Proposal for a diploma thesis

Automatic model selection in classification using the Minimum Description Length Principle

Daniel Renz
renz@informatik.hu-berlin.de

Supervisor:
Dr. Florin Popescu
popescu@first.fraunhofer.de
Prof. Ulf Leser
leser@informatik.hu-berlin.de

21. May 2007

Motivation

In classification tasks, one tries to assign the best-fitting of several models to each data class. The goal in model selection is thus to choose a model (out of several candidate models) that best represents data. The best-fitting model has to be estimated based on the number of observations (evidence) at hand. This is the classic problem of model selection, known as the bias-variance dilemma. It is most often formulated as finding the trade-off between goodness-of-fit (of a model with respect to observations of one class) and complexity (of a model). This is important, because often a complex model can describe the evidence much better than a simple model, but at the same time, it will generalize to new observations much worse. In other words, the complex model *overfits*.

Many different model selection criteria are available. Some examples are cross-validation, A Information Criterion, Bayes Information Criterion and Minimum Description Length (MDL), the latter three being modifications of Maximum-Likelihood-Estimation (see [1], [2] and [3]). All of these criteria find a different trade-off between goodness-of-fit and complexity and it is not known, which (if any) of the model selection criteria is generally the best. In this work, only MDL will be considered - to the author it seems to be a promising model selection criterion, because it allows one to sparsify features and choose confidence intervals for each of the parameters to be compressed with highest chance of generalization. Also, MDL does not impose too many constraints on the properties of the data to be compressed/classified/predicted.

The first central idea of MDL is that any regularity in the data can be used to compress the data. The more regularities there are, the more the data can be compressed. The second central idea is that 'learning' can be equated with 'finding regularities': The more we can compress, the more we learn. The simplest form of MDL seeks to minimize over all possible candidate models the sum of the lengths of the

encoding of the data given a model (using some specified code) and the encoding of the model (specified by its parameters). A good introduction to MDL can be found in [4].

Goal

While the theoretical foundations of MDL are well developed, applications of this principle have mainly been restricted to regression problems. The goal of this thesis is to show that MDL can successfully be applied to classification problems as well. A comparison of MDL performance and the performance of established methods, such as Support Vector Machines, will be conducted. Several data sets from the UCI Machine Learning Repository will be used for performance testing (http://www.ics.uci.edu/~mlearn/MLRepository.html).

Approach

In the classification framework, the goal will be compression of the data given class information (i.e. which sample belongs to which class), using generalized linear models, i.e. models which are linear in parameters $\boldsymbol{\theta}$ to be estimated, but non-linear in input variables x (features). The specific form of generalized linear models to be used in this thesis can be defined as

$$\boldsymbol{y} = M_{prod} \times \boldsymbol{\theta}.$$

Here, $\boldsymbol{y}$ is some variable that will be used for classification. M_{prod} is defined as:

$$M_{prod} = \left[\vec{1}\, M \;\; P\right].$$

$\vec{1}$ is a vector containing 1's (the bias), M is the data matrix of size [nr. of features x nr. of samples], and is a matrix whose entries correspond to some non-linear mapping $\varphi(i,j)$ of any two features i, j. P is sometimes referred to as a product feature matrix.

In this framework there are many more parameters than features. Nonetheless, sparsifying the parameter set – which has been successfully done with model selection criteria in auto-regressive models – is equivalent to feature sparsification, i.e., feature selection. This is due to the fact that each parameter corresponds to a function of only zero, one or two input variables (in M_{prod}): parameters set to zero determine which input variables are not used.

In the proposed work, the basic general linear regression can be under-determined, and therefore a search is done over the linear subspace of exact solutions to the regression problems, approximating a set of parameters inside this space. The challenge lies in improving the speed of optimizations of a non-convex function in thousands of variables, so that near-optimal compression rates can be achieved in reasonable time, possibly using a greedy strategy. This is why realistically sized data sets will be analyzed.

In detail, the work will contain the following steps:

1) Improve existing integer & rational codes that can encode the data.
2) Test smooth functions that approximate the length of codes for use in MDL optimization.
3) Apply a classification using product space regression, regularized by MDL: Develop a greedy search strategy for the best lossless data compression in a high-dimensional space to avoid ´the curse of dimensionality´.
4) Attempt feature extraction by iterative generation and elimination.
   Compare performance of properly regularized product space regression to current state of the art techniques, such as SVMs (on several data sets chosen from the UCI Machine Learning Repository).
5) Regression might poorly describe relations between variables that cannot be approximated by smooth functions (e.g. not one-to-one, disconnected sets) and thus it will be interesting to see for what kinds of data kernel based methods – or kernel based regressions - will be better suited than global function approximation methods such as regression.

Bibliography

1. *Model Selection and the Principle of Minimum Description Length.* **Hansen, M.H. and Yu, B.** 454, 2001, Journal of the American Statistical Association, Vol. 96.

2. *Key Concepts in Model Selection: Performance and Generalizability.* **Forster, M.R.** 1995, Journal of Mathematical Psychology, Vol. 38, pp. 3-20.

3. *Schwarz, Wallace and Rissanen: Intertwining Themes in Theories of Model Selection.* **Lanterman, A.D.** 2, 2001, International Statistical Review, Vol. 69, pp. 185-212.

4. *A Tutorial introduction to the minimum description length principle. In: Advances in Minimum Description Length:Theory and Applications.* **Grünwald, P.** s.l. : MIT Press, 2004.