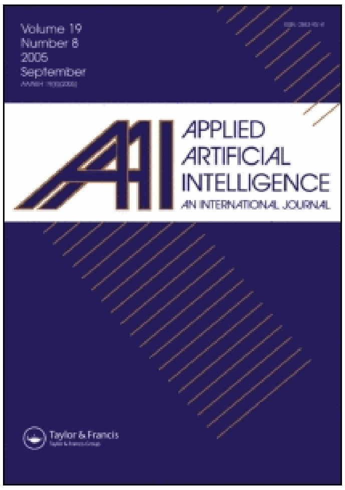

Applied Artificial Intelligence
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713191765

INVESTIGATION OF THE K2 ALGORITHM IN LEARNING BAYESIAN NETWORK CLASSIFIERS

Boaz Lerner[a]; Roy Malka*[a]
[a] Ben-Gurion University, Beer-Sheva, Israel

Online publication date: 22 January 2011

To cite this Article Lerner, Boaz and Malka*, Roy(2011) 'INVESTIGATION OF THE K2 ALGORITHM IN LEARNING BAYESIAN NETWORK CLASSIFIERS', Applied Artificial Intelligence, 25: 1, 74 — 96
To link to this Article: DOI: 10.1080/08839514.2011.529265
URL: http://dx.doi.org/10.1080/08839514.2011.529265

Applied Artificial Intelligence, 25:74–96, 2011
Copyright © 2011 Taylor & Francis Group, LLC
ISSN: 0883-9514 print/1087-6545 online
DOI: 10.1080/08839514.2011.529265

INVESTIGATION OF THE K2 ALGORITHM IN LEARNING BAYESIAN NETWORK CLASSIFIERS

Boaz Lerner and Roy Malka[*]
Ben-Gurion University, Beer-Sheva, Israel

□ *We experimentally study the K2 algorithm in learning a Bayesian network (BN) classifier for image detection of cytogenetic abnormalities. Starting from an initial BN structure, the K2 algorithm searches the BN structure space and selects the structure maximizing the K2 metric. To improve the accuracy of the K2-based BN classifier, we investigate the K2 algorithm initial ordering, search procedure, and metric. We find that BN structures learned using random initial orderings, orderings based on expert knowledge, or a scatter criterion are comparable and lead to similar classification accuracies. Replacing the K2 search with hill-climbing search improves the accuracy as does the inclusion of hidden nodes in the BN structure. Also, we demonstrate that though the maximization of the K2 metric solicits structures providing improved inference, these structures contribute to only limited classification accuracy.*

INTRODUCTION

Fluorescence in situ hybridization (FISH) offers numerous advantages compared with conventional cytogenetic techniques because it allows chromosome abnormalities to be detected during normal cell interphase (Nath and Johnson 2000). One of the most important applications of FISH for the detection of numerical abnormalities, such as Down and Patau syndromes, is dot counting, that is, the enumeration of signals (dots) within the nuclei, as the dots in the image represent the inspected DNA sequences. Manual dot counting is a time-consuming, laborious, and tedious procedure—hence the need in automation.

It was proposed (Lerner et al. 2001; Lerner, Koushnir, and Yeshaya 2007; Lerner, Yeshaya, and Koushnir 2007) to base FISH dot counting on

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

[*]Current address Roy Malka, Department of Computer Science & Applied Mathematics, Weizmann, Institute of Science, Rehovot 76100, Israel.

Address correspondence to Boaz Lerner, Department of Industrial Engineering and Management, Ben-Gurion University, Beer-Sheva 84105, Israel. E-mail: boaz@bgu.ac.il

a classifier discriminating between valid (real) signals and artifacts, thereby allowing the enumeration of only real FISH signals and the automation of dot counting for genetic diagnosis of numerical abnormalities. We focused our previous efforts to accomplish this task on learning Bayesian network classifiers (BNCs) (Lerner 2004; Malka and Lerner 2004). One study (Lerner 2004) demonstrated simplicity and accuracy of the naive Bayesian classifier (NBC) in FISH signal classification, however not the expected domain interpretability.[1] To alleviate the NBC independence restriction that may have weakened interpretability and accuracy, we allowed (Malka and Lerner 2004) the BNC to capture dependencies in the domain. The unrestricted BNC was constructed by using expert knowledge or learned from the data using the K2 algorithm (Cooper and Herskovits 1992), in the latter case, however, with inferior accuracy.

The motivation for the current research is to experimentally explore ways to improve the accuracy of the learned-form-data K2-based BNC, for several reasons. First is the aforementioned inferiority to the expert-based BNC, demonstrating that the common use of the K2 learning algorithm in classification is less than optimal. Second is the inferiority—at least for this domain—of the BNC to a neural network (NN) performing the same task (Lerner et al. 2001), hinting that the BNC has not exploited the full information hidden in the domain. Third is the lack of robust prior (expert) knowledge about the domain, preventing the construction of a highly accurate expert-based BNC. Finally, there are advantages to the BNC over NN and other classifiers with respect to representability and interpretability of the cytogenetic domain, which we do not want to waive.

To improve the accuracy of a BNC learned using the K2 algorithm, we investigate some aspects of the algorithm, such as its dependence on an initial topological ordering, search procedure, and score. By ranking the domain features based on their degree of separability, we establish an expert-free initial ordering for the K2 algorithm that is both data-driven and classification-oriented. By replacing the K2 search by a hill-climbing search, we enable the exploration of larger structure spaces and incorporation of prior knowledge to the search and also dispense with the requirement for an initial ordering. In addition, we demonstrate the limitation of using the K2 score (metric) in structure learning for classification tasks. Finally, 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. All these aspects are experimentally studied here as detailed in the Experiment and Results section and outlined schematically in Figure 1.

There are several contributions to this study. First is the application of unrestricted BNCs to enhance the accuracy of cytogenetic image classification. We focus on studying and optimizing the BNC to FISH dot

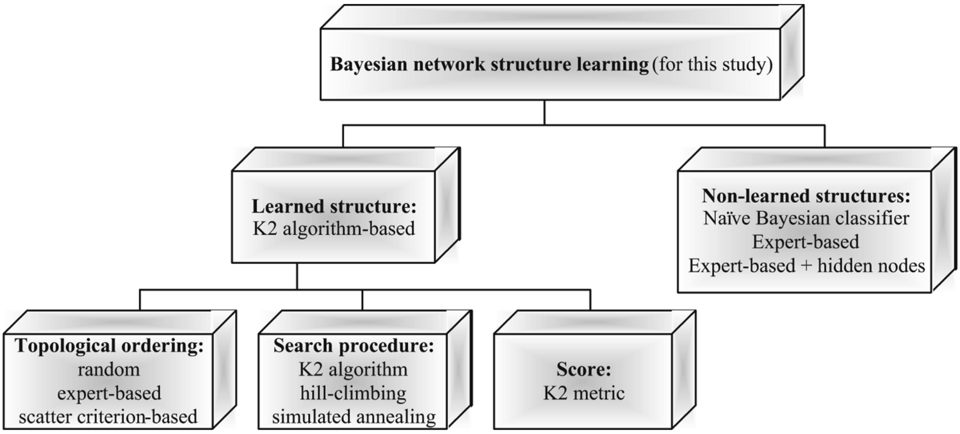

FIGURE 1 Aspects and settings that are experimentally investigated in the study. Note that "learning" here is with respect to structure learning, and we defer exploring parameter learning to another study.

