Towards the Optimal Feature Selection in High-Dimensional Bayesian Network Classifiers

Tatjana Pavlenko, Mikael Hall, and Dietrich von Rosen

Research Report
Centre of Biostochastics

Swedish University of
Agricultural Sciences

Report 2004:1
ISSN 1651-8543

Towards the Optimal Feature Selection in High-Dimensional Bayesian Network Classifiers

Tatjana Pavlenko[1]
Department of Engineering, Physics and Mathematics
Mid Sweden University, SE-85170, Sundsvall, Sweden

Mikael Hall
Department of Information Technology and Media
Mid Sweden University, SE-85170, Sundsvall, Sweden

Dietrich von Rosen
Centre of Biostochastics
Swedish University of Agricultural Sciences
Box 7013, SE-75007, Uppsala, Sweden

Abstract

Incorporating subset selection into a classification method often carries a number of advantages, especially when operating in the domain of high-dimensional features. In this paper, we focus on Bayesian network (BN) classifiers and formalize the feature selection from a perspective of improving classification accuracy. To exploring the effect of high-dimensionality we apply the *growing dimension asymptotics*, meaning that the number of training examples is relatively small compared to the number of feature nodes. In order to ascertain which set of features is indeed relevant for a classification task, we introduce a *distance-based scoring measure* reflecting how well the set separates different classes. This score is then employed to feature selection, using the weighted form of BN classifier. The idea is to view weights as inclusion-exclusion factors which eliminates the sets of features whose separation score do not exceed a given threshold. We establish the asymptotic optimal threshold and demonstrate that the proposed selection technique carries improvements over classification accuracy for different a priori assumptions concerning the separation strength.

Keywords: Bayesian network, augmenting, separation strength, growing dimension asymptotic, weighted classifier, subset selection, limiting error probability.

AMS 1991 subject classifications: Primary 62H30; Secondary 62F12

[1] E-mail address all correspondence and requests for reprints to: tatjana@fmi.mh.se

1 Introduction

This paper is about techniques for improving the performance accuracy of the classification methods in high-dimensional framework. Such methods operate on a given set of examples to produce a *classifier*, sometimes also called classification rule, or in the machine-learning literature, a prediction algorithm. The goal is to find a classifier with high performance accuracy, that is a low misclassification rate on a separate test set. In our analysis, we employ Bayesian Network (BN) models, which have an increasing number of applications in classification theory (e.g. Cowell et al. 1999) as well as in decision analysis and artificial intelligence (Korb and Nicholson 2003), offering complementary advantages such as ability to deal effectively with uncertain and high-dimensional examples.

The focus of the study is on *feature selection*, which involves identifying a set of feature nodes of the input examples that are highly relevant for classification task. More generally, feature selection can be viewed as a problem of setting discrete structural parameters associated with specific classification method. We subscribe here to the view that feature selection is not merely for reducing the computational load associated with a high-dimensional classification problem but can be tailored primarily to improve performance accuracy.

The common need for all subset selection procedures is an evaluation function by which a separation strength of a feature, or a subset of features, is assessed. Assuming that the performance measure is defined by the misclassification risk, the latter seems to be a most appealing function for this. However, it was shown in (Pavlenko and von Rosen 2001), that under rather mild regularity conditions for class probability densities, the asymptotic misclassification risk can be expressed as a monotone transform of the cross-entropy distance between the classes. This theoretically justifies the use of a distance-based scoring measure since it induces, over the set of all potential features, the same ranking as the one induced by the misclassification risk. The form of feature selection we develop in this study is an extended version of the *weighting technique* recently proposed for augmented BN classifiers in order to make more pronounced the inputs of highly informative sets and thereby increase the performance accuracy, (Pavlenko and von Rosen 2002). Given the separation score of each set of features, we redefine the weight-functions as *inclusion-exclusion* factors which depend on a selection threshold and reflect whether the set is chosen for the classification. The main objective is to specify the optimal threshold (score), so that the classifier will then be based only on the sets that have this score or higher.

We emphasize that the feature selection is carried out jointly and discriminatively together with estimation of the specific augmented BN classifier. This type of feature selection is, clearly affected by inaccuracy of estimates involved. The effect is especially pronounced in a high-dimensional setting, i.e. when the number of training examples is relatively small compared to the number of feature nodes. We employ a *growing dimension asymptotics* approach and show how it can accommodate BN classifies and enables us to evaluate the asymptotic distribution of the classifier and optimize the threshold, simultaneously taking into account bias and variance effects.

The paper is organized as follows: in Section 2 we present augmented BN models and introduce the growing dimension asymptotics. Distance-based scoring measure is derived in Section 3 together with asymptotically optimal weighting schemes. These results are then used in Section 4 to define the selection technique and specify the optimal threshold. We conclude in Section 5.

