ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

# Relevance learning for mental disease classification

Barbara Hammer[1], Andreas Rechtien[2], Marc Strickert[3], and Thomas Villmann[4]

(1) Clausthal University of Technology, Institute of Computer Science, Clausthal-Zellerfeld, Germany, hammer@in.tu-clausthal.de
(2) University of Osnabrück, Department of Mathematics/Computer Science, Osnabrück, Germany, arechtie@uni-osnabrueck.de
(3) IPK Gatersleben, Pattern Recognition Group, Gatersleben, Germany, stricker@ipk-gatersleben.de
(4) University of Leipzig, Clinic for Psychotherapy and Psychosomatic Medicine, Leipzig, Germany, villmann@informatik.uni-leipzig.de

**Abstract**. In medical classification tasks, it is important to gain information about how decisions are made to ground and reflect therapies based on this knowledge. Neural black box mechanisms are not suitable for such tasks, whereas symbolic methods which extract explicit rules are, though their tolerance with respect to noise is often smaller since they do not rely on distributed representations. In this article, we test three hybrid prototype-based neural models which combine neural representations with explicit information representation in comparison to classical decision trees for mental disease classification. Depending on the model, information about relevant input attributes and explicit rules can be derived.

## 1 Introduction

An important problem in psychotherapy is to identify border-line patients which are heavily disturbed since a specific treatment and therapy management is necessary for these patients [2]. This classification has to take place before the treatment. Commonly, therapists make their decision depending on the clinical impression of the patients within an initial session. As an alternative which does not require human intervention, the self-report questionnaire *Borderline Personality Inventory* (BPI) has been developed which allows automatic classification of border-line patients [3]. The questionnaire is answered by the patients and a resulting cut-off-value indicates a borderline typed disturbance. The classification of human experts and the results of the questionnaire, however, differ significantly in the given database, indicating that decisions are based on different factors or a different weighting scheme of the factors. Now the question occurs which factors are considered relevant for the respective decision such that an explicit model can be built which can be the starting point for a reliable, optimized, and automatic classification.

The results of three different additional questionnaires are available for each patient which describe potential factors for psychological disturbances, i.e., which reflects different topics of the overall personality and personality relations of the patients. The first one is the *Giessen-Test-Questionnaire* (GT) which is for self-assessment of patients with respect to their emotional feeling, basic self-understanding of personality and social competence, whereby the aspects of self-

ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

perception and self-ideal discrepancies can be detected [1]. The second questionnaire is the *Freiburg Personality Inventory* (FPI) reflecting a set of personality properties including social orientation, inhibition, contentment, aggressiveness, stress, somatic complaints, openess, sensitivity and emotionality/neuroticism [9]. The set of questionnaires is completed by the *Invetory for Assesment of Interpersonal Relations (IIP)* which judges the interpersonal relations.

Based on these data, the question for machine learning is to identify the relevant attributes and a possible hypothesis how experts and BPI, respectively, make their decision. We tackle this problem learning classifiers from the given data, which allow insight into their classification and which provide a hypothesis for the respective decision rule.

## 2 GRLVQ, decision trees, and rule extraction

*Generalized relevance learning vector quantization* (GRLVQ) has been developed by the authors [6] as an improvement of LVQ as introduced by Kohonen [8]. On the one hand, GRLVQ obeys a global cost function also for a continuous data distribution which assures convergence of the algorithm and, in addition, an optimization of the margin during training, i.e. good generalization ability [5]. On the other hand, it substitutes the euclidean metric by a weighted metric with relevance terms which are automatically adapted during training. Apart from an improved accuracy and flexibility, this weighting scheme allows insight into the behavior of the classifier since the learned relevance profile of the input dimensions directly indicates based on which features decisions are taken.