counting and defer the actual implementation of dot counting to another study. Second is the detailed comparison for this problem of expert-based and learned-form-data structures of BNCs. Third is the extensive experimental investigation of the K2 algorithm and the evaluation of different aspects of the algorithm for this domain. Finally is the conclusion, drawn experimentally for this real-world cytogenetic problem, that structures learned by maximizing the K2 metric may excel in general inference problems but do not necessarily yield the most accurate classifiers. To the best of our knowledge this is the first empirical evidence to this conclusion in real–world applications, and to our understanding this conclusion is not restricted only to the cytogenetic domain.

The Bayesian Network Learning and Inference section of the article introduces the Bayesian network (BN), strategies for learning the BN structure and parameters, and inference using the BN. The FISH Signal Representation and Classification section demonstrates the cytogenetic domain, elaborating on FISH signal representation and classification. The Experiments and Results section investigates experimentally learning a BN structure using the K2 algorithm for FISH signal classification. The section concentrates on the K2 algorithm initialization, search procedure, and metric, and in using hidden variables for learning structures. Finally, conclusions for the study and outline of future research are given in the Discussion section.

BAYESIAN NETWORK LEARNING AND INFERENCE

BNs are probabilistic graphical models that provide interpretability of the explored domain by extracting and manifesting dependences, independences, and causal relationships among variables representing the domain. In addition, the models readily combine knowledge acquired from the data with prior information. Using the graph, the joint probability distribution over the variables can be decomposed, rendering probabilistic inference a simple task.

Introduction to Bayesian Networks

A BN model $\mathscr{B}$ for a set of n variables $\boldsymbol{X}=\{X_1, X_2, \ldots, X_n\}$ each having a finite set of mutually exclusive states consists of two main components, $\mathcal{B} = (\mathcal{G}, \boldsymbol{\theta})$. The first component $\mathcal{G}$ is a structure that is a directed acyclic graph (DAG) because it contains no directed cycles. The nodes of $\mathcal{G}$ correspond to the variables of $\boldsymbol{X}$, and thus a variable and its corresponding node are usually referred interchangeably. An edge connecting two nodes in $\mathcal{G}$ manifests the existence of direct causal influence between the corresponding variables, and the lack of a possible edge in $\mathcal{G}$ represents conditional independence (d-separation) between the corresponding variables.

The second component of BN is a set of parameters, $\boldsymbol{\theta}$, that specify all the conditional probability distributions (or densities) that quantify graph edges. The probability distribution of each $X_i, \in \boldsymbol{X}$ conditioned on its parents in the graph $\mathbf{Pa}_i \subset \boldsymbol{X}$ is $P(X_i|\mathbf{Pa}_i) \in \boldsymbol{\theta}$, where we use X_i and $\mathbf{Pa}_i$ to denote a node and its parent set, respectively.

The joint probability distribution for $\boldsymbol{X}$ given a structure $\mathcal{G}$ that is assumed to encode this distribution is given using the set of parameters by (Cooper and Herskovits 1992; Heckerman 1995; Pearl 1988):

$$P(\boldsymbol{X}|\mathcal{G}) = \prod_{i=1}^{n} P(X_i|\mathbf{Pa}_i, \mathcal{G}) \tag{1}$$

The computation of the joint probability distribution and any probability distribution related to the joint (e.g., the posterior probability) is conditioned on the structure. Therefore, we first learn a structure and then estimate its parameters. Once a structure is learned, parameter learning is usually straightforward (see the Learning the BN Parameters section), so usually most of our efforts are concentrated on structure learning. We note that the theory of BNs is well established (Cooper and Herskovits 1992; Heckerman 1995; Pearl 1988), several applications of BNs have been suggested (Luo and Boutell 2005; Pena, Lozano, and Larranaga 1999; Zhang and Ji 2005), and methods of structure and parameter learning are very central to BN research (Cheng, Bell, and Liu 1997; Cooper and Herskovits 1992; Friedman, Geiger, and Goldszmidt 1997; Heckerman 1995; Heckerman, Geiger, and Chickering 1995; Keogh and Pazzani 2002; Pazzani 1996; Pearl and Verma 1991; Spirtes, Glymour, and Scheines 2000; Yehezkel and Lerner 2009).

BNs that were originally used in knowledge representation and general probabilistic inference have recently been applied also to classification (Friedman, Geiger, and Goldszmidt 1997; Greiner et al. 2005; Grossman and Domingos 2004; Gurwicz and Lerner 2006; Kontkanen et al. 1999; Yehezkel and Lerner 2009). Without limiting the generality, we identify the class variable with the first variable $X_1 = C$ and define $\boldsymbol{X}\backslash C$ and $\mathbf{Pa}_i\backslash C$ as the sets of graph nodes and parents of X_i excluding C, respectively, to apply Eq. (1) to classification,

$$P(C|\boldsymbol{X}\backslash C,\mathcal{G}) = \prod_{i=2}^{n} P(X_i|(C,\mathbf{Pa}_i\backslash C),\mathcal{G})\frac{P(C|\mathcal{G})}{P(\boldsymbol{X}\backslash C|\mathcal{G})} \tag{2}$$

Equation (2) assumes a model in which no edges are pointed onto C, which is common and beneficial to BNCs (see also the K2 Initial Ordering section 4.1). To perform probabilistic inference [as in Eqs. (1) and (2)], we first obtain a structure from expert knowledge or learn it from the data and then estimate the corresponding parameters.

Learning the BN Structure

Expert-Based Structure

Until very recently the common approach for constructing a BN structure relied on identifying variables in and extracting dependencies and independencies from the problem domain using expert knowledge. This expert-based approach is probably the most intuitive way to construct a structure. However, although straightforward in principle, the expert-based structure may be different from expert to expert and difficult to obtain in the absence of an expert or because expert knowledge is usually only implicit to the designer. This structure may be inaccurate for data representing a slightly modified environment and also time-consuming to construct, even when the domain is known, because knowledge should be collected and used manually. Thus, expert-based BNs are usually limited to small, known domains.

Learned-From-Data Structure

There are advantages in learning a BN structure directly from data, especially when the domain is large or reliable expert knowledge is unavailable. A BN learned from data may facilitate and expedite the construction of an expert-based BN by providing an initial structure that is further modified using expert knowledge. Also, the learned structure can be used to judge between structures derived using competing or disagreeing experts. Standing for itself, the learned-from-data structure can be used to gain insights about dependence relations within the domain and to perform causal or probabilistic inference and thereby also for decision-making.

