The Most Generative Maximum Margin Bayesian Networks: Supplementary Material

Robert Peharz* ROBERT.PEHARZ@TUGRAZ.AT
Sebastian Tschiatschek* TSCHIATSCHEK@TUGRAZ.AT
Franz Pernkopf PERNKOPF@TUGRAZ.AT
Signal Processing and Speech Communication Laboratory, Graz University of Technology
*These authors contributed equally to this paper

1. Dataset Descriptions

- **UCI data (Frank & Asuncion, 2010).** In case of the datasets *chess, letter, mofn-3-7-10, satimage, segment, shuttle-small, waveform-21, abalone, adult, car, mushroom, nursery, and spambase*, a test set was used to estimate the accuracy of the classifiers. For all other datasets, classification accuracy was estimated by 5-fold cross-validation.

- **TIMIT data (Pernkopf et al., 2012).** This data set is extracted from the TIMIT speech corpus. Utterances from 16 male and 16 female speakers are frame-wise classified into either four or six phonetic classes with 110134 and 121629 samples separately. Each sample consist of 20 mel-frequency cepstral coefficients and wavelet-based features. Subsets of the data that consist of either male speakers (M) or female speakers (F) are considered.

- **USPS data (Hastie et al., 2003).** This data set contains 11000 handwritten digit images from zip codes of mail envelopes. The data set is split into 8000 images for training and 3000 for testing. Each digit is represented as a 16×16 grayscale image. Each pixel is considered as feature.

2. Implementation Details for Projected Gradient Method

In this section we provide more details on the implementation of our projected gradient method. For convenience we re-state the problem formulation for the ML-BN-SVM:

$$\begin{aligned}
\min_{\boldsymbol{\omega},\boldsymbol{\xi}} \quad & -\mathbf{n}^T\boldsymbol{\omega} + \lambda \sum_{m=1}^{M} \xi_m \\
\text{s.t.} \quad & \left(\phi_{c^m}(\mathbf{x}^m) - \phi_c(\mathbf{x}^m)\right)^T \boldsymbol{\omega} + \xi_m \geq \gamma \quad \forall m, c \neq c^m \\
& \log \sum_{j'} \exp\left(\omega^i_{j'|\mathbf{h}}\right) \leq 0 \qquad \begin{array}{l}\forall\, 0 \leq i \leq N \\ \forall\, \mathbf{h} \in \mathbf{val}(\mathbf{Pa}_i)\end{array} \\
& \xi_m \geq 0 \qquad\qquad \forall m
\end{aligned}$$

As stated in the main paper, the main restriction are the $(|\mathbf{val}(C)| - 1)\,M$ linear margin constraints. By expressing the slacks as $\xi_m = \max\left(\max_{c \neq c^m}\left[\gamma - \left(\phi_{c^m}(\mathbf{x}^m) - \phi_c(\mathbf{x}^m)\right)^T\boldsymbol{\omega}\right], 0\right)$, we can eleminiate these constraints, or in other words, they are absorbed into the objective. Since the *hinge function* $\max(\cdot, 0)$ and the $\max_{c \neq c^m}$ are not differentiable, we replace them by smooth approximations. The soft-hinge used in the paper is defined as

$$h_R(\zeta) = \begin{cases} 0 & \zeta < \mu \\ \zeta & \zeta > \mu + \frac{R}{\sqrt{2}} \\ R - \sqrt{R^2 - (\zeta - \mu)^2} & \text{o.w.} \end{cases} \tag{1}$$

The construction of the soft-hinge, by fitting a circle segment at the discontinuity, is illustrated in Figure 1. The derivative of of $h_R(\cdot)$ is given as

$$\frac{\partial h_R(\zeta)}{\partial \zeta} = \begin{cases} 0 & \zeta < \mu \\ 1 & \zeta > \mu + \frac{R}{\sqrt{2}} \\ \frac{\zeta - \mu}{\sqrt{R^2 - (\zeta - \mu)^2}} & \text{o.w.} \end{cases}. \tag{2}$$

The max function is approximated using the following soft-max function:

$$\operatorname*{smax}_{\zeta_1,\ldots,\zeta_L} = \frac{1}{\eta}\log\sum_{i=1}^{L}\exp(\eta\,\zeta_i) \tag{3}$$

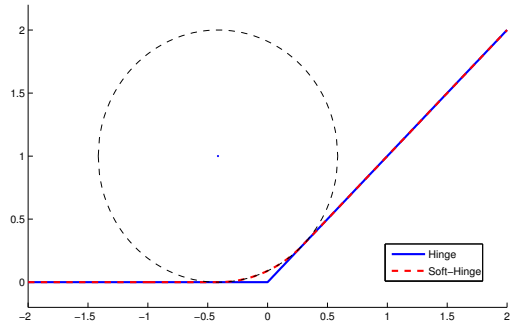

Figure 1. Construction of the soft-hinge by fitting a circle segment (here with radius $R = 1$) at the discontinuity of the (hard) hinge function.