2 BN classifier in high-dimensional framework

A BN model $\mathcal{M} = \langle \mathcal{G}, \mathcal{F}_{\mathcal{G}} \rangle$ for a set of random variables $\mathbf{x} = \{x_1, \ldots, x_p\}$ is a set of joint probability distributions, specified via two components: a structure $\mathcal{G}$ and a set of local distribution families $\mathcal{F}_{\mathcal{G}}$. The structure $\mathcal{G}$ for $\mathbf{x}$ is a directed acyclic graph having for every variable x_i in $\mathbf{x}$ a node labelled by x_i with parents labelled by $Pa_i^{\mathcal{M}}$. In this way $\mathcal{G}$ represents a set of conditional independence assertions which implies a factorization of the joint distribution of $\mathbf{x}$ into $F(\mathbf{x}) = \prod_{i=1}^{p} F(x_i|Pa_i^{\mathcal{M}})$, where $F(x_i|Pa_i^{\mathcal{M}})$ are the p conditional and marginal probability distributions and each $F(x_i|Pa_i^{\mathcal{M}})$ belongs to the specific family of allowable local distributions $\mathcal{F}_{\mathcal{G}}$. We assume that $\mathbf{x}$ consists of continuous random variables and each local probability distribution is selected from a family $\mathcal{F}_{\mathcal{G},\Theta}$ which depends on a finite set of parameters $\theta \in \Theta$. The parameters for a local distribution are a set of real numbers that completely determine the functional form of $F(x_i|Pa_i^{\mathcal{M}})$, given $\mathcal{M}$. Consequently, the joint probability density for a BN model is represented by

$$f(x_1, \ldots, x_p; \theta) = \prod_{i=1}^{p} f(x_i; \theta_i | Pa_i^{\mathcal{M}}),$$

where $\theta_1, \ldots, \theta_p$ are subsets of θ and $f(x_i; \theta_i|Pa_i^{\mathcal{M}})$ are conditional local densities.

In the current study, BN models will be embedded in the *classification framework* where the outcome of interests, c, falls into ν unordered classes, which for convenience we denote by the set $\{1, 2, \ldots, \nu\}$. The goal is to build a rule for assessing the class membership of an item based on p feature variables $\mathbf{x} \in R^p$, whose joint conditional probability density in each class is represented by a BN model, $\mathcal{M}$, having its own set of parameters, but sharing a common structure. Using Bayes' theorem and flipping the densities into class posterior probabilities $\mathsf{Pr}(c\,|\mathbf{x})$ we construct the classification rule

$$c = j \quad \text{if} \quad \mathsf{Pr}(c = j\,|\mathbf{x}) = \max_k \mathsf{Pr}(c = k\,|\mathbf{x}), \tag{1}$$

where $\mathsf{Pr}(c = j|\mathbf{x}) \propto \pi_j f(\mathbf{x}; \theta^j)$, $\mathsf{Pr}(c = j) = \pi_j$ are class prior probabilities, $j = 1, \ldots, \nu$ and $\propto$ denotes proportionality. This is in fact the definition of a *general Bayesian network classifier* (BN classifier) commonly found in the literature (e.g. Cowell et al. 1999).

2.1 Augmenting via binary classification

A well-known example of the BN classifier is the *naive Bayesian classifier*, which is a network with one arc from the class node c to each of the feature nodes x_i. In this case, $\mathcal{G}$ represents the assumption that the feature variables are conditionally independent, given the class variable, from which it is immediate that $f^j(\mathbf{x}; \theta) = \prod_{i=1}^p f_i^j(\mathbf{x}_i; \theta)$, $j = 1, \ldots, \nu$. Hence, $\mathsf{Pr}(c = j\,|\mathbf{x}) \propto \pi_j \prod_{i=1}^p f_i^j(\mathbf{x}_i; \theta)$. It is worth noting that the naive BN classifier does surprisingly well when only a finite sample of training observation is available. This behavior has been noted in (e.g. Hastie et al. 2001). The naive Bayesian approach also turns out to be effective when studying the high dimensionality effect; see for instance, (Friedman 1997), where the bias and variance induced on the class density estimation by the naive Bayes decomposition and their effect on classification have been studied.

However, the total conditional independence inherent to the naive BN is far from being realized in most applications. To relax this assumption we use the methodology proposed by (Friedman et. al 1997) where the problem was approached by *augmenting* the naive BN model by allowing additional arcs between the nodes that capture possible dependencies among them. In this way, the original set of nodes is decomposed into several subsets and, requiring the subsets to be *non-overlapping*, the network structure $\mathcal{G}$ forms a decomposition of both $\mathbf{x}$ and θ^j into κ pairwise disjoint, independent, m-dimensional subsets, so that $p = \kappa m$ and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_\kappa)$, $\theta^j = (\theta_1^j, \ldots, \theta_\kappa^j)$ where $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$, $\theta_i^j = (\theta_{i1}^j, \ldots, \theta_{im}^j)$ $i = 1, \ldots, \kappa$, $j = 1, 2$. We call these structures *augmented Bayesian networks* (augmented BN) and the subsets *blocks*.