Learning a BN structure is usually accomplished by constraint-based (Cheng, Bell, and Liu 1997; Pearl and Verma 1991; Spirtes, Glymour, and Scheines 2000; Yehezkel and Lerner 2009) or search-and-score (S&S) (Cooper and Herskovits 1992; Heckerman 1995; Heckerman, Geiger, and Chickering 1995; Keogh and Pazzani 2002) methods. Constraint-based methods use statistical tests such as chi-squared or mutual information to find conditional independence relationships among the variables and use these relationships and causality-driven orientation rules (Pearl and Verma 1991) in constructing the BN (see the IC [Pearl and Verma 1991], PC [Spirtes, Glymour, and Scheines 2000], TPDA [Cheng, Bell, and Liu 1997], or RAI [Yehezkel and Lerner 2009] algorithms for details). This work does not deal with constraint-based algorithms.

S&S methods comprise two elements: a search procedure for a network structure and a score (metric) evaluating each structure found in the search. In the brute-force (exhaustive) search approach, every possible DAG is scored. This approach provides a "gold standard" in comparing search algorithms but is limited to structures with small numbers of nodes ($n \leq 5$) because the number of possible structures grows more than exponentially with the number of structure nodes (Cooper and Herskovits 1992). Without limiting the number of parents each node may have to one, learning a structure is NP-hard, although it can be accomplished using heuristic search algorithms (e.g., greedy search) (Heckerman, Geiger, and Chickering 1995).

Search can be performed by starting from a specific point (structure) in space (randomly chosen or based on prior knowledge) and considering all neighboring structures obtained from the current structure by adding, deleting, or reversing a single edge at every iteration of the search algorithm. The search progresses to the neighboring structure having the highest value of a score if this value is higher than that of the current structure. This procedure—called hill-climbing search (HCS)—stops when reaching a local maximum. One method of escaping a local maximum is a greedy search with random restarts, that is, random perturbation of the structure whenever getting stuck at a local maximum. Other approaches for escaping local maxima, such as simulated annealing (Kirkpatrick, Gelatt, and Vecchi 1983) and best-first search, are described in Heckerman (1995). Alternatively, if we knew a total ordering on the nodes, finding the optimal structure would be equivalent to the selection of the best set of parents for each node independently. This is the idea behind the heuristic K2 algorithm.

The K2 Learning Algorithm

The S&S K2 algorithm (Cooper and Herskovits 1992) uses a greedy search and may impose no restriction on the number of parents a node has. The K2 search begins by assuming that a node (representing a discrete variable) has no parents and then adds incrementally that parent from a given ordering whose addition increases the score of the resulting structure the most. We stop adding parents to the node when the score stops to increase.

A common scoring metric is the Bayesian score that is, in principle, the posterior probability of a structure $\mathcal{G}$ given a random sample $D = \{d_1, d_2, \ldots, d_N\}$ from the joint distribution of $\boldsymbol{X}$,

$$P(\mathcal{G}|D) = \frac{P(D|\mathcal{G})P(\mathcal{G})}{P(D)} = \frac{P(\mathcal{G}, D)}{P(D)}$$

Because $P(D)$ does not depend on the structure, we may use the marginal likelihood $P(D|\mathcal{G})$ (Heckerman 1995) or joint probability (also called relative posterior probability) $P(\mathcal{G}, D)$ (Cooper and Herskovits 1992; Heckerman, Geiger, and Chickering 1995) as scoring metrics. The joint probability is the Bayesian Dirichlet (BD) metric (Heckerman, Geiger, and Chickering 1995):

$$P(\mathcal{G}, D) = P(\mathcal{G}) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

where $P(\mathcal{G})$ is the structure prior probability that is constant for each $\mathcal{G}$. n, q_i, r_i, and N_{ijk} are, respectively, the numbers of nodes in the graph, configurations (states) of the parents of the ith node, mutual exclusive states of the ith node, and instances of the ith node being in the kth state when its parents are in their jth configuration. $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. The hyper-parameters of the Dirichlet distribution, $\alpha_{ijk} > 0$, correspond to the a priori probability distribution of X_i taking on its kth state while its parents are in their jth configuration. $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$. The gamma function satisfies $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1) = 1$.

Assigning values to α_{ijk} $\forall i,j,k$ is infeasible, and by the uninformative assignment $\alpha_{ijk} = 1$ $\forall i,j,k$ we turn the BD metric to the simple K2 metric (Cooper and Herskovits 1992; Heckerman, Geiger, and Chickering 1995):

$$P(\mathcal{G}, D) = P(\mathcal{G}) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \qquad (3)$$

Assuming the parameters associated with each variable are mutually independent, the K2 metric is decomposable. That is, the metric can be written as a product of independent subscores, $g(X_i, \mathbf{Pa}_i)$, one for each variable and its set of parents (measuring the degree of dependence between the variable and its parents), in the form

$$P(\mathcal{G}, D) = P(\mathcal{G}) \prod_{i=1}^{n} g(X_i, \mathbf{Pa}_i)$$

The K2 algorithm finds the structure that maximizes each factor (subscore). This maximization is achieved because a node X_j is added to X_i's parent set $\mathbf{Pa}_i$ if following the addition $g(X_i, \mathbf{Pa}_i) > g(X_i, \emptyset)$ for an empty parent set or $g(X_i, \mathbf{Pa}_i) > g(X_i, \mathbf{Pa}_i \backslash X_j)$ for a nonempty parent set. Maximizing each factor also maximizes their product, which is the K2 metric [Eq. (3)].

Interestingly, by assigning $\alpha_{ijk} = 1\ \forall i, j, k$ the K2 metric prefers simpler structures (Borgelt and Kruse 2001). That is, assuming a uniform prior distribution, all possible parents to be included in a variable parent set have equal probabilities. Thus, the number of possible parents is higher than if a non-uniform prior would have restricted some of the parents. As more parents are considered, less instances (N_{ijk}) can affect the calculation of $g(X_i\ \mathbf{Pa}_i)$, so the dependence of a variable on a possible parent is reduced. This is demonstrated in the rejection of possible parents from inclusion within the parent set that leads to simple structures (having fewer edges), as is exemplified experimentally in section 4. A method of controlling this tendency of the metric to select simpler structures is suggested in Borgelt and Kruse (2001). On the other hand, a similar metric to BD—called BDe (Heckerman, Geiger, and Chickering 1995)—encourages more complex structures (Borgelt and Kruse 2001).