Here η is a approximation parameter, where for $\eta \to \infty$ the soft-max converges to the (hard) max. The derivative of the soft-max is given as

$$\frac{\partial \operatorname*{smax}_{\zeta_1,\ldots,\zeta_L}}{\partial \zeta_i} = \frac{\exp(\eta\zeta_i)}{\sum_{l=1}^{L}\exp(\eta\zeta_l)}. \tag{4}$$

The smooth version of the ML-BN-SVM is

$$\begin{aligned} \min_{\boldsymbol{\omega}} \quad & -\mathbf{n}^T\boldsymbol{\omega} + \\ & \lambda \sum_{m=1}^{M} h_R\left(\operatorname*{smax}_{c\neq c^m}\left[\gamma - \left(\phi_{c^m}(\mathbf{x}^m) - \phi_c(\mathbf{x}^m)\right)^T\boldsymbol{\omega}\right]\right) \\ \text{s.t.} \quad & \log\sum_{j'}\exp\left(\omega^i_{j'|\mathbf{h}}\right) \leq 0 \qquad \begin{array}{l}\forall\ 0 \leq i \leq N \\ \forall\ \mathbf{h} \in \mathbf{val}(\mathbf{Pa}_i)\end{array} \end{aligned} \tag{5}$$

The objective

$$\begin{aligned} O(\boldsymbol{\omega}) = & -\mathbf{n}^T\boldsymbol{\omega} + \\ & \lambda \sum_{m=1}^{M} h_R\left(\operatorname*{smax}_{c\neq c^m}\left[\gamma - \left(\phi_{c^m}(\mathbf{x}^m) - \phi_c(\mathbf{x}^m)\right)^T\boldsymbol{\omega}\right]\right) \end{aligned} \tag{6}$$

is continuously differentiable, where the derivative is given as

$$\begin{aligned} & \frac{\partial O(\boldsymbol{\omega})}{\partial \omega^i_{j|\mathbf{h}}} = \\ & -n^i_{j|\mathbf{h}} - \lambda \sum_{m}^{M} \frac{\partial h_R}{\partial \operatorname{smax}} \cdot \sum_{c\neq c^m} \frac{\partial \operatorname{smax}}{\partial \xi^m_c} \cdot \left(\nu^{i,m}_{j|\mathbf{h}} - \nu^{i,m,c}_{j|\mathbf{h}}\right), \end{aligned} \tag{7}$$

where $\xi^m_c := \gamma - \left(\phi_{c^m}(\mathbf{x}^m) - \phi_c(\mathbf{x}^m)\right)\boldsymbol{\omega}$ and $\nu^{i,m,c}_{j|\mathbf{h}}$ is defined as $\nu^{i,m,c}_{j|\mathbf{h}} := \mathbb{1}(x^{m,c}_i = j \ \wedge\ \mathbf{x}^{m,c}(\mathbf{Pa}_i) = \mathbf{h})$, with $\mathbf{x}^{m,c} = [c, \mathbf{x}(\mathbf{Z})]$. The gradient is used in conjugate gradient descent, where $\boldsymbol{\omega}$ is projected onto the set of sub-normalized vectors after each gradient step.

Algorithm 1 Projection onto subnormalized set

Input: $\boldsymbol{\zeta}^*$, $\boldsymbol{\zeta}_0$ with $\log\sum_l \exp(\zeta_{0,l}) = 0$, $\rho > 0$
Output: $\boldsymbol{\zeta} = \arg\min_{\boldsymbol{\zeta}} \|\boldsymbol{\zeta}^* - \boldsymbol{\zeta}\|$, s.t. $\log\sum_l \exp\zeta_l \leq 0$

if $\log\sum_i \exp(\zeta^*_i) \leq 0$ **then**
  $\boldsymbol{\zeta} \leftarrow \boldsymbol{\zeta}^*$
  return
end if
$\boldsymbol{\zeta} \leftarrow \boldsymbol{\zeta}_0$
$\mathbf{g} \leftarrow \exp(\boldsymbol{\zeta})$
$\mathbf{g} \leftarrow \frac{\mathbf{g}}{\|\mathbf{g}\|_2}$
$\mathbf{d} \leftarrow \boldsymbol{\zeta}^* - \boldsymbol{\zeta}$
$\mathbf{d} \leftarrow \frac{\mathbf{d}}{\|\mathbf{d}\|_2}$
while $\mathbf{g}^T\mathbf{d} < 1$ **do**
  $\boldsymbol{\mu} = \boldsymbol{\zeta} - \rho\,\mathbf{g}$
  $\bar{\boldsymbol{\zeta}} = \boldsymbol{\mu} + \rho\,\mathbf{d}$
  if $\log\sum_l \exp(\bar{\zeta}_l) \leq 0$ **then**
    find κ: $\log\sum_l \exp\left(\bar{\zeta}_l + \kappa\,(\zeta^*_l - \bar{\zeta}_l)\right) = 0$
    $\boldsymbol{\zeta} \leftarrow \bar{\boldsymbol{\zeta}} + \kappa\,(\boldsymbol{\zeta}^* - \bar{\boldsymbol{\zeta}})$
  else
    find κ: $\log\sum_l \exp\left(\bar{\zeta}_l + \kappa\,(\zeta_l - \bar{\zeta}_l)\right) = 0$
    $\boldsymbol{\zeta} \leftarrow \bar{\boldsymbol{\zeta}} + \kappa\,(\boldsymbol{\zeta} - \bar{\boldsymbol{\zeta}})$
  end if
  $\mathbf{g} \leftarrow \exp(\boldsymbol{\zeta})$
  $\mathbf{g} \leftarrow \frac{\mathbf{g}}{\|\mathbf{g}\|_2}$
  $\mathbf{d} \leftarrow \boldsymbol{\zeta}^* - \boldsymbol{\zeta}$
  $\mathbf{d} \leftarrow \frac{\mathbf{d}}{\|\mathbf{d}\|_2}$
end while