Augmenting the network followed by the block independence implies that the joint probability density of $\mathbf{x}$ for each class can be decomposed into a product of local interaction models, i.e. $f(\mathbf{x}; \theta^j) = \prod_{i=1}^\kappa f_i(\mathbf{x}_i; \theta_i^j)$, where the local m-dimensional density $f_i(\mathbf{x}_i; \theta_i^j)$ belongs to a family $\mathcal{F}_{\mathcal{G},\Theta}$ which depends on a finite set of parameters $\theta_i^j \in \Theta$, $i = 1, \ldots, \kappa$, $j = 1, 2$. This makes it possible to build a classification model separately for each block, and then combine all local classifiers.

To learn the augmented BN classifier we need to choose parametric families for representing the local class conditional densities. Given $\mathcal{G}$ we assume that the family $\mathcal{F}_{\mathcal{G},\Theta}$ satisfies the following regularity conditions: for each $\mathbf{x}_i$, the function $\ell_i(\mathbf{x}_i; \theta_i^j) := \ln f_i(\mathbf{x}_i; \theta_i^j)$ is three times differentiable in the components of θ_i^j and all first-, second- and third- order derivatives with respect to θ_i^j of $\ell(\mathbf{x}_i; \theta_i^j)$ are integrable with respect to $f(\mathbf{x}; \theta_j)\, d\mathbf{x}$, $j = 1, 2$.

In what follows we restrict ourselves to *binary classification*, the special (but common) case in which $\nu = 2$ and assign to the class that wins the most pairwise comparisons. Further, we will make use of the decision boundaries that are expressed in terms of a logarithmic difference between two densities, i.e. the *discriminant score*,

$$\mathcal{C}(\mathbf{x}; \theta^1, \theta^2) = \ell(\mathbf{x}; \theta^1) - \ell(\mathbf{x}; \theta^2).$$

To motivate why this representation of the classifier is attractive, we note first that the score $\mathcal{C}(\mathbf{x}; \theta^1, \theta^2)$ preserves the ordering of the class posterior probabilities leading to the decision rule

$$c(\mathbf{x}) = \begin{cases} 1 & \text{whenever} \quad \mathcal{C}(\mathbf{x}; \theta^1, \theta^2) > \ln \frac{\pi_2}{\pi_1}, \\ 2 & \text{otherwise.} \end{cases} \tag{2}$$

which is equivalent to (1).

$$\mathcal{E}_1 = \mathsf{Pr}(\mathcal{C}(\mathbf{x}; \theta^1, \theta^2) \leq \ln \frac{\pi_2}{\pi_1} | c(\mathbf{x}) = 1), \tag{3}$$

$$\mathcal{E}_2 = \mathsf{Pr}(\mathcal{C}(\mathbf{x}; \theta^1, \theta^2) > \ln \frac{\pi_2}{\pi_1} | c(\mathbf{x}) = 2). \tag{4}$$

These in turn form the *Bayes risk* $\mathcal{R}_{\mathcal{C}(\mathbf{x};\theta^1,\theta^2)} = \pi_1\mathcal{E}_1 + \pi_2\mathcal{E}_2$, which gives a straightforward way of judging the classification accuracy. Note also that in the symmetric case with equal prior probabilities both class-wise error rates are equal, and the minimum attainable Bayes risk is $\mathcal{R}_{\mathcal{C}(\mathbf{x};\theta^1,\theta^2)} = \frac{1}{2}(\mathcal{E}_1 + \mathcal{E}_2)$.

2.2 High-dimensional framework and estimates

A theoretically sound way to deal with the high-dimensional problem is to turn to a general asymptotic approach, meaning that a relationship between dimensionality and sample size satisfies the condition:

$$\lim_{n_j \to \infty} \lambda(p, n_j) < \infty,$$

where $\lambda(p, n_j)$ is a positive function increasing along p and decreasing along n_j, $j = 1, 2$. Since the increase of p and n_j is somehow simultaneous in a high-dimensional setting, the asymptotic approach we are going to work with can be based on the ratio

$$\lim_{n_j \to \infty} \frac{p}{n_j} = \eta, \tag{5}$$

where $0 < \eta < \infty$ is a certain constant for each $j = 1, 2$. This approach is often referred to under the name of *growing dimension asymptotics* (Pavlenko and von Rosen 2001) and the goal is to apply this to explore the high-dimensionality effect on the classification performance. Regarding n_j, in this study we assume the same rate of growing for both samples sizes so that $n_1 = n_2 = n$.