Inclusion of Hidden Nodes

Most often, the inclusion of hidden nodes into a BN yields a richer and more interpretable model than that without these nodes. Hidden nodes may reduce the number of edges, and thus parameters needed to be learned, and thereby diminish the curse-of-dimensionality and time of learning. If the existence of a hidden variable and its relations to other variables is known, we can introduce it to an expert-based network and use the incomplete data and the EM algorithm (Dempster, Laird, and Rubin 1977) to estimate the sufficient statistics defining the local conditional probability distributions (Ghahramani and Jordan 1994; Heckerman 1995). Methods of automatic discovery of hidden variables exist as well (Elidan et al. 2001; Friedman 1997; Silva et al. 2006; Spirtes, Glymour, and Scheines 2000), constructing structures usually having representability and scores higher than those of their counterpart structures having no hidden variables.

Learning the BN Parameters

Equation (1) summarizes the joint probability over the graph as a product of local probability distributions (densities), one for each node (variable) conditioned on its parents. In the cytogenetic domain, all except one of the variables are continuous (see the FISH Signal Representation and Classification section); hence, we quantize the variables, as required by the K2 algorithm, and estimate the distributions using the relative frequencies in the data (Malka and Lerner 2004) (i.e., the maximum likelihood solution [Heckerman 1995]).

Inference

Inference in BNs is the task of calculating the conditional probability distribution of a subset of the nodes in the graph (the "hidden" nodes[2]) given another subset of the nodes (the "observed" nodes). In a classification problem a hidden node represents the class variable, the observed nodes represent the features, and inference is conducted using Eq. (2). We use for inference the junction tree algorithm (Huang and Darwiche 1994; Lauritzen and Spiegelhalter 1988), though other methods (Pearl 1988) may do as well (yet not exact for the nontree structures used here).

FISH SIGNAL REPRESENTATION AND CLASSIFICATION

FISH data preparation and image analysis were described thoroughly in Lerner et al. (2001) and hence are avoided here. Red and green signals, corresponding to Down and Patau syndromes, respectively, were extracted from 400 images collected from five slides. After nuclei segmentation the system identified 944 objects within these images as nuclei, of which 613 also contained signals (the remaining 331 objects were unfocused nuclei that therefore contained no signals). After signal segmentation, 3144 objects within the nuclei were identified as signals. Based on labels provided by expert inspection, 1145 of the signals were considered as "reals" (among them 551 were red) and 1999 as "artifacts" (among them 1224 were red). Aiming at the discrimination between real and artifact signals of the two syndromes, we establish a four-class classification problem.

Twelve features were measured to represent the signals to the classifier. The features are as follows (Lerner et al. 2001):

(1) area,
(2) eccentricity (a shape feature measuring the signal similarity to an ellipse), and a number of spectral features. We computed at the specific color plane three RGB (red-green-blue) intensity-based measurement:

(3) total and
(4) average channel intensities and
(5) channel intensity standard deviation. We also computed four HSI (hue-saturation-intensity) hue-based measurements:
(6) maximum hue,
(7) average hue,
(8) hue standard deviation, and
(9) delta hue. Delta hue is the difference between the maximum and average hue normalized by the average hue. Two additional features
(10 and 11) are the coordinates of the eigenvector corresponding to the largest eigenvalue of the red and green intensity components of the signal. The last feature is
(12) average gray intensity, that is, average intensity over the three color channels.

EXPERIMENTS AND RESULTS

In all experiments the BN structure is composed of nodes for the observable variables, representing the features of section 3, and the class node (variable), taking on four states associated with the possible classes determined for real or artifact signals of Down or Patau syndromes. We concentrate on structure learning and simplify parameter learning for the learned structure by using the maximum likelihood solution (see the Learning the BN Parameters section). All experiments to evaluate classifier accuracy are held using 10-fold cross-validation (CV10), and the BN implementation is aided by the Bayes net toolbox (BNT) (Murphy 2001). The number of parents a node may have when experimented with the K2 algorithm is not limited a priori.

We preliminary studied three types of structures. NBC is a learning-free structure that when represented as a BN and adopted to FISH signal classification is shown in Figure 2 (top left) (the NBC structure). Having no expert to guide structure learning, NBC can be considered as the most generic expert-based structure. The result of applying expert knowledge to improve NBC by adding necessary and removing unnecessary edges is shown in Figure 2 (top right) (the expert-based structure). Added edges from the maximum hue (6) and average hue (7) nodes to the delta hue (9) node, as well as from the area (1) and channel intensity [total (3) and average (4)] nodes to the average gray intensity (12) node reflect expert knowledge. The third structure (Figure 2, bottom left) is based on the K2 algorithm being initialized using one of the possible orderings coinciding with the expert knowledge.

The first three rows of Table 1 present only slight differences between the accuracies of classifiers based on the above three structures. The similarity in accuracies between NBC and the expert-based classifier is

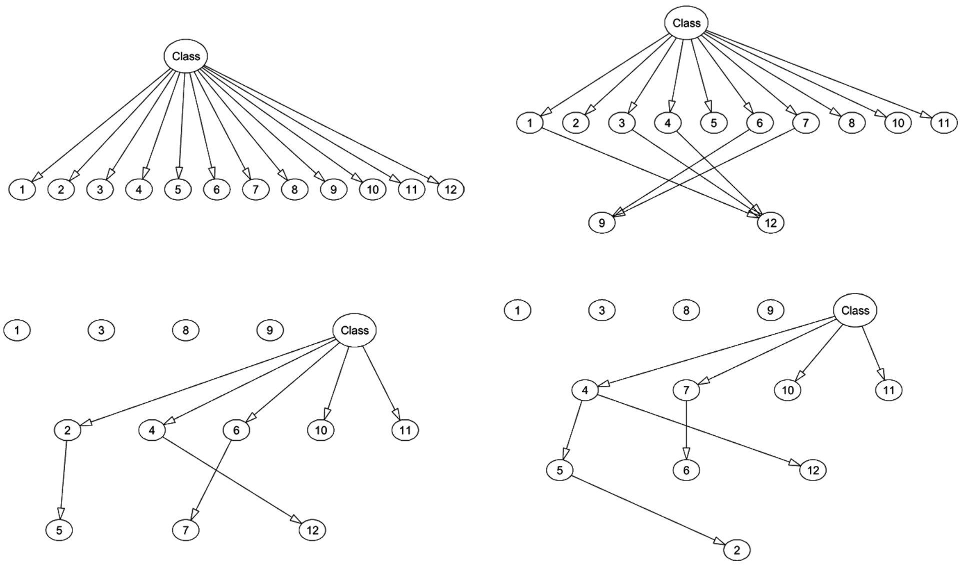

FIGURE 2 Structures constructed for FISH signal classification using NBC (top left) or expert knowledge (top right) as well as learned from the data using the K2 algorithm having an initial ordering based on expert knowledge (bottom left) or features ranked using the J_3 scatter criterion (bottom right). Node numbers correspond to the features given in the FISH Signal Representation and Classification section.