This can be done for each CPT individually. For projecting, we use a variant of the algorithm described in (Lin, 2003), which projects an arbitrary vector onto the intersection of strictly convex sets. Here, we have the set $\mathcal{M} = \{\boldsymbol{\zeta} \mid \log \sum_l \exp(\zeta_l) \leq 0\}$, which is only a *single* strictly convex set. The algorithm is depicted in Algorithm 1, where $\boldsymbol{\zeta}^*$ is some arbitrary input vector, i.e. some CPT which has to be projected onto $\mathcal{M}$. The solution vector $\boldsymbol{\zeta}$ is initialized with some arbitrary vector $\boldsymbol{\zeta}_0$, with $\log \sum_l \exp(\zeta_{0,l}) = 0$. Vector $\mathbf{g}$ is the normalized gradient vector of the $\log \sum \exp(\cdot)$ function at the current solution vector $\boldsymbol{\zeta}$, which is the normal vector of $\mathcal{M}$. Vector $\mathbf{d}$ is the normalized residual vector. As easily shown via the KKT conditions, $\boldsymbol{\zeta}$ is optimal when $\mathbf{g} \propto \mathbf{d}$, as checked in step 10. Following (Lin, 2003), in each iteration, $\mathcal{M}$ is locally approximated with a ball of radius ρ and center $\boldsymbol{\mu}$, and the projection $\bar{\boldsymbol{\zeta}}$ onto this ball is calculated. In our experiments we used a radius $\rho = 1$. When $\bar{\boldsymbol{\zeta}}$ is feasible (steps 14-15), this solution is improved by finding the point closest to $\boldsymbol{\zeta}^*$ on the line segment $[\bar{\boldsymbol{\zeta}}, \boldsymbol{\zeta}^*]$. When $\bar{\boldsymbol{\zeta}}$ is infeasible (steps 17-18), a feasibility restoration is performed as depicted in (Lin, 2003). In both cases, the Newton-Raphson method is used to find scalar κ.

The projection algorithm interacts nicely with the projected gradient method, since we use the solution of the previous gradient step as initialization $\boldsymbol{\zeta}_0$. Therefore, since in each iteration of Algorithm 1 the distance $\|\boldsymbol{\zeta}^* - \boldsymbol{\zeta}\|$ is reduced (see (Lin, 2003)), we do not need to run the projection algorithm until convergence, but only for some few iterations (in fact, a single iteration is sufficient).

3. Detailed Classification Results

In the main paper we omitted results for the datasets "corral", "iris", "mofn-3-7-10", "mushroom", "glass2", and combined results for all "TIMIT" datasets. Table 1 shows all results for TAN structures in detail. The results for NB structures are shown in Table 2. Furthermore, in Table 3, we provide pairwise comparisons of all methods conducted on the UCI datasets: Plain numbers denote the number of times where the algorithm in the row outperforms the algorithm in the column at a significance level of 68%. Bold face numbers denote a significance level of 95%. When using 5-fold cross-validation for testing, we used a one-sided t-test, otherwise we used a one-sided binomial test for testing significance. Tables 4 and 5 show the corresponding results, when 50% and 90% percent of features are missing in the test data, respectively. Similar as in the main paper, these results demonstrate the robustess against missing features of ML and ML-BN-SVM parameters.

4. Effect of Early Stopping

In the main paper, we compared our method with state-of-the art maximum margin (MM) training for BNs (Pernkopf et al., 2012). In (Pernkopf et al., 2012), MM training was proposed with early stopping. This makes it hard to assess, to which part the classification performance stems from the problem formulation, and to which part from the early stopping heuristic. Therefore, in the main paper, we performed all experiments *without* early stopping. However, early stopping is easy to use, and an effective method to improve classification results. Here we show results for MM and ML-BN-SVM training when using early stopping; for both methods we performed gradient descent until convergence, but maximally for 10000 iterations, recording the performance on the validations set and storing maximizing parameter vectors. Finally, we used those parameters achieving the highest performance over all iterations and hyperparameters (γ and λ in our method, λ and κ for MM, see (Pernkopf et al., 2012)). Table 6 compares results with and without early stopping. We see that for NB, the ML-BN-SVM performs in 25 cases better than MM, while MM performs better in 9 cases. For TAN, the ML-BN-SVM performs in 22 cases better than MM, while MM performs better in 12 cases. We see that also in the case of early stopping the ML-BN-SVM performs favorable in comparison to MM. Furthermore, we see that early stopping tends to improve classification results significantly. In cases where methods with early stopping perform worse than the version without early stopping, the degradation is small.

References

Frank, A. and Asuncion, A. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2010. URL http://archive.ics.uci.edu/ml.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, August 2003.

Lin, A. A class of methods for projection on a convex set. *Advanced Modeling and Optimization (AMO)*, 5(3), 2003.

Pernkopf, F., Wohlmayr, M., and Tschiatschek, S. Maximum margin Bayesian network classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):521–531, 2012.