Formally, a GRLVQ network is characterized by a number of prototypes $\vec{w}^i \in \mathbb{R}^n$, $i = 1, \ldots, c$ which are masked with labels $y(\vec{w}^i) \in \{1, \ldots, C\}$. A point $\vec{x} \in \mathbb{R}^n$ is mapped to $y(\vec{w}^i)$ where the distance $|\vec{x} - \vec{w}^i|^2_\lambda$ is minimal.

$$|\vec{x}^i - \vec{w}^j|^2_\lambda = \sum_k \lambda_k^2 (x_k^i - w_k^j)^2$$

is the weighted euclidean metric with relevance terms $\lambda_i \geq 0$ with $\sum_i \lambda_i^2 = 1$. The learning algorithm adapts both, prototypes $\vec{w}^i$ and relevance terms $\lambda_i$ by minimizing the cost function

$$E_G = \sum_{\vec{x}^i} \mathrm{sgd}\left(\left(d^+(\vec{x}^i) - d^-(\vec{x}^i)\right) / \left(d^+(\vec{x}^i) + d^-(\vec{x}^i)\right)\right)$$

for a given training set $(\vec{x}^i, y^i) \in \mathbb{R}^n \times \{1, \ldots, C\}$, $i = 1, \ldots, p$, by a stochastic gradient descent, where $d^+(\vec{x}^i)$ is the squared weighted euclidean distance to the closest correct prototype and $d^-(\vec{x}^i)$ is the squared weighted euclidean distance to the closest wrong prototype. It has been shown in [5] that the generalization ability of GRLVQ-networks can be expressed in terms of the quantity $\min_i |d^+(\vec{x}^i) - d^-(\vec{x}^i)|$, the hypothesis margin.

*Induction of decision trees* constitutes a standard symbolic machine learning tool for classification of data [10]. A decision tree consists of a tree which interior nodes are labeled by a dimension number and the connections to the node's children are labeled by real values which split the dimension into intervals. The leaves contain class labels. Given a data point $\vec{x}$, a decision is based on consecutive decisions provided by the interior nodes until the class information is reached at the leaves. For a given training set, a decision tree can recursively be induced by the choice of an interior node and an induced split of the training set until a widely uniform classification is possible at the leaves. Thereby, an appropriate measure such as the entropy guides the choice of the splitting dimension. In this paper, we use the algorithm J48 provided within the open source Weka data mining software [11]. Unlike GRLVQ, decision tree learners consider one attribute at a time, such that relevance distributed among several attributes cannot be detected. However, they provide explicit rules and an ordering of the dimensions with respect to their importance for the decision tree by means of their depth within the tree.

GRLVQ comes together with a relevance profile, i.e. a weighting scheme for the dimensions such that the extraction of a decision tree which approximates a trained GRLVQ network is easily possible with only little extra cost, without the need of further training or induction. The following recursive method, the *BB-tree algorithm*, has been introduced and tested by the authors in [4]: we start with the dimension with highest relevance factor as the root of the decision tree and split between consecutive prototypes (whereby we ignore prototypes which distance in the considered dimension is smaller than average) until the remaining set of data points assigned to this strand can almost uniquely be classified by one class or further separation has no effect. Depending on the respective setting, the complexity of the resulting tree can be controlled and extraction of further interior nodes be stopped, assigning the majority class to the resulting leaf. In this BB-tree algorithm, the order of the dimensions within the tree is given by the order of the relevance factors determined by GRLVQ. Thus the BB-tree algorithm is built on the ability of GRLVQ to judge groups of dimensions at a time, i.e. relevant input groups will be highlighted and the involved dimensions be used for splitting. Induction of decision trees, on the contrary, does only consider one dimension at a time independent of their connection to other splitting possibilities, since the information gain of each dimension is computed separately. However, since the order of the dimensions within BB-trees is priorly fixed, a different weighting scheme for subsets of points cannot be achieved. In addition, the number of prototypes and thus the potential complexity of the resulting tree is determined a priori.