attributed to the coupling (in this study) between the corresponding structures as explained above. It is also the class node Markov blanket[3] in the expert structure that separates this node from all nodes that are not its children. That is, nodes (9) and (12) of the expert structure (Figure 2, top right) do not participate in the classification. Because these two features are almost irrelevant to FISH signal classification when the other features are being used (Lerner et al. 2001; Lerner 2004), the accuracy of NBC is

TABLE 1 Accuracy of FISH Signal Classification Based on BNCs Using NBC, Expert Knowledge, Different Settings for the K2 Algorithm or Different Combinations of Hidden Variables

Model	Classification accuracy (mean [std] in %)
NBC	78.0 (2.1)
Expert-based BNC	79.5 (2.1)
Expert-initialized K2-based BNC	78.0 (2.1)
J_3-initialized K2-based BNC	74.5 (2.0)
BNC based on NBC, HCS, and K2 metric	80.1 (2.0)
Expert-based BNC with intensity hidden node	78.1 (1.7)
Expert-based BNC with color hidden node	80.5 (2.5)
Expert-based BNC with intensity and color hidden nodes	80.9 (2.3)

See the Experiment and Results section for full details.

deteriorated compared with that of the expert-based classifier (Table 1). The K2-based classifier is inferior to the expert-based classifier because the corresponding structure of the former (Figure 2, bottom left) avoids several of the features (1, 3, 8, 9) used by the structure corresponding to the latter, also keeping other features (5, 7, 12) from participating in the classification (again, due to the Markov property). Although it may be justified for some of the overlooked features (Lerner et al. 2001; Lerner 2004), it cannot be justified for other relevant features, and thus the K2-based classifier achieves lesser accuracy than the expert-based structure.

The incapability of the K2-based classifier in improving the NBC and expert-based classifier accuracies has led to this study. To improve accuracy of K2-based classifiers, we investigate three aspects of the K2 algorithm, namely initial ordering, algorithm search, and score (Figure 1). Following, we experimentally study different initializations (see the K2 Initial Ordering section) and search procedures (see the K2 Search section) to the algorithm. We also extend the ability of the structure in representing the domain by including hidden variables manifesting relations between variables not straightforwardly evident (see the Hidden Nodes section). In addition, we demonstrate the limitation in using the K2 score for learning BN classifiers (see the K2 Score section).

K2 Initial Ordering

The K2 algorithm requires, and thus depends on, an initial topological ordering, that is, an ordering in which a parent precedes its children (see the K2 Learning Algorithm section). However, this ordering does not necessarily accommodate the optimal node ordering. Moreover, it is not uniquely determined thus have to be based on prior knowledge if exists or otherwise set arbitrarily. For example, the initial ordering of the K2-based structure shown in Figure 2 (bottom left) was determined, in the absence of any other a priori information, based on the 12 features given in the FISH Signal Representation and Classification section ordered 1, 2, . . . , 12. This ordering may also represent NBC and may accommodate the previous expert knowledge that variables 1, 3, and 4 should precede variable 12 as well as 6 and 7 should precede 9. This can partially explain the resemblance in accuracy between the three classifiers as reflected in Table 1.

Another source for the similarity of the structures (and hence also their accuracies) is that for all the above examined orderings we positioned the class node before all 12 variables. It corresponds to the assumption that the class variable is the most significant factor in classification and thus should be positioned first in the ordering making this variable a potential parent of all variables. It also allows the computation using (2). Moreover, it was noted (Cheng and Greiner 1999; Friedman, Geiger, and Goldszmidt 1997; Madden 2003; Singh and Valtorta 1995) that structures learned when placing the class variable first in the ordering may have smaller values of the K2 metric but will lead to higher predictive accuracies. Madden (2003) called a structure learned according to this scheme a "selective BN augmented NBC."

We further suggest ranking the domain features based on the degree of separability they provide in FISH signal classification and establish an initial ordering for the K2 algorithm based on this ranking and hence discrimination-oriented. This data-driven ordering is also useful when prior knowledge to guide the determination of the ordering is missing or not robust enough, as in our case. Therefore, we sort the individual features by the values they get for the class separability criterion J_3 (Devijver and Kittler 1982). A high value of the J_3 criterion indicates a feature contributing to high class separability, that is, classes that are concentrated around the corresponding feature expectation values and are far away from each other. The K2 search algorithm adds incrementally for a node that parent from a given ordering whose addition increases the score of the resulting structure the most (see the K2 Learning Algorithm section). It is therefore reasonable to believe that by placing features that contribute significantly to classification high on the ordering (i.e., potential parents of other variables), we assist learning structures providing good discrimination. In addition, this method of initialization is fast and simple and also advantageous computationally compared to repetitive random initialization of the ordering.

However, comparing K2-based structures derived when the algorithm is initialized using orderings based on expert knowledge (Figure 2, bottom left) or J_3 scatter criterion (Figure 2, bottom right), we observe no substantial differences between the structures. Although starting using different topological orderings, in both cases the K2 algorithm considers the same variables (1, 3, 8, and 9) as irrelevant to FISH signal classification and independent on the other variables and thus leaves these variables unconnected to the structure. The algorithm finds approximately the same variables (4, 6 [or 7 as the two features are almost interchangeable {Lerner et al. 2001}], 10, and 11) as contributors to the K2 score, independently of the method of initialization, and thus connects them to the class variable. In addition, it identifies the same edges between pairs of variables (2–5, 6–7, and 4–12), although the directions of the edges may be different. Moreover, structures similar to these have also been found for random initial orderings. Therefore, the similarity of K2-based structures could be explained by either marginal sensitivity of the algorithm to the initial ordering or the position of the class variable first in the ordering. We also note that unexpectedly, the J_3-initialized K2-based classifier yielded lower accuracy than the expert-initialized K2-based classifier (fourth and third rows in Table 1, respectively).

Finally, and as explained in section 2.2.3 after Eq. (3), the K2 metric prefers simple structures. After highly dependent variables, yielding a high value of $g(X_i, \mathbf{Pa}_i)$, are connected as parent and child nodes, the K2 algorithm usually rejects other parents in the ordering from being included in the child–parent set. This is because additional parents should raise the K2 metric to be included in the set, but the highly dependent first parents already provided unbeatable high values for the K2 metric. This is exemplified in Figure 2 (bottom left), for example, in the relations $g(5, 2) > g(5, 2 \cup 3)$, $g(5, 2 \cup 4)$. Thus, not only are the differently initialized K2-based structures similar to each other, it is also unlikely to find a K2-based structure that is not a tree, which due to the Markov blanket of the class variable turns to be a degenerate NBC (also called selective Bayesian classifier [Langley and Sage 1994]). Figure 2 (bottom) supports this insight. Hence, and for all these reasons, we report our findings, which are important to research in the field, but see no point in further research on the impact of initial ordering on the K2 algorithm for the cytogenetic domain.