Table 1. Detailed classification rates with 95% confidence intervals for BN parameters, using **TAN** structures. ML: maximum likelihood, MCL: maximum condition likelihood, MM: maximum margin BN parameters (Pernkopf et al., 2012), ML-BN-SVM: proposed method, Linear SVM: support vector machine without kernel, SVM: support vector machine with Gauss kernel.

dataset	ML	MCL	MM	ML-BN-SVM	Linear SVM	SVM
abalone	57.70 ± 1.58	57.92 ± 1.65	57.78 ± 0.96	58.69 ± 1.86	58.42 ± 1.77	59.29 ± 1.40
adult	85.70 ± 0.66	86.65 ± 0.64	86.54 ± 0.65	86.76 ± 0.64	86.86 ± 0.64	86.87 ± 0.64
australian	81.67 ± 2.66	81.97 ± 3.70	85.49 ± 3.40	84.76 ± 3.78	85.78 ± 1.69	86.80 ± 2.34
breast	95.56 ± 2.06	95.56 ± 1.45	96.59 ± 0.50	96.00 ± 2.31	96.15 ± 1.51	97.19 ± 0.41
car	94.24 ± 1.50	98.08 ± 0.75	97.79 ± 0.79	98.08 ± 1.07	93.84 ± 0.65	99.65 ± 0.30
chess	92.19 ± 1.62	97.65 ± 0.81	97.43 ± 0.79	97.99 ± 0.92	97.02 ± 0.82	99.50 ± 0.25
cleve	79.43 ± 6.34	77.74 ± 7.53	79.09 ± 7.56	80.79 ± 7.58	83.57 ± 5.29	82.19 ± 6.37
corral	97.53 ± 4.61	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	93.36 ± 4.55	100.00 ± 0.00
crx	84.04 ± 4.64	80.32 ± 5.20	83.89 ± 5.89	84.20 ± 4.56	85.75 ± 3.20	85.75 ± 2.65
diabetes	74.35 ± 4.23	74.22 ± 5.50	73.31 ± 5.71	74.35 ± 5.42	73.96 ± 4.46	74.48 ± 4.65
flare	81.57 ± 1.27	81.48 ± 1.91	84.45 ± 0.28	83.30 ± 1.06	84.45 ± 0.28	84.45 ± 0.28
german	71.90 ± 1.83	69.50 ± 3.54	73.20 ± 4.01	72.60 ± 2.89	76.10 ± 1.11	75.80 ± 2.80
glass	72.68 ± 5.29	68.55 ± 4.03	71.71 ± 10.88	72.61 ± 6.35	71.61 ± 5.50	73.24 ± 5.33
glass2	81.38 ± 9.20	82.00 ± 8.05	80.75 ± 10.51	80.75 ± 10.51	79.38 ± 4.27	79.96 ± 8.90
heart	80.74 ± 10.36	77.04 ± 10.61	77.41 ± 9.81	81.48 ± 9.34	84.81 ± 4.11	81.85 ± 9.40
hepatitis	86.17 ± 10.00	86.08 ± 11.48	86.08 ± 3.38	86.17 ± 6.31	87.42 ± 10.89	88.67 ± 6.37
iris	94.00 ± 1.85	94.00 ± 1.85	92.67 ± 4.53	94.00 ± 1.85	93.33 ± 2.93	93.33 ± 2.93
letter	86.21 ± 0.84	87.65 ± 0.80	89.58 ± 0.74	88.57 ± 0.77	90.07 ± 0.73	94.07 ± 0.58
lymphography	80.77 ± 7.36	75.38 ± 10.86	80.66 ± 11.11	76.92 ± 10.54	83.57 ± 10.44	86.48 ± 9.99
mofn-3-7-10	92.62 ± 1.37	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
mushroom	100.00 ± 0.07	100.00 ± 0.07	100.00 ± 0.07	100.00 ± 0.07	100.00 ± 0.07	99.82 ± 0.19
nursery	92.96 ± 0.77	98.31 ± 0.40	98.84 ± 0.33	98.68 ± 0.35	93.31 ± 0.76	100.00 ± 0.04
satimage	85.79 ± 1.92	81.52 ± 0.95	86.82 ± 2.66	86.98 ± 1.30	88.36 ± 1.58	90.59 ± 1.59
segment	94.89 ± 1.02	94.37 ± 1.57	96.02 ± 1.21	95.76 ± 0.62	96.19 ± 0.73	96.84 ± 1.17
shuttle	99.88 ± 0.05	99.84 ± 0.06	99.91 ± 0.05	99.92 ± 0.04	99.96 ± 0.03	99.96 ± 0.03
soybean-large	91.88 ± 1.28	82.66 ± 4.59	90.77 ± 2.16	91.87 ± 2.26	91.15 ± 3.72	93.54 ± 1.19
spambase	92.97 ± 0.85	92.99 ± 1.10	93.62 ± 0.80	94.03 ± 0.84	94.27 ± 0.72	95.04 ± 0.37
TIMIT4CF	90.70 ± 0.42	87.25 ± 0.48	91.70 ± 0.40	91.59 ± 0.40	92.05 ± 0.39	92.38 ± 0.39
TIMIT4CM	90.47 ± 0.43	88.57 ± 0.46	85.62 ± 0.51	92.58 ± 0.38	92.88 ± 0.38	93.16 ± 0.37
TIMIT6CF	83.18 ± 0.52	80.92 ± 0.54	84.27 ± 0.50	84.89 ± 0.49	85.57 ± 0.48	85.74 ± 0.48
TIMIT6CM	83.05 ± 0.52	80.98 ± 0.54	85.45 ± 0.49	85.91 ± 0.48	86.66 ± 0.47	86.56 ± 0.47
USPS	91.20 ± 0.93	90.46 ± 0.97	95.98 ± 0.65	95.98 ± 0.65	95.82 ± 0.66	91.80 ± 0.90
vehicle	70.60 ± 2.00	69.64 ± 3.69	69.04 ± 4.30	69.88 ± 2.41	70.12 ± 1.26	69.76 ± 2.43
vote	94.37 ± 2.62	94.15 ± 2.04	96.01 ± 2.45	95.31 ± 2.74	94.85 ± 2.20	95.54 ± 3.18
waveform-21	82.36 ± 0.71	80.55 ± 1.00	82.86 ± 0.51	83.48 ± 0.56	84.78 ± 1.77	85.16 ± 1.29

