Learning Bayesian Networks for Cytogenetic Image Classification

Boaz Lerner and Roy Malka
Pattern Analysis and Machine Learning Laboratory
Department of Electrical & Computer Engineering
Ben-Gurion University, Beer-Sheva, Israel

Abstract

We experimentally learn structures of Bayesian networks classifying signals enabling genetic abnormality diagnosis. Structures learned based on the naive Bayesian classifier, expert knowledge or using the K2 algorithm are compared. Inferiority of the K2-based classifier has motivated an investigation of the algorithm initial ordering, search procedure and metric. Replacing the K2 search with hill-climbing search improves accuracy as does the inclusion of hidden variables into the structure. However, it is proved experimentally that this inferiority of the K2-based classifier is mainly due to the K2 metric soliciting structures having enhanced representability but limited classification accuracy.

1 Introduction

We experimentally investigate structure learning for Bayesian network classifiers (BNCs) through a case study in genetic abnormality diagnosis. By classifying fluorescence in-situ hybridization (FISH) signals in cytogenetic images, we enable efficient diagnosis of abnormalities, such as Down and Patau syndromes. A previous study [6] has demonstrated simplicity and accuracy in FISH signal classification of the naive Bayesian classifier (NBC) but not the expected interpretability. Another study [8] allowed the BNC to capture dependencies in the domain, thereby to alleviate the NBC independence restriction which may weaken interpretability and performance. The unrestricted BNC was constructed using expert knowledge or learned from the data using the K2 algorithm [2] but with inferior accuracy in the latter case. The motivation in this research is to experimentally explore ways to improve the accuracy of the search-and-score K2-based BNC and for a several reasons. First is the aforementioned inferiority demonstrating that the common operation of the K2-based classifier is less than optimal. Second is the general inferiority for this domain of the BNC to a neural network (NN) [7] hinting that the BNC has not exploited the full information hidden within the data. Third is the lack of robust expert knowledge about the domain preventing the construction of an alternative accurate expert-based BNC. Finally is the advantages of the BNC over the NN with respect to representability and interpretability which we do not want to waive.

We investigate three flaws of the K2 algorithm in classification, namely the requirement to determine an initial ordering, the algorithm search procedure and its score. By ranking the domain features based on the separability or classification accuracy they provide, we establish data-driven and classification-oriented initial orderings for the K2 algorithm. By replacing the K2 search with hill-climbing search (HCS), we dispense with the requirement for an initial ordering and also enable the exploration of larger structure spaces and the incorporation of prior knowledge to the search. By including hidden variables manifesting causal relations between variables not straightforwardly evident a priori, we extend the ability of the structure in representing the cytogenetic domain. Finally, we manifest the inherent limitation of the K2 score in classification.

Section 2 presents the BN model and methods for learning the model, while Section 3 demonstrates experimentally the capabilities of these methods in FISH signal classification. Finally, Section 4 summarizes the work.

2 Learning a BN

A BN model $\mathcal{B}$ for a set of $n$ random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, having each a finite set of mutually exclusive states, consists of two main components, $\mathcal{B} = (\mathcal{G}, \theta)$. The first is a structure $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ that is a directed acyclic graph (DAG) containing no directed cycles. $\mathbf{V}$ is a finite set of nodes of $\mathcal{G}$ corresponding to the variables of $\mathbf{X}$, and $\mathbf{E}$ is a finite set of directed edges of $\mathcal{G}$. $\mathbf{E}$ represents the dependence relations modeled by $\mathcal{G}$, where the presence or absence of an edge in $\mathbf{E}$ demonstrates dependence or independence, respectively, of two variables of $\mathbf{X}$ corresponding to the nodes being connected or not by that edge. The second component is a set of parameters, $\theta$, that specify the con-
ditional probability distributions that quantify graph edges. The probability distribution of each \( X_i \in \mathbf{X} \) conditioned on its parents in the graph \( \mathbf{Pa}_i \in \mathbf{X} \) is \( P(X_i | \mathbf{Pa}_i) \in \theta \).