K2 Search

Another source for the similarity of the K2-based structures is the K2 search procedure. In each step of the algorithm, it includes in X_i's set of parents either the class variable (because in our case it always precedes X_i) or the variable that together with X_i's current set of parents affects X_i the most, as measured by the highest value of $g(X_i, \mathbf{Pa}_i)$ (see the K2 Learning Algorithm section). That is, measuring $g(X_i, \mathbf{Pa}_i)\ \forall i$ and independently of the type of ordering, we expect (1) the variables relevant to classification to be connected to the class variable by an edge (e.g., $class \rightarrow 2$ in Figure 2 (bottom left) because $g(2, class) > g(2)$, where $g(2)$ is the initial metric value (Cooper and Herskovits 1992) for variable 2); (2) those variables that depend on variables already connected to the structure more than on the class variable to be connected to these variables and not to the class variable (e.g., the edge $2 \rightarrow 5$ in the same structure since $g(5, 2) > g(5) > g(5, class)$); and (3) those variables irrelevant to classification that also depend only weakly on all nonclass variables not to be connected to the structure at all (e.g., variable 8 in the same structure because $g(8) > g(8, class), g(8, 1), \ldots, g(8, 7)$). The results of this pattern of edge connection performed by the K2 search is evident in the similarity between the two structures in the bottom of Figure 2, although they were initialized using different orderings.

Hence, and due to the failure to improve accuracy through the algorithm initial ordering, we consider replacing the K2 search with HCS. HCS is less restrictive and enables the exploration of larger structure spaces and incorporation of prior knowledge through the initialization of the structure. Alternatively, HCS dispenses with the requirement for an initial ordering. Using NBC as the initial structure to HCS, the resulted structure is shown in Figure 3 and the corresponding accuracy in Table 1 (fifth row). We note that the HCS-based structure has higher K2 metric value than the NBC structure as the latter is the starting point for the former and the two structures are different. We also note that the NBC-initialized HCS-based structure provides richer representation of the domain than its counterpart constructed without learning. This representation is also translated into approximately 2% accuracy improvement (first and fifth rows in Table 1). Trying to avoid being trapped in a local maximum when searching a structure using the K2 search or HCS, we replace HCS with the simulated annealing (Kirkpatrick, Gelatt, and Vecchi 1983) search. This allows searching broader spaces and escaping local maxima easily; however, we find no improvement to the classification accuracy from using this computationally expensive method.

Hidden Nodes

Before studying the aspect of the K2 metric, we investigate the impact of including hidden variables on the representation of the cytogenetic domain and according to three expert beliefs. First, by incorporating into an expert-based structure a color hidden node, acting as a parent of all

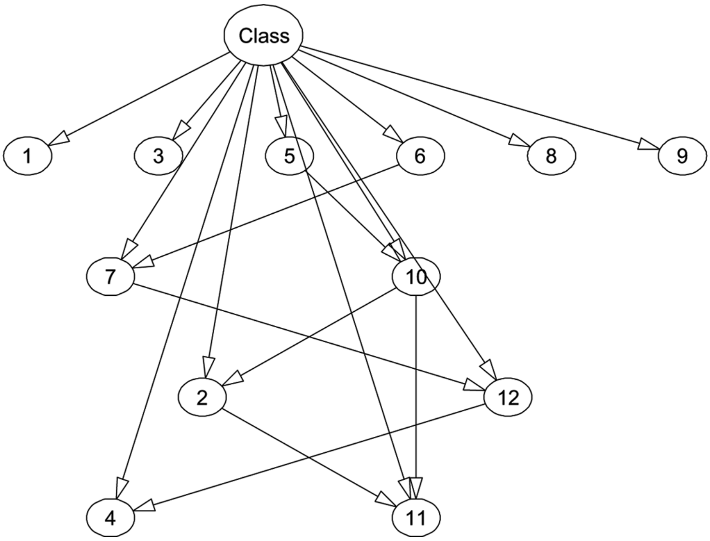

FIGURE 3 A structure learned using the K2 metric and NBC as a starting point for hill-climbing search. Node numbers correspond to the features given in the FISH Signal Representation and Classification section.

color nodes, we accommodate a belief that all the used hue features are the result of a single source (Figure 4, top–left). Second, by introducing an intensity hidden node to another expert-based structure, we address a belief that the intensity variables originate from a single source (Figure 4, top–right). A modified combination of these two beliefs using two hidden nodes is shown in Figure 4 (bottom). From judging the accuracies in Table 1, we summarize that the first belief regarding color is probably more significant for improving the classification accuracy. Nevertheless, further experimentation is needed to establish this conclusion.

K2 Score

Replacing the restricted K2 search with the more flexible HCS relieved the dependence on an initial ordering and led to some accuracy improvement; however, we would like to improve accuracy further. Ruling out both the K2 initial ordering and search procedure as the main causes to inferior accuracy, we examine now the suitability of the K2 metric to BNC structure learning.

We recorded the structures as well as the K2 metric values and classification accuracies that correspond to these structures for subsequent iterations of HCS. Figure 5 shows for one particular fold of the CV10

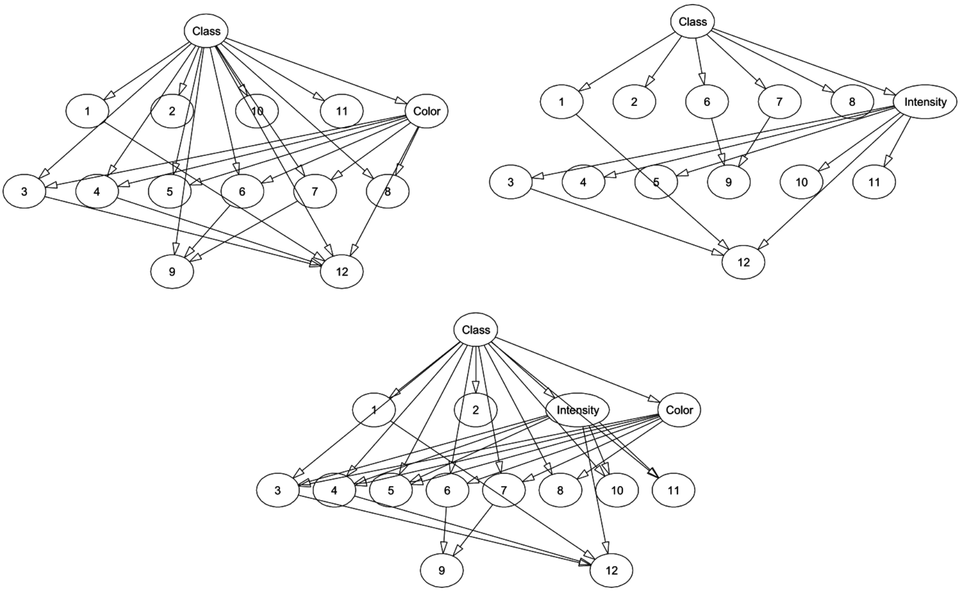

FIGURE 4 Structures based on expert knowledge with the inclusion of a color hidden node (top–left), intensity hidden node (top–right), or both nodes (bottom). Node numbers correspond to the features given in the FISH Signal Representation and Classification section.