Table 2. Detailed classification rates with 95% confidence intervals for BN parameters, using **NB** structures.
ML: maximum likelihood, MCL: maximum condition likelihood, MM: maximum margin BN parameters (Pernkopf et al., 2012), ML-BN-SVM: proposed method, Linear SVM: support vector machine without kernel, SVM: support vector machine with Gauss kernel.

dataset	ML	MCL	MM	ML-BN-SVM	Linear SVM	SVM
abalone	53.64 ± 1.45	59.12 ± 1.71	56.62 ± 0.88	59.12 ± 1.69	58.42 ± 1.77	59.29 ± 1.40
adult	83.37 ± 0.71	86.90 ± 0.64	86.92 ± 0.64	86.94 ± 0.64	86.86 ± 0.64	86.87 ± 0.64
australian	85.92 ± 2.92	84.02 ± 2.76	85.34 ± 2.64	87.24 ± 2.86	85.78 ± 1.69	86.80 ± 2.34
breast	97.63 ± 1.01	95.56 ± 1.45	95.85 ± 2.22	97.04 ± 1.45	96.15 ± 1.51	97.19 ± 0.41
car	85.64 ± 1.59	93.43 ± 1.76	93.78 ± 1.63	92.73 ± 1.14	93.84 ± 0.65	99.65 ± 0.30
chess	87.45 ± 2.57	97.11 ± 1.02	97.58 ± 0.86	97.68 ± 1.21	97.02 ± 0.82	99.50 ± 0.25
cleve	82.87 ± 6.79	82.52 ± 6.36	82.17 ± 6.94	82.53 ± 7.64	83.57 ± 5.29	82.19 ± 6.37
corral	89.16 ± 8.67	93.36 ± 4.55	93.36 ± 4.55	93.36 ± 4.55	93.36 ± 4.55	100.00 ± 0.00
crx	86.84 ± 3.29	85.13 ± 4.10	84.82 ± 3.71	86.06 ± 3.54	85.75 ± 3.20	85.75 ± 2.65
diabetes	73.96 ± 4.17	75.40 ± 5.41	74.61 ± 4.94	74.87 ± 3.47	73.96 ± 4.46	74.48 ± 4.65
flare	76.58 ± 1.04	83.40 ± 1.02	82.63 ± 1.79	83.11 ± 0.82	84.45 ± 0.28	84.45 ± 0.28
german	74.20 ± 3.58	75.10 ± 1.42	76.50 ± 1.52	75.30 ± 3.12	76.10 ± 1.11	75.80 ± 2.80
glass	71.66 ± 3.58	68.05 ± 0.63	68.03 ± 1.91	70.61 ± 3.63	71.61 ± 5.50	73.24 ± 5.33
glass2	81.29 ± 10.50	82.63 ± 8.12	80.09 ± 9.96	82.63 ± 8.12	79.38 ± 4.27	79.96 ± 8.90
heart	81.85 ± 9.40	82.59 ± 5.77	81.85 ± 5.73	83.33 ± 5.14	84.81 ± 4.11	81.85 ± 9.40
hepatitis	88.58 ± 6.57	86.08 ± 3.38	84.92 ± 8.69	92.33 ± 6.75	87.42 ± 10.89	88.67 ± 6.37
iris	93.33 ± 2.93	92.67 ± 3.46	93.33 ± 2.93	93.33 ± 2.93	93.33 ± 2.93	93.33 ± 2.93
letter	74.95 ± 1.05	85.97 ± 0.84	82.53 ± 0.92	85.79 ± 0.85	90.07 ± 0.73	94.07 ± 0.58
lymphography	84.23 ± 5.60	84.23 ± 4.47	82.80 ± 5.54	82.80 ± 4.39	83.57 ± 10.44	86.48 ± 9.99
mofn-3-7-10	87.31 ± 1.94	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
mushroom	98.04 ± 0.54	100.00 ± 0.07	100.00 ± 0.07	99.78 ± 0.20	100.00 ± 0.07	99.82 ± 0.19
nursery	89.97 ± 0.91	92.38 ± 0.80	92.98 ± 0.77	93.03 ± 0.77	93.31 ± 0.76	100.00 ± 0.04
satimage	81.56 ± 1.80	87.29 ± 1.11	88.82 ± 1.26	88.41 ± 1.33	88.36 ± 1.58	90.59 ± 1.59
segment	92.68 ± 1.78	94.29 ± 0.77	94.98 ± 1.66	95.37 ± 0.86	96.19 ± 0.73	96.84 ± 1.17
shuttle	99.62 ± 0.09	99.91 ± 0.05	99.94 ± 0.04	99.95 ± 0.04	99.96 ± 0.03	99.96 ± 0.03
soybean-large	93.35 ± 1.91	92.98 ± 3.88	92.79 ± 1.59	91.50 ± 3.81	91.15 ± 3.72	93.54 ± 1.19
spambase	90.03 ± 1.11	93.73 ± 0.95	94.01 ± 0.97	94.08 ± 0.75	94.27 ± 0.72	95.04 ± 0.37
TIMIT4CF	87.88 ± 0.47	92.04 ± 0.39	91.90 ± 0.40	91.95 ± 0.39	92.05 ± 0.39	92.38 ± 0.39
TIMIT4CM	88.86 ± 0.46	93.04 ± 0.37	92.88 ± 0.38	92.71 ± 0.38	92.88 ± 0.38	93.16 ± 0.37
TIMIT6CF	82.20 ± 0.53	85.50 ± 0.49	85.20 ± 0.49	85.49 ± 0.49	85.57 ± 0.48	85.74 ± 0.48
TIMIT6CM	82.43 ± 0.53	86.24 ± 0.48	86.04 ± 0.48	86.50 ± 0.47	86.66 ± 0.47	86.56 ± 0.47
USPS	86.89 ± 1.11	94.37 ± 0.76	95.44 ± 0.69	95.08 ± 0.71	95.82 ± 0.66	91.80 ± 0.90
vehicle	61.57 ± 1.44	68.67 ± 3.03	69.76 ± 2.56	67.95 ± 6.00	70.12 ± 1.26	69.76 ± 2.43
vote	90.16 ± 4.70	94.61 ± 2.21	95.78 ± 2.21	94.61 ± 3.19	94.85 ± 2.20	95.54 ± 3.18
waveform-21	81.14 ± 1.05	85.10 ± 1.53	85.43 ± 1.34	85.14 ± 1.52	84.78 ± 1.77	85.16 ± 1.29