The joint probability distribution for \( \mathbf{X} \) given a structure \( \mathcal{G} \) that is assumed to encode this probability, is \([4, 9]\)

\[
P(\mathbf{X} = \mathbf{x} | \mathcal{G}) = \prod_{i=1}^{n} P(X_i = x_i | \mathbf{Pa}_i = \mathbf{pa}_i, \mathcal{G})
\]

(1)

where \( \mathbf{x} \) is the assignment of states to all variables, \( \mathbf{pa}_i \) the assignment of states to \( X_i \)'s parents \( \mathbf{Pa}_i \) and the terms in the product compose the set of parameters \( \theta \) quantifying the dependence relations modeled by \( \mathcal{G} \). In classification, and without limiting the generality, we may identify the class variable with the first variable \( X_1 = \mathbf{C} \) and \( \mathbf{X} \setminus \mathbf{C} \) with the sets of nodes and parents of \( X_i \) excluding \( \mathbf{C} \), respectively, in order to apply (1) to classification,

\[
P(\mathbf{C} | \mathbf{X} \setminus \mathbf{C}, \mathcal{G}) = \prod_{i=2}^{n} P(X_i | (\mathbf{C}, \mathbf{Pa}_i \setminus \mathbf{C}), \mathcal{G}) \frac{P(\mathbf{C} | \mathcal{G})}{P(\mathbf{X} \setminus \mathbf{C} | \mathcal{G})}.
\]

(2)

Indeed, BNs that were originally employed in knowledge representation and probabilistic inference have recently been applied to classification. The computation of the joint probability distribution and any related probability (e.g., the posterior probability) is conditioned on the graph. Therefore, we must first obtain a structure and estimate its parameters in order to represent the domain accurately.

**Structure Learning** Constructing the BN structure is usually performed using expert knowledge to identify variables in, and extracting dependencies and independencies from, the domain. But, the expert-based structure may be different from expert to expert, difficult to obtain and time-consuming to be manually constructed. Hence, there are advantages in learning a BN structure from the data.

One common approach to structure learning is score-based \([2, 4]\). It comprises of a search procedure for a structure and a score evaluating each structure. An exhaustive search in which every possible DAG is scored is limited to small structures. A sub-optimal hill-climbing search method is the K2 algorithm \([2]\). The algorithm is initialized using node ordering and assuming a node has no parents. It adds incrementally that parent from the ordering whose addition increases the score of the resulting structure the most. The algorithm stops adding parents to a node when the addition of no single parent can increase the score.

The score used in the K2 algorithm is the K2 metric \([2]\):

\[
P(\mathcal{G}, D) = P(\mathcal{G}) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_j - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!
\]

(3)

where \( D \) is a data set of instances, \( P(\mathcal{G}) \) is the structure prior probability and \( q_i \) and \( r_i \) are the numbers of parent states and mutual exclusive states of the \( i \)th node, respectively. \( N_{ijk} \) is the number of instances of the \( i \)th node being in the \( k \)th state when its parents are in their \( j \)th state and \( N_{ijk} = \sum_{k=1}^{r_i} N_{ijk} \). This score is decomposable since it is assumed that the parameters associated with each variable are mutually independent. Thus, the score can be written as a product of independent sub-scores one for each variable.

**Parameter Learning** Equation 1 summarizes the joint probability over the graph as a product of local probability distributions one for each node. We model each such distribution using the unrestricted multinomial model, where the distribution parameters are estimated using the maximum-likelihood solution which is the sample frequency \([2, 4]\).

**Inference** In inference we calculate the conditional probability distribution of a subset of the nodes in the graph (‘hidden’ nodes) given another subset of the nodes (‘observed’ nodes). Regarding classification, a hidden node represents the class variable, the observed nodes represent the features and inference is performed using (2). For exact inference, we use the well-known junction tree algorithm \([5]\).

### 3 Experimentation and results

Real signals and artifacts of red and green fluorophores manifesting Down and Patau syndromes, respectively, are classified in FISH images. FISH preparation and image acquisition and analysis are described in \([7]\). Twelve features are measured for the signals. These include Area (1), Eccentricity (2) and a number of spectral features. We compute at the specific color plane three RGB (red-green-blue) intensity-based features: the Total (3) and Average (4) Channel Intensities and the Channel Intensity Standard Deviation (5). We also compute four HSI (hue-saturation-intensity) features: Maximum Hue (6), Average Hue (7), Hue Standard Deviation (8) and Delta Hue (9). Delta Hue is the difference between the Maximum and Average Hue normalized by the Average Hue. Additional features are the two coordinates of the eigenvector (10 and 11) corresponding to the largest eigenvalue of the red and green intensity components of the signal. Finally is the Average Gray Intensity (12), i.e., average intensity over the three color channels. These features are represented by the observable nodes of the BN, whereas the (hidden) class variable takes up four states associated with the possible classes.