In order to completely specify the learning method in the context of augmented BN model, we define the asymptotic properties of estimates $\hat{\theta}_i^j$ to be plugged-in into $\mathcal{C}(\mathbf{x}; \theta^1, \theta^2)$. We introduce the statistics $T_i^j = n^{1/2}(\hat{\theta}_i^j - \theta_i^j)' I^{1/2}(\theta_i^j)$, which for each $i = 1, \ldots, \kappa$ describes the standardized bias of the estimate $\hat{\theta}_i^j$, where $I^j = I(\theta^j)$ is the Fisher information matrix which is positive definite for all $\theta^j \in \Theta^j$ and whose eigenvalues are bounded from above. By the network structure, the matrices are of block-diagonal form with blocks $I_i^j = I(\theta_i^j)$ of dimension $m \times m, j = 1, 2$. We assume that the estimate $\hat{\theta}_i^j$ is such that for each j uniformly in i:

1. $\lim_{n \to \infty} \max_i |\mathsf{E}[T_i^j]| = 0$.

2. All eigenvalues of the matrices $n\mathsf{E}[(\hat{\theta}_i^j - \theta_i^j)(\hat{\theta}_i^j - \theta_i^j)']$ are bounded from above so that

$$\begin{aligned} & \lim_{n \to \infty} \max_i |n\mathsf{E}[(\hat{\theta}_i^j - \theta_i^j)' I(\theta_i^j)(\hat{\theta}_i^\nu - \theta_i^j)] - m| \\ = \ & \lim_{n \to \infty} \max_i |\mathsf{E}[\langle T_i^j, T_i^j \rangle] - m| = 0, \end{aligned}$$

   where $\langle \bullet, \bullet \rangle$ denotes the scalar product.

3. $\max_i \mathsf{E}[|T_i^j|^3] = \mathcal{O}(\frac{1}{n^{3/2}})$. (6)

4. The asymptotic distribution of T_i^j converges to $\mathcal{N}_m(0, I)$ as n approaches infinity.

These assumptions form the standard set of “good” asymptotic properties, of which first three reflect unbiasedness, efficiency and boundness of the third absolute moment of $\hat{\theta}_i^j$, uniformly in i as $n \to \infty$.

Let us now in this framework analyze classifier $\mathcal{C}(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)$, given the structure $\mathcal{G}$. Since we fix the block-size to the constant m, the total number of blocks, κ, must grow together with n according to (5) in such a way that

$$\lim_{n\to\infty}\frac{\kappa}{n}=\rho, \quad \text{where} \quad 0<\rho<\infty \tag{7}$$

and $\eta=m\rho$. This assumption being designed for augmented BN, is just a particular case of (5).

Further, by the block independence we get

$$\mathcal{C}(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)=\sum_{i=1}^{\kappa}\mathcal{C}_i(\mathbf{x}_i;\hat{\theta}^1_i,\hat{\theta}^2_i), \tag{8}$$

where $\mathcal{C}_i(\mathbf{x}_i;\hat{\theta}^1_i,\hat{\theta}^2_i)=\ell_i(\mathbf{x}_i;\hat{\theta}^1_i)-\ell_i(\mathbf{x}_i;\hat{\theta}^2_i)$, which implies that the classifier induced by augmenting BN is *log additive* in each block and the corresponding procedure is within the frame of the *Generalized Additive Models*; see (Hastie et al. 2001). The main advantage of the additive structure of the augmented classifier is that in the asymptotic framework specified by (7), $\mathcal{C}(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)$ can be viewed as a sum of a growing numbers (κ grows together with n) of independent random variables and, under rather mild regularity conditions imposed on the family of local densities $\mathcal{F}_{\mathcal{G},\Theta}$, we may state the convergence of this sum towards a Gaussian distribution. This methodology has been studied in details in (Pavlenko and von Rosen 2001), where the asymptotic distribution of $\mathcal{C}(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)$ was used to establish the minimum misclassification risk

$$\mathcal{R}_{\mathcal{C}(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)}\longrightarrow\Phi\Big(-\frac{\sqrt{J}}{2}\frac{1}{\sqrt{1+\frac{2m\rho}{J}}}\Big), \tag{9}$$

as $n\longrightarrow\infty$ by (7), where $\Phi(y)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y}\exp(-z^2/2)dz$ and J denotes the cross-entropy distance between the classes defined as

$$J=\int\ln\frac{f^1(\mathbf{x};\theta^1)}{f^2(\mathbf{x};\theta^2)}\Big(f^1(\mathbf{x};\theta^1)-f^2(\mathbf{x};\theta^2)\Big)d\mathbf{x}. \tag{10}$$

3 Separation score and weighted BN classifier

Given the augmented structure $\mathcal{G}$ specified by m and κ, the cross-entropy distance $J^{\mathcal{G}}$ defined by (10) is additive and decomposable as $J^{\mathcal{G}}=\sum_{i=1}^{\kappa}J_i$, where

$$J_i=\int\ln\frac{f_i(\mathbf{x}_i;\theta^1_i)}{f_i(\mathbf{x}_i;\theta^2_i)}(f(\mathbf{x};\theta^1)-f(\mathbf{x};\theta^2))d\mathbf{x} \tag{11}$$

is the input of ith block of nodes towards $J^{\mathcal{G}}$. Our idea is to employ this quantity to evaluate the relevance of the block with respect to (for) classification: We define the *separation score* of the ith block by the value $\frac{nJ_i}{2}$, and its sample based analogue $\frac{n\hat{J}_i}{2}$ with estimated parameters in $f_i(\mathbf{x}_i; \theta_i^j)$. Normalization by n is to ensure that $0 < \frac{nJ_i}{2} < \infty$ as $n \to \infty$ according to (7).