Table 3. Number of times classifier in row outperforms classifier in column with significance 68 % (plain) and 95 % (bold), when no features are missing.

	ML		MCL		MM		ML-BN-SVM		SVM	
	NB	TAN	NB	TAN	NB	TAN	NB	TAN	Linear	Gauss
ML NB	–	9/**5**	8/**4**	11/**7**	9/**4**	11/**4**	7/**2**	9/**5**	5/**1**	5/**0**
ML TAN	20/**18**	–	8/**3**	14/**8**	8/**3**	8/**1**	7/**3**	4/**1**	6/**1**	4/**1**
MCL NB	21/**18**	20/**10**	–	19/**11**	10/**2**	13/**6**	8/**1**	13/**2**	5/**1**	5/**2**
MCL TAN	17/**14**	8/**6**	7/**5**	–	7/**4**	8/**0**	7/**5**	1/**0**	6/**4**	3/**1**
MM NB	20/**18**	15/**11**	14/**8**	17/**11**	–	15/**8**	9/**3**	12/**4**	8/**2**	4/**2**
MM TAN	18/**18**	18/**12**	12/**7**	19/**12**	10/**6**	–	10/**8**	8/**3**	8/**4**	3/**2**
ML-BN-SVM NB	24/**19**	21/**11**	15/**9**	21/**14**	14/**7**	20/**8**	–	15/**4**	9/**3**	7/**1**
ML-BN-SVM TAN	19/**18**	21/**15**	13/**8**	21/**16**	12/**8**	15/**3**	12/**6**	–	10/**4**	3/**2**
LinSVM	21/**18**	22/**14**	19/**7**	21/**14**	16/**6**	15/**7**	15/**7**	15/**8**	–	6/**2**
SVM	23/**18**	26/**18**	20/**14**	25/**18**	18/**12**	25/**13**	17/**10**	21/**11**	17/**9**	–

Table 4. Number of times classifier in row outperforms classifier in column with significance 68 % (plain) and 95 % (bold), with 50% missing features.

	ML		MCL		MM		ML-BN-SVM		SVM	
	NB	TAN	NB	TAN	NB	TAN	NB	TAN	Linear	Gauss
ML NB	–	8/**2**	23/**19**	20/**13**	25/**18**	25/**15**	14/**7**	8/**5**	11/**5**	12/**3**
ML TAN	20/**13**	–	24/**20**	25/**16**	26/**21**	28/**17**	18/**13**	9/**3**	13/**4**	13/**3**
MCL NB	6/**2**	2/**0**	–	13/**5**	11/**7**	16/**9**	2/**0**	2/**2**	4/**3**	4/**2**
MCL TAN	10/**7**	4/**1**	15/**9**	–	15/**13**	19/**8**	9/**6**	5/**2**	5/**3**	5/**2**
MM NB	5/**3**	5/**2**	14/**11**	12/**7**	–	16/**10**	4/**4**	3/**1**	2/**1**	3/**2**
MM TAN	5/**4**	2/**1**	11/**9**	11/**6**	10/**6**	–	6/**4**	4/**2**	1/**0**	1/**0**
ML-BN-SVM NB	12/**6**	7/**1**	25/**19**	18/**12**	23/**15**	22/**13**	–	7/**2**	11/**4**	11/**6**
ML-BN-SVM TAN	18/**11**	13/**3**	25/**19**	24/**20**	26/**22**	27/**17**	18/**11**	–	10/**4**	10/**6**
LinSVM	17/**11**	12/**6**	26/**22**	25/**18**	26/**23**	27/**17**	18/**11**	14/**7**	–	10/**3**
SVM	16/**12**	13/**9**	25/**22**	24/**19**	26/**20**	25/**17**	17/**12**	14/**9**	15/**7**	–

Table 5. Number of times classifier in row outperforms classifier in column with significance 68 % (plain) and 95 % (bold), with 90% missing features.