In this article, we introduce a further fully *hybrid method* based on the BB-tree algorithm. The method iterates GRLVQ training for a minimum number of prototypes (only one per class, thus a very fast version of GRLVQ), followed by the induction of a new node and the corresponding splitting of the data set as in the BB-tree algorithm. Thus, GRLVQ is started with one prototype per class for a fixed number of steps. The most relevant dimension serves as splitting dimension, whereby the splits are located between the prototypes ignoring prototypes which are closer than average in the respective dimension. For the resulting data split, GRLVQ is restarted from initial positions independently for each part of the data set until a decision is reached. Since the method iteratively splits and trains afresh, it is more costly than a direct tree extraction from a fully trained GRLVQ network. However, this method allows more flexible trees since local weighting factors evolve for each data split, thus the split-dimensions

| | BPI-CUT (test) | BPI-CUT (full) | DIAG (test) | DIAG (full) |
|---|---|---|---|---|
| GRLVQ(1) | 72 | 76 | 71 | 72 |
| GRLVQ(2) | 71 | 77 | 70 | 73 |
| BB(1) | 67 | 74 | 68 | 69 |
| BB(2) | 64 | 74 | 67 | 68 |
| HYBRID | 64 | 77 | 67 | 71 |
| J48 | 67 | 97 | 60 | 98 |

Table 1: Mean training accuracy (in %) for different algorithms and settings, The mean error on the test set and the training error on the whole data set are reported

are independent for each interior node, whereas they are identical for all interior nodes of the same depth within the BB-tree algorithm. In addition, the initial complexity of GRLVQ need not fit the complexity of the final classifier for this hybrid version, since the initial setting just serves as a starting point for an automatic modularization of the problem. The overall complexity of the classifier evolves automatically and can be controlled during training.

## 3 Experiments

Data from 334 patients are available who are characterized by 28 attributes describing their personality and personal relations collected within self-assessment using the three different questionnaires GT, FPI and IIP. As mentioned above, the questionnaires focus on different aspects of personality and personal relations, whereby attributes 0 to 7 stem from the GT, attributes 8 to 19 come from the FPI, and the attricutes 20 to 27 come from IIP. These inputs are classified according to the BPI classification (data set BPI-CUT) which indicates 38% as border-line patients, and they are classified according to diagnosis made by therapist (data set DIAG) who set 32% as border-line patients.

These two data sets are tackled as classification tasks for machine learning whereby specific emphasize is laid on the respective attributes indicated as relevant. Data has been normalized before training by a z-transformation. Training is done on five different random splits of the data sets into training and test set, each comprising 50%, and, in addition, once on the data set as a whole (without test set). The training algorithms include:

- J48: decision trees (J48 with default parameters as set in WEKA), the resulting depth of the trees varies about 10, resulting into 70-100 nodes,
- GRLVQ(1), GRLVQ(2): GRLVQ with one and with two prototypes per class, respectively, whereby the parameters have been roughly optimized over all runs (initialization in the centre points of the classes, learning rate 0.05, learning rate for the relevance terms 0.005, about 1000 updates),
- BB(1), BB(2): decision trees extracted from GRLVQ(1) and GRLVQ(2) with maximum depth given by the number of significant relevance terms

| | BPI-CUT (test) | BPI-CUT (full) | DIAG (test) | DIAG (full) |
|---|---|---|---|---|
| GRLVQ(1) | 19>15>11 | 17>19>15>11 | 1 | 2>0 |
| GRLVQ(2) | 19>17>15>11 | 17>15>11 | 23>1>0>18 | 2>7>0 |
| HYBRID | 19 – (5,17) | 17 – (7,11,15,19) | (0,1,2) – (1,2,10,14) | 2 – (16,17) |
| J48 | 15 – (12,17) – (1,2,8) | | 15 – (2,12) – (18,24) | |

Table 2: Dimensions ranked as important by the algorithms. For GRLVQ, dimensions with average weighting $> 0.1$ are reported in the order of their relevance; thereby, some dimensions change for the runs because the information can be substituted by others. For BB, the dimensions ranked as important are the same, since training is identical. For HYBRID, the dimensions in the first two layers of the extracted decision trees are reported; these are different for HYBRID in different runs and all found dimensions are accumulated in the presentation. For J48, the dimensions indicated as important are the same for subsets and the full set and they are reported for the first three layers.

(usually 3),

- HYBRID: a hybrid GRLVQ-BB algorithm as described earlier with roughly optimized parameters (learning rate 0.05 for prototypes and relevance terms, 400 steps for each training, resetting of training after each split, maximum depth 3).