In the growing dimension asymptotics framework it is worthwhile introducing a distribution function of the block scores as

$$H_n(u) = \frac{1}{\kappa} \sum_{i=1}^{\kappa} \mathbf{1}_{[\frac{nJ_i}{2}, \infty)}(u),$$

where $\mathbf{1}_{\{A\}}$ is the indicator function of the set A. We suppose also that the convergence $\lim_{n\to\infty} H_n(u) = H(u)$ takes place uniformly in u and $H(u)$ is a known distribution. Observe that using the distribution $H(u)$ in asymptotics (7) we can conclude that

$$J = \lim_{n\to\infty} J^{\mathcal{G}} = \lim_{n\to\infty} \sum_{i=1}^{\kappa} J_i = 2\rho \int u dH(u), \tag{12}$$

where J is the limiting value of the cross-entropy distance for given $\mathcal{G}$.

To incorporate the block separation strength into classification we specify the weight-function of the ith block by $w_i := w(\frac{n\hat{J}_i}{2})$ where $w_i(u)$ is nonnegative and bounded for $u > 0$ and define the *weighted* BN classifier as

$$\mathcal{C}_w(\mathbf{x}; \hat{\theta}^1, \hat{\theta}^2) = \sum_{i=1}^{\kappa} w_i \mathcal{C}_i(\mathbf{x}_i; \hat{\theta}_i^1, \hat{\theta}_i^2), \tag{13}$$

which provides us with the natural extension of the augmented BN model: each local classifier $\mathcal{C}_i(\mathbf{x}_i; \hat{\theta}_i^1, \hat{\theta}_i^2)$ is weighted by the correspondent block separation score $\frac{n\hat{J}_i}{2}$.

When weighting the network by separation score in practical situations, it is especially important to investigate the asymptotic properties of estimates, $\frac{n\hat{J}_i}{2}$, since we generally can not observe $\frac{nJ_i}{2}$. An impression about the bias induced in the separation score by the plug-in estimative approach for high-dimensional data is given by considering the asymptotic distribution of $\frac{n\hat{J}_i}{2}$, established in (Pavlenko 2003). It is shown that uniformly in i as $n \to \infty$ the distribution of $\frac{n\hat{J}_i}{2}$ converges to the non-central χ^2 distribution, $\chi(u; m, \gamma_i^2)$ with m degrees of freedom and non-centrality parameter γ_i^2, $i = 1, \ldots, \kappa$. This asymptotic result reveals a remarkable property of $\frac{n\hat{J}_i}{2}$ in high-dimensional case: For the non-central χ^2 distribution one can see that

$$\mathsf{E}[\frac{n\hat{J}_i}{2}] = \gamma_i^2 + m + \mathcal{O}(n^{-3/2}), \tag{14}$$

where $\gamma_i^2 = \frac{nJ_i}{2}$, (see, for instance Johnson et al. 1995) which implies that $\frac{n\hat{J}_i}{2}$ *overestimates* the true value of the separation score up to the order m, the block size. Furthermore, the accumulation of the bias over the increasing number of blocks in asymptotics (7) leads to the bias of the classifier of order $\mathcal{O}(\kappa/n)$, which in turn can severely hurt the classification accuracy. To help with this problem, i.e. to take into account the bias induced by plug-in estimation, we derive a down-weighting procedure which can be provided by a properly chosen weight-function w in (13). This function is obtained by minimizing the misclassification risk $\mathcal{R}_w$ over all possible type of weighting assuming that $\mathcal{E}_{1,w} = \mathcal{E}_{2,w}$, i.e. when $\pi_0 = 0$, in which case $\mathcal{R}_w = \Phi(- \frac{D_w}{2})$, where $D_w = \frac{E_w}{\sqrt{V_w}}$,

$$E_w = \rho \int \gamma^2 [\int w(u) \chi(u; m+2, \gamma^2) du] dH(\gamma^2), \tag{15}$$

$$V_w = 2\rho \int [\int u w^2(u) \chi(u; m, \gamma^2) du] dH(\gamma^2). \tag{16}$$

These in turn give the optimal type of weighting as

$$w_0(u) = \frac{\int \gamma^2 \chi(u; m+2, \gamma^2) dH(\gamma^2)}{u \int \chi(u; m, \gamma^2) dH(\gamma^2)}. \tag{17}$$