	ML		MCL		MM		ML-BN-SVM		SVM	
	NB	TAN	NB	TAN	NB	TAN	NB	TAN	Linear	Gauss
ML NB	–	3/**0**	22/**16**	20/**15**	26/**18**	24/**17**	18/**10**	16/**4**	23/**12**	24/**13**
ML TAN	8/**4**	–	22/**16**	20/**14**	26/**18**	24/**17**	19/**11**	18/**4**	24/**13**	25/**14**
MCL NB	0/**0**	0/**0**	–	8/**4**	14/**7**	13/**7**	6/**2**	5/**1**	7/**4**	8/**4**
MCL TAN	3/**1**	2/**1**	15/**10**	–	13/**8**	15/**8**	7/**5**	2/**1**	13/**5**	11/**6**
MM NB	0/**0**	0/**0**	11/**8**	10/**5**	–	15/**8**	7/**3**	6/**2**	7/**5**	8/**4**
MM TAN	0/**0**	0/**0**	9/**5**	8/**6**	9/**3**	–	4/**3**	3/**1**	9/**6**	10/**6**
ML-BN-SVM NB	1/**0**	2/**0**	18/**11**	14/**7**	19/**9**	20/**13**	–	6/**2**	16/**6**	14/**7**
ML-BN-SVM TAN	5/**3**	3/**1**	19/**14**	20/**11**	22/**14**	20/**14**	17/**10**	–	23/**11**	23/**11**
LinSVM	2/**2**	1/**1**	17/**10**	13/**8**	17/**10**	19/**9**	7/**3**	5/**1**	–	7/**4**
SVM	3/**2**	2/**1**	15/**11**	14/**7**	17/**9**	19/**9**	8/**3**	6/**1**	11/**4**	–

Table 6. Classification results for MM (Pernkopf et al., 2012) and ML-BN-SVM (this paper), with and without early stopping.

