

LINNÉ, a Software System for Automatic Classification

N. Christlieb¹ and L. Wisotzki

*Hamburger Sternwarte, Gojenbergsweg 112, D-21029 Hamburg,
Germany*

G. Graßhoff

*Institut für Wissenschaftsgeschichte, Georg-August-Universität
Göttingen*

A. Nelke and A. Schlemminger

Philosophisches Seminar, Universität Hamburg

Abstract. We report on the software system LINNÉ, which has been designed for the development and evaluation of classification models. LINNÉ is used for the exploitation of the Hamburg/ESO survey (HES), an objective prism survey covering the entire southern extragalactic sky.

1. Introduction

The Hamburg/ESO survey (HES) was originally conceived as a wide angle objective prism survey for bright quasars. It is carried out with the ESO Schmidt telescope and its 4° prism and covers the total southern extragalactic sky. For a description of the survey see Wisotzki et al. (1996).

A few years ago we started to develop methods for the systematic exploitation of the *stellar* content of the survey by means of automatic spectral classification. A short overview of the scientific objectives is given in Christlieb et al. (1997) and Christlieb et al. (1998), where also a detailed description of the classification techniques can be found. In this paper we report on the software system LINNÉ, that has been designed for the development and evaluation of classification models.

2. Classification models

A *classification model* (CM) consists of the following components:

Class definitions are given by means of a learning sample, implicitly including the number N and names of the defined classes.

¹E-mail: nchristlieb@hs.uni-hamburg.de

Class parameters include the *a priori* probabilities $p(\omega_i)$, $i = 1 \dots N$, of the N defined classes and the parameters of the multivariate normal distribution of the class-conditional probabilities $p(\vec{x}|\omega_i)$.

Classification aim can be one of the following items:

- (1) Perform a “simple”, i. e. Bayes-rule classification.
- (2) Compile a complete sample of class $\omega_{\text{target}} \in \{\omega_1, \dots, \omega_N\}$ with minimum cost rule classification.
- (3) Detect “un-classifiable” spectra, i. e. spectra to which the reject option (Christlieb et al. 1998) applies. Note that e. g. quasar spectra belong to this class.

Feature space The space of features in which the search for the optimal subset is carried out. Note that in certain cases one may want to exclude available features beforehand to avoid biases, so that the feature space is not necessarily identical to the *total* set of available features.

Optimal feature set for the given classification aim.

Optimal loss factors In case of classification aim (3) a set of three optimal loss factors – weights for different kinds of misclassifications – has to be stated (Christlieb et al. 1998).

Once a CM is established, it is straightforward to derive from it a *classification rule* for the assignment of objects of unknown classes to one of the defined classes.

The aim of LINNÉ is to permit easy and well controlled access to the variation of the model components and effective means to evaluate the resulting quality of classification. The performance of a model with classification aim (1) can be evaluated by e. g. the total number of misclassifications, estimated with the *leaving-one-out* method (Hand 1981); in case of aim (3) the model is usually assessed by the number of misclassifications between the target class and the other classes.

3. Description of the system

The core of LINNÉ was implemented in an object-oriented extension to Prolog, with the numerical routines – e. g. for estimation of the parameters of the multivariate normal distributions – written in C. To facilitate user interaction and to ensure effective control over the model components and performance, a graphical user interface (GUI) for LINNÉ was developed (see Figure 1). After the first implementation, using Prolog’s own graphical library (SWI-Prolog plus XPCE), we recently started to switch to Java for reasons of system independence and remote access via WWW. At present, LINNÉ has a server-client architecture, the Prolog server communicating with a Java client through TCP/IP sockets. The server keeps the learning sample data, read in from MIDAS via an interface and converted into Prolog readable terms. It is not yet possible to select *all* model components interactively via the GUI, so that partly pre-designed models have to be used. They are also provided from the server side.

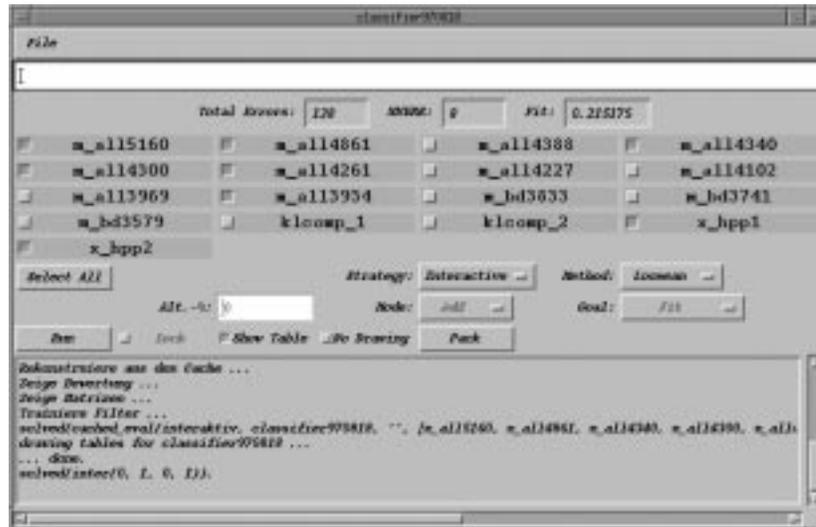


Figure 1. Main control panel of LINNÉ. The three upper text fields show model performance parameters. Below them the feature selection area is placed (m_all15160... x_hpp2). The automatic feature selection is controlled by the menus above the Prolog server messages window.

