Testing a Simulated Annealing Algorithm in a Classification Problem

Karsten Luebke and Claus Weihs

University of Dortmund, Department of Statistics,
44221 Dortmund, Germany
luebke@statistik.uni-dortmund.de
http://www.statistik.uni-dortmund.de

Abstract. In this work we develop a new classification algorithm based on simulated annealing. The new method is evaluated and tested in a variety of situations which are generated and simulated by a Design of Experiments. This way, it is possible to find data characteristics that influence the relative classification performance of different classification methods. It turns out that the new method improves the classification performance of the classical Linear Discriminant Analysis (LDA) significantly in some situations. Moreover, in a real life example the new algorithm appears to be better than LDA.

Keywords: Simulated annealing, classification, desgin of experiments, latent factors.

1 Introduction

Classification (or supervised learning) is a ubiquitous challenge which is tackled by many different methods like for example Linear Discriminant Analysis (LDA), Support Vector Machines (SVM) or decision tree classifier. Surprisingly, the simple classification method of LDA does perform well even in situations, where the underlying premises are not met. Our new method is closely related to LDA (so that we keep the good classification performance) but does not make use of the premises. It projects the original observation first on some latent factors and then transforms these latent factors in order to discriminize the classes. We call the new method Classification Pursuit Projection (ClPP). It is a computer intensive method as it is using a stochastic algorithm, namely Simulated Annealing as a function minimizing algorithm.

In order to compare the performance of the new method to LDA an experimental design is used. To achieve most general results 7 characteristics of a classification problem are varied in the design.

The paper is organized as follows: In the next section we introduce the underlying scoring function of the classification problem which is minimized by simulated annealing. In section 3 the implementation of the Simulated Annealing Algorithm is described. The characteristics of the classification problem which are varied are introduced in section 4. The results of the simulations are reported in section 5. The new method was also tested on a real world example in section 6. After that a brief outlook on future work will be given as well as some concluding remarks.

A. Albrecht and K. Steinhöfel (Eds.): SAGA 2003, LNCS 2827, pp. 61–70, 2003.

2 Optimal Scoring with Latent Factors

Linear Discriminant Analysis is a statistical method for classification. In LDA the classification is based on the calculation of the posteriori probabilities of a trial point. The class with the highest posteriori probability is chosen. To calculate the posteriori probability it is assumed that the data comes from a multivariate normal distribution where the classes share a common covariance matrix but have different mean vectors. Hastie et.al. show in [4] that LDA is equivalent to canonical correlation analysis and optimal scoring. This is possible as LDA can be seen as a special linear regression. One of the problems in LDA is that the estimated covariance matrix of the data points has to be inverted for the classification. Especially in a high dimensional problem this can cause numerical problems. In the paper of Hastie et.al. they try to overcome possible numerical problems by using a penalysing term. This may not be optimal as the covariance matrix is transformed away from singularity without using information in the data. In the new method the data is first projected on (few) latent factors so that the original covariance matrix need not to be inverted and so there are no problems with singular matrices.

Assume that there are n observations with p variables in the predictor space and k classes. Let

- $X \in R^{n\times p}$: Predictor variables.
- $Y \in R^{n\times k}$: Indicator matrix of the classes.

The basic idea is as follows: Assign $l \leq k-1$ scores to the classes and regress these scores on X. We are looking for scores (of the k classes) and a suitable regression of these scores on the predictor variables so that the residuals are small for the true class and large for the wrong. The average squared residual function is (see Hastie et.al. ([5], p. 392):

$$ASR(H,M) = \frac{1}{n}||YH - XM||^2, \tag{1}$$

where

- $H \in R^{k\times l}$ is the score matrix of the classes, and
- $M \in R^{p\times l}$ is the regression parameter matrix.

To avoid trivial solutions the constraint

$$H'(Y'Y/n)H = I_l \tag{2}$$

is used.
To tackle the numerical problems in calculating M so called latent factors are derived. Latent factors are linear combinations of the original predictor variables $Z = XG$ with $G \in R^{p\times r}, r < p$.

These latent factors must fulfill the side condition that they are orthonormal, i.e.

$$Z'Z = (XG)'(XG) = I_r. \tag{3}$$