	without early stopping				with early stopping			
	MM		ML-BN-SVM		MM		ML-BN-SVM	
dataset	NB	TAN	NB	TAN	NB	TAN	NB	TAN
abalone	56.62 ± 0.88	57.78 ± 0.96	59.12 ± 1.69	58.69 ± 1.86	58.16 ± 0.96	58.11 ± 1.65	58.88 ± 1.71	58.90 ± 1.49
adult	86.92 ± 0.64	86.54 ± 0.65	86.94 ± 0.64	86.76 ± 0.64	86.89 ± 0.64	86.38 ± 0.65	86.96 ± 0.64	86.47 ± 0.65
australian	85.34 ± 2.64	85.49 ± 3.40	87.24 ± 2.86	84.76 ± 3.78	85.48 ± 3.57	85.04 ± 2.33	86.80 ± 2.75	85.93 ± 1.95
breast	95.85 ± 2.22	96.59 ± 0.50	97.04 ± 1.45	96.00 ± 2.31	97.04 ± 0.65	96.59 ± 1.05	97.04 ± 0.92	96.74 ± 1.67
car	93.78 ± 1.63	97.79 ± 0.79	92.73 ± 1.14	98.08 ± 1.07	93.84 ± 1.68	98.26 ± 0.92	92.97 ± 1.43	97.85 ± 0.83
chess	97.58 ± 0.86	97.43 ± 0.79	97.68 ± 1.21	97.99 ± 0.92	97.21 ± 0.94	97.40 ± 0.62	97.62 ± 1.33	97.93 ± 0.84
cleve	82.17 ± 6.94	79.09 ± 7.56	82.53 ± 7.64	80.79 ± 7.58	81.51 ± 7.16	83.90 ± 4.95	82.53 ± 7.49	83.55 ± 7.08
corral	93.36 ± 4.55	100.00 ± 0.00	93.36 ± 4.55	100.00 ± 0.00	87.73 ± 10.44	100.00 ± 0.00	93.36 ± 4.55	100.00 ± 0.00
crx	84.82 ± 3.71	83.89 ± 5.89	86.06 ± 3.54	84.20 ± 4.56	86.21 ± 3.96	84.81 ± 5.20	86.37 ± 3.26	84.97 ± 3.64
diabetes	74.61 ± 4.94	73.31 ± 5.71	74.87 ± 3.47	74.35 ± 5.42	74.22 ± 4.01	73.96 ± 4.14	73.44 ± 4.14	74.61 ± 5.09
flare	82.63 ± 1.79	84.45 ± 0.28	83.11 ± 0.82	83.30 ± 1.06	81.09 ± 2.92	84.26 ± 0.73	83.88 ± 0.34	84.17 ± 0.57
german	76.50 ± 1.52	73.20 ± 4.01	75.30 ± 3.12	72.60 ± 2.89	74.10 ± 1.42	72.00 ± 2.15	74.60 ± 2.46	74.70 ± 4.09
glass	68.03 ± 1.91	71.71 ± 10.88	70.61 ± 3.63	72.61 ± 6.35	71.61 ± 6.96	71.13 ± 5.18	72.16 ± 4.60	72.13 ± 6.23
glass2	80.09 ± 9.96	80.75 ± 10.51	82.63 ± 8.12	80.75 ± 10.51	83.98 ± 6.92	83.34 ± 6.52	81.29 ± 10.50	84.00 ± 7.38
heart	81.85 ± 5.73	77.41 ± 9.81	83.33 ± 5.14	81.48 ± 9.34	82.96 ± 6.97	80.74 ± 9.97	81.48 ± 8.13	82.22 ± 10.61
hepatitis	84.92 ± 8.69	86.08 ± 3.38	92.33 ± 6.75	86.17 ± 6.31	89.83 ± 8.95	89.92 ± 6.86	96.17 ± 4.35	87.42 ± 7.63
iris	93.33 ± 2.93	92.67 ± 4.53	93.33 ± 2.93	94.00 ± 1.85	93.33 ± 2.93	94.00 ± 1.85	93.33 ± 2.93	94.67 ± 2.27
letter	82.53 ± 0.92	89.58 ± 0.74	85.79 ± 0.85	88.57 ± 0.77	82.40 ± 0.92	89.55 ± 0.74	86.06 ± 0.84	90.25 ± 0.72
lymphography	82.80 ± 5.54	80.66 ± 11.11	82.80 ± 4.39	76.92 ± 10.54	83.52 ± 11.07	82.91 ± 10.65	86.54 ± 10.49	82.14 ± 5.76
mofn-3-7-10	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.90 ± 0.27	100.00 ± 0.00	100.00 ± 0.00
mushroom	100.00 ± 0.07	100.00 ± 0.07	99.78 ± 0.20	100.00 ± 0.07	99.56 ± 0.27	100.00 ± 0.07	99.67 ± 0.24	100.00 ± 0.07
nursery	92.98 ± 0.77	98.84 ± 0.33	93.03 ± 0.77	98.68 ± 0.35	92.66 ± 0.79	98.80 ± 0.34	92.92 ± 0.78	98.38 ± 0.39
satimage	88.82 ± 1.26	86.82 ± 2.66	88.41 ± 1.33	86.98 ± 1.30	89.17 ± 1.39	88.33 ± 1.60	88.61 ± 1.42	87.68 ± 1.47
segment	94.98 ± 1.66	96.02 ± 1.21	95.37 ± 0.86	95.76 ± 0.62	94.94 ± 1.21	95.80 ± 1.15	95.15 ± 0.62	95.54 ± 0.94
shuttle	99.94 ± 0.04	99.91 ± 0.05	99.95 ± 0.04	99.92 ± 0.04	99.94 ± 0.04	99.91 ± 0.05	99.96 ± 0.03	99.91 ± 0.05
soybean-large	92.79 ± 1.59	90.77 ± 2.16	91.50 ± 3.81	91.87 ± 2.26	92.62 ± 1.61	91.32 ± 3.30	92.24 ± 1.80	92.79 ± 1.95
spambase	94.01 ± 0.97	93.62 ± 0.80	94.08 ± 0.75	94.03 ± 0.84	93.99 ± 0.66	94.27 ± 0.59	93.97 ± 0.80	94.06 ± 0.39
TIMIT4CF	91.90 ± 0.40	91.70 ± 0.40	91.95 ± 0.39	91.59 ± 0.40	91.82 ± 0.40	87.46 ± 0.48	91.95 ± 0.39	91.78 ± 0.40
TIMIT4CM	92.88 ± 0.38	85.62 ± 0.51	92.71 ± 0.38	92.58 ± 0.38	92.89 ± 0.38	85.84 ± 0.51	92.88 ± 0.38	92.62 ± 0.38
TIMIT6CF	85.20 ± 0.49	84.27 ± 0.50	85.49 ± 0.49	84.89 ± 0.49	85.20 ± 0.49	83.86 ± 0.51	85.21 ± 0.49	84.99 ± 0.49
TIMIT6CM	86.04 ± 0.48	85.45 ± 0.49	86.50 ± 0.47	85.91 ± 0.48	85.98 ± 0.48	85.68 ± 0.49	86.47 ± 0.47	86.04 ± 0.48
USPS	95.44 ± 0.69	95.98 ± 0.65	95.08 ± 0.71	95.98 ± 0.65	94.89 ± 0.73	95.77 ± 0.67	95.68 ± 0.67	95.44 ± 0.69
vehicle	69.76 ± 2.56	69.04 ± 4.30	67.95 ± 6.00	69.88 ± 2.41	66.99 ± 3.10	70.60 ± 1.93	68.80 ± 4.41	70.72 ± 1.70
vote	95.78 ± 2.21	96.01 ± 2.45	94.61 ± 3.19	95.31 ± 2.74	96.01 ± 3.50	95.32 ± 2.72	95.31 ± 3.86	94.37 ± 2.40
waveform-21	85.43 ± 1.34	82.86 ± 0.51	85.14 ± 1.52	83.48 ± 0.56	85.29 ± 1.26	84.18 ± 0.59	85.55 ± 0.98	84.00 ± 0.90