The achieved classification results are reported in table 1. It is obvious, that the classification task is difficult, since it is affected by severe noise. From visual inspection of the data, it can be seen that the correlation of a single dimension of the data with the classification is low. Correspondingly, the neural method GRLVQ which takes more than one dimension into account at a time shows better performance than decision trees (4% resp. 10% improvement on the test set compared to J48). Turning from GRLVG to decision trees gives competitive results in case of BPI-CUT and superior results in case of DIAG with respect to J48. J48 shows superior results on the full training set due to the fact that the complexity of the trees was not restricted for J48, whereas it is for the other methods. Increasing the latter complexity would also allow a classification close to 100% on the training set, but without generalization to the test set.

The algorithms emphasize dimensions by weighting (GRLVQ, BB) or split according to the dimensions at the top of the decision tree (J48, HYBRID). The respective dimensions are reported in table 2. Naturally, the ranking of the algorithms is different since their result is nondeterministic and information is spread among the attributes. Nevertheless clear tendencies can be observed for all runs: the decision for BPI-CUT is mainly influenced by dimensions 11, 15, 17, and 19, all quantities are contained in the second questionnaire FPI. On the contrary, the decisions of therapists collected in DIAG also depend on dimensions 0, 1, and 2, all attributes of the first questionnaire GT. The third questionnaire starting from dimension 20 only plays minor a role.

## 4 Discussion

Obiously, the decision for borderline typed disturbance is made on different diognostic views. The BPI-based decision indicates scales obtained from FPI which are in particular the items inhibition (11), somatic complaints (15), openess (17) and emotionality/neuroticism (19). In contrast, the decisions made by the therapist reflect the items provided by the GT outlining the items social resonance (0), dominance/pliancy (1) and control as opposite to compulsion (2). So there is some evidence that the clinical view by therapist is dominated by *active* aspects of personality whereas the BPI-cut-decision emphasizes more *passive* aspects representing the emotial and somatic feeling of patients.

From a technical point of view, GRLVQ, the hybrid variant GRLVQ-BB and J48-decision trees obtain competitive results as discussed above. Yet, one advantage of the GRLVQ-based approaches could be seen in the availability of the more flexible description by characteristic prototypes.

## References

[1] D. Beckmann, E. Brähler, and H.-E. Richter. Der Gießen-Test. Huber-Verlag, Bern, 1991.

[2] F. Leichsenring. BPI - das Borderline-Persönlichkeits-Inventar. in E. Brähler, J. Schuhmacher and B. Strauß (eds.) Diagnostische Verfahren der Psychotherapie, Bd. I. Hogrefe-Verlag Göttingen, Germany. 60–63, 2002.

[3] F. Leichsenring. Development and first results of the Bordeline Personality Inventory (BPI) - a self-report instrumnet for assessing borderline personality and borderline schizophrenia. *Journal of Personality Assessment* 73:45–63, 1999.

[4] B. Hammer, A. Rechtien, M. Strickert, and T. Villmann. Rule extraction from self-organizing maps. In: J.R. Dorronsoro (ed.), Artificial Neural Networks - ICANN 2002, pages 370–375, Springer, 2002.

[5] B. Hammer, M. Strickert, and T. Villmann. On the generalization ability of GRLVQ-networks. To appear in *Neural Processing Letters*.

[6] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks* 15:1059–1068, 2002.

[7] L.M. Horowitz, S.E. Rosenberg, B.A. Bauer, G. Ureno and V. Villasensor. Inventory of interperosnal problems - psychometric properties and clinical applications. *Journal of Clinical and Consulting Psychology* 56: 885–892, 1988.

[8] T. Kohonen, Self-Organizing Maps, Springer, 1997.

[9] F. Ostendorf. Freiburger Persönlichkeitsinventar - Revidierte Fassung (FPI-R). *Zeitschrift für differentielle und diagnostische Psychologie* 18:81–85, 1991.

[10] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

[11] I.H. Witten and E. Frank, Data Mining: Practical machine learning tools with Java implementations, Morgan Kauffmann, 2000.