The results of the classification model evaluation are presented on the client. A confusion matrix and loss matrix window assists the user in the analysis of the model. The user may then alter components and repeat the evaluation to improve the model step by step.

The search for the optimal feature set can also be done automatically. Since the set of available features may easily become too large to perform exhaustive search among all possible combinations, apart from the exhaustive search a hill-climbing like, stepwise search has been implemented. It can be controlled from the client side, using different strategies and branching parameters.

LINNÉ also provides a tool for the systematic and efficient adjustment of loss factors (see Figure 2).

4. Application of classification models

Once a CM has been established, evaluated, and the evaluation has pleased the user, its parameters can be exported to MIDAS tables. The classification of spectra of unknown classes can then be carried out under MIDAS. The typical computing time for the classification of all spectra on one HES plate – mapping $5^\circ \times 5^\circ$ of the sky and yielding typically $\sim 10,000$ non-disturbed spectra with $S/N > 10$ – is less than 5 min on a Linux PC with a Pentium 133 MHz processor.

So far LINNÉ has been used for some first test applications, i. e. compilation of a sample of extremely metal poor halo stars and a search for FHB/A stars. It

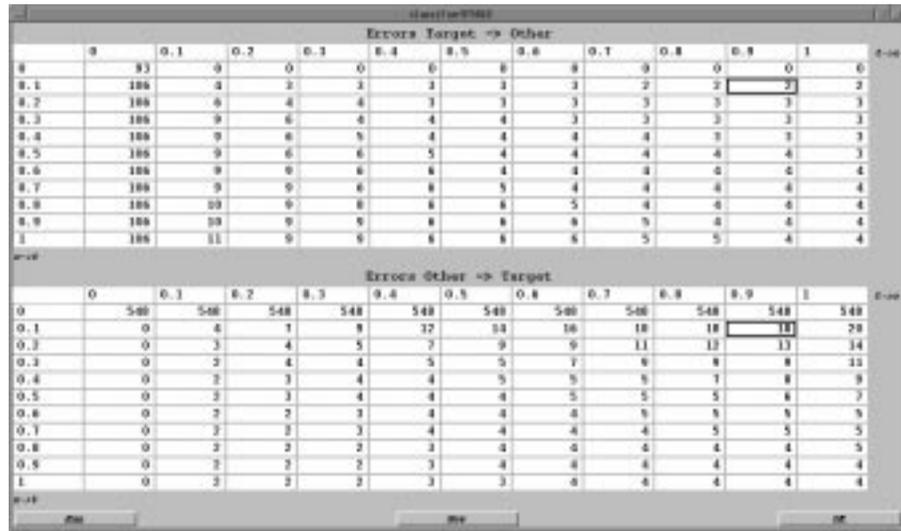


Figure 2. Tool for interactive adjustment of loss factors. The upper half of the window shows the number of target class spectra which have been erroneously assigned to one of the other classes in dependence of the loss factors $c_{\text{target} \rightarrow \omega_i}$ (abscissa) and $c_{\omega_i \rightarrow \text{target}}$ (ordinate). The lower half shows the same for the target class contamination. The third loss factor, $c_{\omega_i \rightarrow \omega_j}$, does not have to be adjusted but can be held constant at a small value.

will be developed further and extended in functionality and will be applied to the exploitation of the huge HES data base, which will finally consist of $\sim 5,000,000$ digitised objective prism spectra.

Acknowledgments. N.C. acknowledges an accommodation grant by the conference organizers. This work was supported by the Deutsche Forschungsgemeinschaft under grants Re 353/40-1 and Gr 968/3-1.

References

- Christlieb, N. et al. 1997, in *Wide-Field Spectroscopy*, ed. Kontizas, E. et al., Kluwer, Dordrecht, 109
- Christlieb, N. et al. 1998, to appear in *Data Highways and Information Flooding, a Challenge for Classification and Data Analysis*, ed. Balderjahn, I. et al., Springer, Berlin.
- Hand, D. 1981, *Discrimination and Classification*, Wiley & Sons, New York.
- Wisotzki, L. et al. 1996, *A&AS*, 115, 227