal{AIC}$ criterion may be profitable in such cases. Also the resampling techniques help with the problem of low number of observation, therefore they have been applied to the project.

Whole thesis consists of eleven chapters, which are divided into two basic parts: theoretical and practical.

1. **The theoretical part** (chapters 1-4) presents the whole formalism of AIC, feature selection, as well as characteristics of terms used in the further part of the work, it contains explanations and examples, which were designed by author.

2. **The practical part**, i.e. chapters 5-11, is the author's input in whole, it contains the proposed approach to feature selection on microarray data, description and computational complexity of algorithms used. Analysis of problems connected with different statistical methods and final conclusions are also included in this part. The last chapter contains the headers of all procedures written in R language with explanations.

The practical part consists of a *feature selection path*, with such steps:

1. **Filter selection** - at this early stage we pay attention on the relevance of each feature to the classifier. We compute some relevance measure individually for each existing feature and filter out the worse features in an accordance to the measure. This step assumes reduction of the dimension of the data from thousands to hundreds. The methodology used: ANOVA and Levene's test.

2. **Experimental selection** - since the dimension of the data is still high, we are not convinced to use the wrapper modeling. We will pay attention on correlation between features and so called *predictive accuracy improvement* which results from adding a second feature to a single feature model.

The experimental methods are based on a superstition, that if each two features from some feature set have good predictive accuracy, the whole feature set might have good predictive accuracy too. This step should reduce the dimension of the data from hundreds to tens. It consists of three algorithms:

(a) Removing high correlations with respect to $\mathcal{AIC}$,

(b) Filtering the most helpful features according to $\mathcal{AIC}$,

(c) Selecting proposed subsets of features.

Summarizing, after processing these algorithms we receive such subsets of features, in which

- each feature is helpful to one another in a sense of $\mathcal{AIC}$ *improvement*,
- none of the features share some α or higher percentage of information with one another. The α level have to be taken arbitrarily.

3. **Wrapper selection** in cooperation with bootstrap resampling technique - we apply a wrapper and develop candidate models each basing on a different subset of features achieved in the previous step. Since the proportion of the number of observations to the number of features is low we apply $\mathcal{AIC}$c criterion to each learning task. We explore two experimental paths:

   (a) external bootstrap - when a model is developed on a whole data set and bootstrap is applied to calculate the $\mathcal{AIC}$c index of the model

   (b) internal bootstrap - when a model is developed inside each bootstrap sample and the best model shows up as a consequence of bagging.

2 Objectives already achieved

The process of developing the project lasted according to such schedule:

date	step
February 2009	formulating the objectives, literature research, the project schema
March 2009	literature research, getting started with implementation of filtering methods and reading microarray data sets
April 2009	exploring the data, finding solutions for experimental selection, implementing the wrapper methods, testing the algorithms
May 2009	implementing the experimental selection algorithms, writing a document, optimizing and improving all algorithms, processing whole feature selection path on the real data

Previously there was no intention to develop the experimental selection part, instead of which some famous filter methods had been considered. After several tests and experiments author decided to introduce the experimental approach, based on the $\mathcal{AIC}$ *improvement*, to feature selection before the real wrapper modeling.

3 Planning of the work to be done to complete the project

Actually there is a few things left, which has to be done in June.

date	step
June 2009	applying whole selection path to more data examples, the discussion and the conclusions, finishing the document, writing a help() procedure, providing all algorithms with exception handle blocks

The project will be finished by the second half of June 2009. The results of each selection process applied to the real microarray data sets will be discussed and collectively compared.