4 Selection technique

We now extend the formulations to accommodate feature selection. We denote by κ_0 the putative number of irrelevant blocks and assume that for all $i = 1, \ldots, \kappa_0$, $p\lim_{n\to\infty} \frac{n\hat{J}_i}{2} = \gamma_0^2$ and $\lim_{n\to\infty} \frac{\kappa_0}{\kappa} = \psi$, where $p\lim$ means limit in probability and $\psi > 0$ is a fixed small constant. The first assumption is to reflect that the irrelevant blocks suppose to have close sample characteristics, i.e. low sample separation score with the same limit value, γ_0^2 and the second one is to ensure that the number of irrelevant blocks is sufficiently small. In fact, the notion "*number of irrelevant blocks*" is subtle in the growing dimension framework. When reasoning in a usual way, certain number of relevant or irrelevant blocks is a measure of absolute growth rather than relative. On the other hand, it seems unsound to make a finite selection from an infinite number of potential feature nodes. By normalizing κ_0 with the total number of blocks, κ we determine the notion - *fraction of irrelevant features*, denoted by ψ, which suits better the needs of our current investigation.

The method adapted in this paper to incorporating the subset selection step into classification is based on the replacing the weight-function in (13) with its discrete analog of the form $w_i(u) = \mathbf{1}_{[\gamma_0^2,\infty)}(u)$, so that

$$\mathcal{C}(\mathbf{x}; \hat{\theta}^1, \hat{\theta}^2) = \sum_{i=1}^{\kappa} \mathbf{1}_{[\gamma_0^2,\infty)}\Big(\frac{n\hat{J}_i}{2}\Big)\mathcal{C}_i(\mathbf{x}_i; \hat{\theta}_i^1, \hat{\theta}_i^2). \tag{18}$$

The indicator form of $w_i(u)$ can be seen as a special type of weighting and thus works as an *inclusion-exclusion factor* thereby eliminating the blocks whose separation score, $\frac{n\hat{J}_i}{2}$, does not exceed the threshold γ_0^2. Our goal is to determine the optimal subset of $\tilde{\kappa}$ blocks ($\tilde{\kappa} < \kappa$) whose contributions towards classification are essential and using asymptotic results, develop the practically useful selection procedure where the unknown threshold γ_0^2 can be estimated from data.

Since we are looking for the sets of nodes which provide the better classification performance, optimizing the feature selection must be based on minimizing the misclassification risk $\mathcal{R}$. To do this, we first investigate the asymptotic effect of excluding a set of low informative blocks by using $\varepsilon(\mathcal{C}) = \tilde{\mathcal{C}} - \mathcal{C}$, which represents the difference between the classifier $\tilde{\mathcal{C}}$ based on the selected $\tilde{\kappa}$ blocks and the classifier $\mathcal{C}$ where all of the potential κ blocks are used. Further, to relate the difference between $\tilde{\mathcal{C}}$ and $\mathcal{C}$ to the results (9) and (15)-(16) we note that the misclassification risk $\mathcal{R}$ is a function of the first two moments of the the weighted classifier and therefore, to proceed, we need to evaluate $\varepsilon(E_w)$ and $\varepsilon(V_w)$ which is done in the following theorem.

Theorem 1 *If* $\lim_{n\to\infty} \frac{\kappa_0}{\kappa} = \psi \geq 0$ *where* ψ *is small, then with the optimal weighting by (17)*

$$\varepsilon(\rho) = \tilde{\rho} - \rho = -\psi\rho,$$
$$\varepsilon(E_w) = \varepsilon(\rho)[(w(\gamma_0^2)(\gamma_0^2 - m) - 2\gamma_0^2 w'(\gamma_0^2)],$$
$$\varepsilon(V_w) = \varepsilon(\rho)2\gamma_0^2 w^2(\gamma_0^2),$$

and using the weight-function $w(u) = \mathbf{1}_{[\gamma_0^2,\infty)}(u)$

$$\varepsilon(D) = \varepsilon(\rho)\frac{2J_\rho}{(J_\rho + 2\rho m)^2}[\gamma_0^2(J_\rho + 4\rho m) - 2m(J_\rho + 2\rho m)], \tag{19}$$

where $w'(u) = \frac{d}{du}w(u)$, $D = \frac{E_w}{\sqrt{V_w}}|_{w=\mathbf{1}_{[\gamma_0^2,\infty)}(u)}$ *and in the asymptotic framework specified by (7)* $J_\rho = p\lim_{n\to\infty}\sum_{i=1}^{\kappa}\hat{J}_i$.

Due to space consideration we do not represent the detailed proof here.

Now, using the estimation scheme (6) we can specify the selection procedure which minimizes the misclassification risk. Observe that $\mathcal{R}$ being a monotone decreasing function of D can be diminished by the feature selection (18) if and only if $\varepsilon(D) > 0$, i.e. if $\tilde{D} > D$. This implies that in (19) we require that $\gamma_0^2(J_\rho + 4\rho m) - 2m(J_\rho + 2\rho m) < 0$ since $\varepsilon(\rho)$ is negative, and therefore

$$\gamma_0^2 < 2m\frac{J_\rho + 2\rho m}{J_\rho + 4\rho m}, \tag{20}$$

so that the blocks with the limiting separation score lower than $2m\frac{J_\rho+2\rho m}{J_\rho+4\rho m}$ should be excluded.