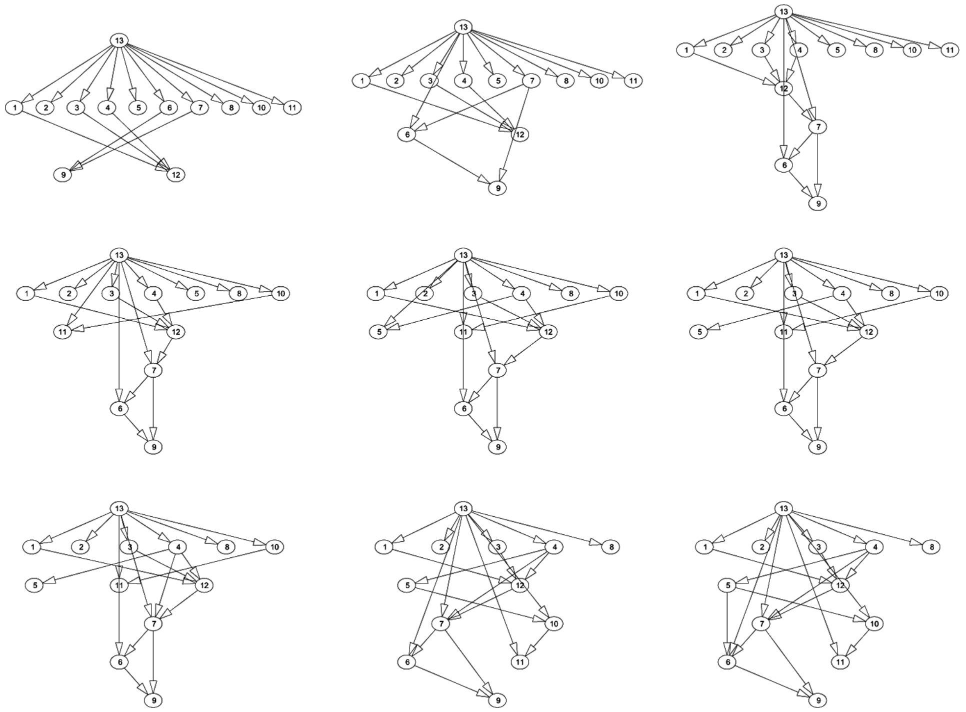

FIGURE 5 Structures learned using the K2 metric and hill-climbing search. Starting from the top left (expert-based structure) and moving to the right and bottom, the figure shows a structure for each of the eight iterations of the search. Node numbers correspond to the features given in the FISH Signal Representation and Classification section and number 13 represents the class variable.

experiment structures derived from the expert-based structure (Figure 2, top right) during eight iterations of HCS. Figure 6 plots the K2 metric values and classification accuracies corresponding to these iterations averaged over the 10 folds of the CV10 experiment. For each iteration of HCS, the structures derived for the different folds of the CV10 experiment yield similar values of the K2 metric and hence the low standard deviation for this criterion, as exemplified in Figure 6. Though having similar K2 metric values, these structures are different enough to yield average classification accuracy having relatively large standard deviation. For most folds of the CV experiment, HCS needed six to seven iterations to achieve its highest K2 metric value. In one particular fold it required eight iterations; thus, the standard deviations of the K2 metric and classification accuracy for eight iterations are zero.

To understand Figure 6, we associate the values of the two criteria–K2 metric and classification accuracy–for additional iterations of HCS to the corresponding learned structures (Figure 5). We note that the initial structure (i.e., the expert-based structure which is first in Figure 5) has low value

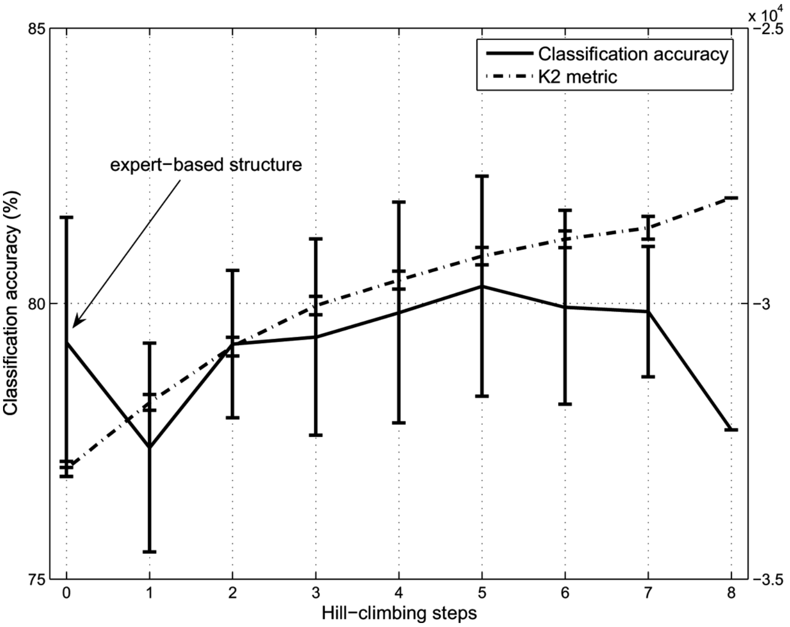

FIGURE 6 K2 metric values of structures learned from the expert-based structure using hill-climbing search, as well as the accuracies of classifiers based on these structures and averaged over a CV10 experiment for increasing numbers of iterations of the search.

of the K2 metric but the expert-based classifier is relatively accurate (as is demonstrated in Figure 6 and the second row in Table 1). In the first HCS iteration (second graph in Figure 5), the edge 7 → 6 is added because it contributes the most to the K2 metric value of the expert structure. This edge reflects the high correlation between these two hue variables (Lerner et al. 2001). However, adding a highly correlated variable may reduce the classification accuracy, because more parameters have to be estimated using the same sample size, as may be implied from Figure 6. In the second iteration of HCS, the edge 12 → 7 (two slightly correlated variables [Lerner et al. 2001]) is added (third graph in Figure 5) because it increases the K2 metric value more than other edges do. This edge connects intensity (12) and hue (7) variables that are relatively important to classification (Lerner et al. 2001), leading to improvement of the classification accuracy. This trend of simultaneous increase in the K2 metric value and classification accuracy is continued until the fifth HCS iteration (subsequently adding edges 10 → 11, 4 → 5 and dropping the edge from the class variable (13) to 5, where all these changes are supported by high dependencies and independencies, respectively [Lerner et al. 2001]). From the sixth iteration, the K2 metric continuous to increase as additional edges reflecting connections of correlated variables (e.g., 4 → 7 and 5 → 10) are added to the structure (three graphs in the bottom of Figure 5). However, the classification accuracy starts to decrease monotonically (Figure 6) because there is no added value for the classifier in the additional edges. For example, modeling the linkage between intensity and hue variables by the edge 4 → 7 (sixth iteration) is redundant because it is already modeled by the existing edge 12 → 7 (both 4 and 12 are intensity features). Furthermore, the last added edges do not contribute to the classification accuracy of the structure, but they raise the number of parameters that should be estimated using the same data. This reduces the classification accuracy because the number of instances representing each combination of variable states that is required to learn a parameter decreases. The reduction reaches 3% when measured between the fifth and eighth HCS iterations (Figure 6). This is a very important result that demonstrates experimentally for the cytogenetic domain that the K2 metric is not an appropriate criterion for BNC structure learning. A structure learned using the K2 metric may represent the dependencies and independencies within the domain precisely and therefore be appropriate for data representation and general inference but is not necessarily the basis of an accurate classifier. Hence, replacing the K2 metric with a classification-oriented score has the highest potential to improve the accuracy of K2-based classification.