First, we construct the NBC having the class node the only parent of all other nodes (Figure 1). Then, we construct a structure using expert knowledge in order to improve the NBC by adding necessary, and removing unnecessary, edges (Figure 2). Third, we apply the K2 algorithm to search for the structure maximizing the K2 metric (3) using an initial ordering coinciding with expert knowledge (Figure 3). We show in Table 1 the (10-fold cross-validation)
accuracy of structures derived using these three methods in classifying FISH signals. The class node Markov blanket (including the node parents, children and children co-parents [9]) in the expert structure separates this node from all nodes that are not its children. That is, variables (9) and (12) do not participate in the classification. Since the relevancy of these features to FISH signal classification is low [6, 7], the accuracy of the expert-based classifier requiring the estimation of fewer parameters is superior to that of the NBC. The expert-based classifier is also superior to the K2-based classifier since the structure of the latter avoids several features relevant to classification. The inferiority of the K2-based classifier is further studied in relation to the algorithm initial node ordering, search procedure and score.

We compare random initial orderings to others based on expert knowledge or feature rankings using scatter criterion $J_B$ [3] and the multilayer perceptron NN [1] classification accuracy. The data-driven orderings are useful when prior knowledge about the ordering is missing or not robust enough. Since no substantial differences between structures and the resultant classification accuracies are shown, we conclude that the initial ordering for the FISH domain has only minor effect on the K2-based structure.

Employing HCS and the NBC as an initial structure, the resulted structure (Figure 4) improves representation, increases the K2 metric and yields enhanced accuracy (Table 1). Also, HCS needs no initial ordering, enables the study of larger structure spaces and the incorporation of prior knowledge through the initial structure.

Including hidden nodes in a BN enriches the model and extends its interpretability. Hidden nodes may reduce the number of edges and thus parameters needed to be learned thereby eliminating the curse-of-dimensionality and time of learning. By incorporating into an expert-based structure color and intensity hidden nodes acting as parents of all color and intensity nodes, respectively, we accommodate expert belief that all color and intensity features are the result of two separate sources (Figure 5). Table 1 shows the classification accuracy when using these two hiddens.

Finally, in order to evaluate the influence of the K2 metric, we plot in Figure 6 the average K2 metric and classification accuracy for increasing steps of the HCS initialized by the expert-based structure. The initial structure has a low value of the K2 metric but high accuracy (Table 1). In the first step, edge 6 → 7, reflecting the high correlation between the variables [7], is added (Figure 4) as it contributes the most to the expert structure. However, adding a highly correlated variable cannot enhance the classification accuracy but to reduce it (Figure 6) as more parameters have to be estimated. Until the fifth step, edges increasing the K2 metric are added and lead to improvement of the accuracy. However, from the sixth step, the K2 metric continuous to increase as additional correlated variables are added but the accuracy declines monotonically since these edges provide no added value to the classifier. Besides not contributing to classification, these edges also raise the number of parameters to estimate leading to the curse-of-dimensionality and accuracy deterioration. This demonstrates experimentally for our domain that the K2 metric is not an appropriate criterion for BNC structure learning. A structure learned using the K2 algorithm may represent the dependencies and independencies within the domain precisely but is not necessarily the basis of an accurate classifier.

4 Discussion

We experimentally learned structures based on the NBC, expert knowledge and the data-driven K2 algorithm for BNCs classifying FISH signals for the diagnosis of genetic abnormalities. Inferiority of the K2-based classifier has motivated experimental exploration of its initial node ordering, search procedure and score. We have found that (a) the K2 algorithm is almost insensitive to its initial ordering (at least for this case study), (b) an NBC-initialized hill-climbing-searched structure as well as the inclusion of hidden nodes in the structure improve FISH signal classification accuracy, and (c) structures selected according to the K2 metric represent the domain well but cannot provide accurate classifiers.

Acknowledgment: This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University, Beer-Sheva, Israel.

References

Figure 1. The NBC. Numbers represent features described in Section 3.

Figure 2. The expert-based structure.

Figure 3. The K2-based structure.

Table 1. Classification accuracy (mean and std) using the structures in Figures (1)-(5).

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBC (1)</td>
<td>78.0 (2.1)</td>
</tr>
<tr>
<td>Expert-based BNC (2)</td>
<td>79.5 (2.1)</td>
</tr>
<tr>
<td>K2-based BNC (3)</td>
<td>78.0 (2.1)</td>
</tr>
<tr>
<td>NBC + HCS + K2 metric (4)</td>
<td>80.1 (2.0)</td>
</tr>
<tr>
<td>Expert structure + hiddens (5)</td>
<td>80.9 (2.3)</td>
</tr>
</tbody>
</table>

Figure 4. A structure learned using the K2 metric and HCS starting from the NBC.

Figure 5. Expert structure with color and intensity hidden nodes.

Figure 6. K2 metric and accuracy of a structure learned from the expert structure by HCS.