The practical implementation of the selection technique requires specification of both separation score and selection threshold from the data. Estimation scheme for $\frac{nJ_i}{2}$ is already specified in (6) and one can redly see that the standard type of estimates like, for example maximum likelihood, satisfy the conditions. To evaluate the threshold in practice we relate (20) to the exact form of the asymptotic cross-entropy distance given by

$$p\lim_{n\to\infty} \hat{J}^{\kappa,n} = J + 2\rho m.$$

This results follows straightforwardly from the representation

$$\hat{J}_i = \frac{1}{n}\langle \gamma_i + T_i^1 - T_i^2, \gamma_i + T_i^1 - T_i^2\rangle + \mathcal{O}(\frac{1}{n^2}), \tag{21}$$

where T_i^j are defined in (4) and $\gamma_i^2 = \frac{nJ_i}{2}$ is, in this context the true separation score distributed by $H_n(u)$. (21) is obtained by the standard Taylor series expansion of $f_i(\mathbf{x}_i; \hat{\theta}_i^j)$ about θ_i^j and taking into account the regularity conditions imposed on $\mathcal{F}_{\mathcal{G},\Theta}$; see the details in (Pavlenko 2003). Using the convergence properties of $H_n(u)$ we further write

$$p\lim_{n\to\infty}\sum_{i=1}^{\kappa}\hat{J}_i = 2\lim_{n\to\infty}\frac{\kappa}{n}\int(\gamma^2 + m)dH_n(\gamma^2) = J + 2\rho m,$$

by (7) and (12), where the bias-term $2\rho m$ highlights the effect of high dimensionality.

With these results we establish the selection procedure: to improve the classification accuracy, the ith block should be excluded if

$$\frac{n\hat{J}_i}{2} < 2m\frac{\hat{J}^{\kappa,n}}{\hat{J}^{\kappa,n} + 2\rho m}. \tag{22}$$

To give an impression of how the proposed feature selection technique effects the

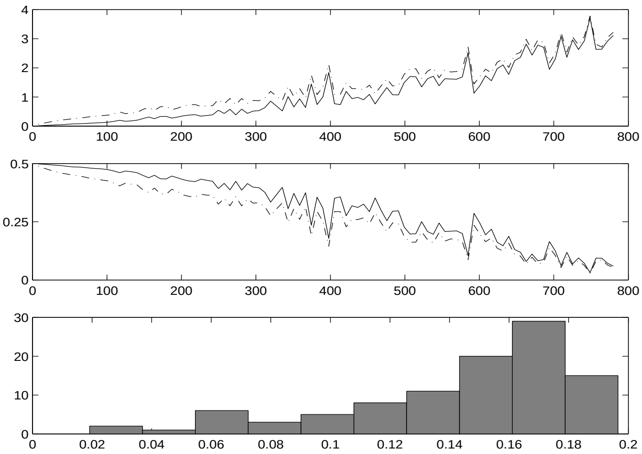

Figure 1: $H(u) = \mathcal{U}_{[0,b]}$ is specified for each $b \in [0; 1.3]$ and for each block. D and $\tilde{D}$ (upper panel) as well $\mathcal{R}$ and $\tilde{\mathcal{R}}$ (middle panel) are plotted as functions of $nb/2$, without selection (solid lines) and with selection (dashed lines), respectively. The bottom panel represents the distribution of $1 - \tilde{\mathcal{R}}/\mathcal{R}$.

classification accuracy, we consider two cases of $H(u)$ specifying our a *priori* knowledge about the block separation strength given $\mathcal{G}$. We first assign the block separation score a uniform distribution, i.e. assume that $H(u) = \mathcal{U}_{[a,b]}(u)$, where a and b are given constants. For the second case, we assume that $H(u) = \chi(u, m, \gamma^2)$, i.e. that the separation strength has the non-central χ^2 distribution with m degrees of freedom and non centrality parameter γ^2. For each case we calculate estimates of the block separation scores, $\frac{n\hat{J}_i}{2}$ and the cross-entropy distance $\hat{J}_n$ using the data produced from the correspondent distribution $H(u)$ and then compare between D and $\tilde{D}$ for different network models represented in terms of $\mathcal{G}$ and m for the training set, and different values of a, b, γ^2 and different training set sizes. Since the proposed selection technique is based on estimates, we have to take into account the high-dimensionality effects when evaluating the classification performance. To do so, we focus on the training sets of the size $n = 1200$ assuming that $\kappa = 100$ and $m = 12$ for both distributions.

Figures 1 and 2 represent the behavior of $\tilde{D}$ and $\tilde{\mathcal{R}}$ (using $\tilde{\kappa}$ selected blocks) as well as D and $\mathcal{R}$ (using all κ blocks), as functions of the sample size normalized by the range of the correspondent distribution $H(u)$; see two upper panels, respectively. Histograms in the bottom panels illustrate the benefit in classification accuracy arising from the feature selection approach for each particular choice of $H(u)$. The selection procedure is running for each sample and since the training examples are

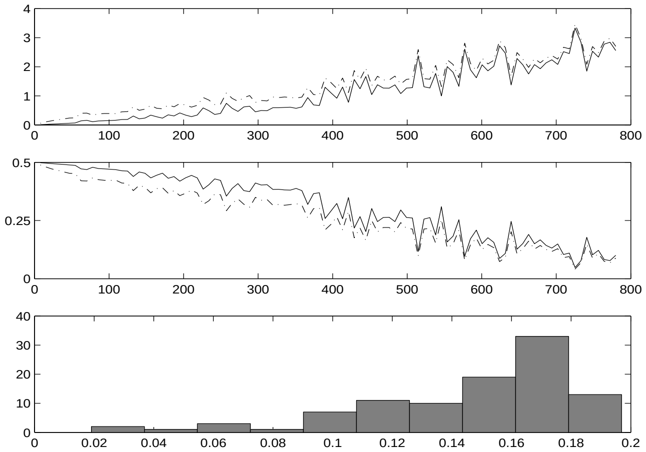

Figure 2: $H(u) = \chi(u; m, \gamma^2)$ is specified for each $u \in [0; 1.3]$ and for each block; m and γ^2 are uniformly drawn from values in $[1; \kappa m]$ and $[0; 10]$, respectively. Otherwise, the plots are produced in the same way as in Figure 1.

different in different trials we expect the number of excluded blocks κ_0 to depend on the trial.

The effect of incorporating the feature selection specified by (22) is clearly seen in terms of more rapid decrease of the asymptotic misclassification risk $\mathcal{R}$ of the selective BN classifier with increasing the sample size for both $\mathcal{U}_{[a,b]}(u)$ and χ^2 distributions,(middle panel in both figures, dashed lines) even if $\frac{\kappa_0}{\kappa}$ is small, i.e. when the small portion of low-informative blocks is excluded. The decrease of the misclassification risk measured as $1-\frac{\tilde{\mathcal{R}}}{\mathcal{R}}$ for the selective BN model versus the model without selection is demonstrated by the histogram in figures 1 and 2 (bottom panel). For both cases of the distribution $H(u)$, the selective BN reveals noticeably (up to 20%) lower asymptotic misclassification rate than that for the classifier without selection.

5 Conclusion

We have presented a theoretically justified subset selection approach which is based on the idea of defining a probabilistic distance measure of the separation score of a set of feature nodes for the augmented BN classifier. Feature selection was developed as an extension of the weighted BN classifier, where weight-functions are viewed as inclusion-exclusion factors. The optimization of the selection procedure was based on minimizing the misclassification risk and is combined with estimating the unknowns for a given network structure $\mathcal{G}$, thereby *jointly* taking into account the high-dimensionality effects. The calculations were shown to be feasible in the context of augmented BN model for different augmenting order m and when classifiers are defined as a discriminant score of the local class-conditional densities satisfying rather mild regularity conditions. The selective BN classifier has shown to achieve a better general performance accuracy in a high-dimensional framework. We have developed an algorithm that approximates our theoretical model and present experimental results which support the contention that the proposed feature selection scheme does substantially improve classification performance of the augmented BN model for different a *priori* assumptions about the block separation properties.

Acknowledgments

This work was supported in part by Swedish Västernorrland County Council under grant FoU: JA-2003-1001 RDC.

References

Cowel, R., Dawid, A.P., Lauritzen, S.L. & Spiegelhalter, D.J. (1999), *Probabilistic Networks and Expert Systems*. New York: Springer.

Friedman, J. (1997), On bias, variance, 0/1 - loss, and curse-of-dimensionality, *Data Mining and Knowledge Discovery* **1**, 55-77.

Friedman, N., Geiger, D. & Goldzmidt, N. (1997), Bayesian network classifiers. *Machine Learning* **29**, 131-163.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions*, Vol. 2, Wiley, New York.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag, New York.

Korb, K.B. & Nicholson, A.E. (2003), *Bayesian Artificial Intelligence*, Chapman and Hall/CRC.

Pavlenko, T. (2003), On Feature Selection, Curse-of-Dimensionality and Error Probability in Discriminant Analysis, *J. of Stat. Planning and Inference* **115**, 565-584.

Pavlenko, T. & von Rosen, D. (2001), Effect of Dimensionality on Discrimination, *Statistics* **35**, 191-213.

Pavlenko, T. & von Rosen, D. (2002), Bayesian network classifiers in a high dimensional framework. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI-02)*, 397-404, San Francisco, CA: Morgan Kaufmann Publishers.