DISCUSSION

We learned BNC structures for automatic signal classification in FISH images that are necessary for the diagnosis of genetic abnormalities. Having at our disposal only partial expert knowledge, the best way to evaluate these structures, especially when used for classification, was using the classification accuracy the structures provide. A structure may be constructed based on a restrictive assumption such as that of NBC or based on expert knowledge setting the actual connections between nodes representing the variables. Because the first approach is usually restricted in complex domains revealing a high degree of dependency among variables and the second approach is biased and time-consuming, there are advantages to learning a BN structure directly from the data. Using an initial ordering on the variables, a search algorithm, and by maximizing a metric measuring the joint probability over the variables, the K2 algorithm learns a structure. We studied here these aspects of the algorithm to improve the classification accuracy of the learned structure.

Using an ordering on the variables allows the K2 algorithm to reduce the combinatorics in structure learning enormously, but any such ordering may be uncertain. We found that all examined orderings provided tree-like structures that are turned into NBC-like structures in the case of classification. This similarity of structures is due to (1) forcing the class variable to be first in the K2 initial ordering, (2) the K2 search, and (3) the K2 metric favoring simple structures. Unfortunately, these orderings failed to improve the accuracy of the K2-based classifier for the cytogenetic domain. We suggest examining other data-driven methods for establishing initial orderings; one such method may use conditional-independence tests similarly to Singh and Valtorta (1995).

To expand the search, we replaced the K2 search with HCS that provided richer representation of the domain and improved accuracy. In addition, the utilization of hidden nodes enabled broader and more accurate modeling of the domain compared with that derived using only expert knowledge. It also enhanced the accuracy. Because expert knowledge in our domain is only partial, we are interested in exploring ways to automatically identify hidden nodes, similar to Elidan et al. (2001), Friedman (1997), and Silva et al. (2006).

Investigating the K2 metric, we experimentally demonstrated for the cytogenetic application that structures selected according to the K2 metric are appropriate for general inference but do not necessarily provide accurate classifiers. There are additional reports that support this limitation of the K2 metric (Herskovits 1991; Singh and Valtorta 1995) and related likelihood-based metrics (Friedman, Geiger, and Goldszmidt 1997; Greiner et al. 2005; Grossman and Domingos 2004; Kontkanen et al. 1999) for BNCs. Classification-driven scores have been suggested (Grossman and Domingos 2004; Kontkanen et al. 1999), and they are also the target of our current research. Preliminary results show superiority of a novel classification-driven score over likelihood-based scores in classifying the cytogenetic data.

It is disappointing that all our efforts to improve the accuracy of the K2-based classifier have led to only limited success. However, these efforts enabled us to throughly explore the K2 algorithm probably as never before explored. We believe this exploration is a considerable contribution to the field, which we wish to share with the community.

BN is a generative model that is very useful in modeling the joint probability distribution. However, for accurate modeling of this distribution, BN requires a large sample size. The relatively small sample size of the cytogenetic database is more appropriate for a discriminative model, such as the neural network (NN) maximizing the posterior probability, rather than for estimating the joint probability distribution. This has been proved for this database and several other BNCs. For example, the TPDA algorithm (Cheng, Bell, and Liu 1997), PC algorithm (Spirtes, Glymour, and Scheines 2000), Chow-Liu multinet (Friedman, Geiger, and Goldszmidt 1997), tree-augmented naive (TAN) Bayes (Friedman, Geiger, and Goldszmidt 1997), RAI algorithm (Yehezkel and Lerner 2009), and *t*BCM[2] (Gurwicz and Lerner 2006) achieved accuracies between 77.2% and 82.9% (Gurwicz and Lerner 2006), which are similar, though sometimes slightly superior, to those reported in the current study. Therefore, we believe that a classification accuracy of 80% to 83% for the cytogenetic database is about the best a BNC can get. We suspect the only way to improve this accuracy toward those of non-BN discriminative models (e.g., NN and SVM providing accuracies of ∼87% [David and Lerner 2005; Lerner et al. 2001]) is by substantially increasing the database size.

Another objective of future research is to investigate the implications of this study to FISH dot counting and clinical genetic diagnosis. Additionally, we are interested in examining the conclusions of the study using other applications so as to further generalize the relative contribution of each aspect of the K2 algorithm to classification.

NOTES

1. The NBC (Langley and Sage 1994) is a Bayesian network (Pearl 1988) that assumes the observable variables are independent conditioned on the class variable.
2. Learning a hidden (latent) concept usually comes in two main flavors: learning a structure that may have hidden variables (see the Inclusion of Hidden Nodes section) or learning the parameters for a set of variables having unobserved (hidden) states where other variables in the structure are fully observed (see the Inference section).
3. The Markov blanket of a node includes the node, its parents, children, and children coparents (Pearl 1988).

REFERENCES

Borgelt, C., and R. Kruse. 2001. An empirical investigation of the K2 metric. In *ECSQARU' 01: Proc. of the 6th European Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, ed. S. Benferhat and P. Besnard, 240–251. London, UK: Springer-Verlag.

Cheng, J., D. A. Bell, and W. Liu. 1997. An algorithm for Bayesian belief network construction from data. In *Proc. of AI & STAT' 97*, 83–90. Ft. Lauderdale, FL.

Cheng, J., and R. Greiner. 1999. Comparing Bayesian network classifiers. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence*, ed. K. B. Laskey and H. Parde, 101–108. Stocholm: Morgan Kaufmann.

Cooper, G. F., and E. Herskovits. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9:309–347.

David, A., and B. Lerner. 2005. Support vector machine-based image classification for genetic syndrome diagnosis. *Pattern Recognition Letters* 26:1029–1038.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39:1–38.

Devijver, P. A., and J. Kittler. 1982. *Pattern recognition–A statistical approach.* Englewood Cliffs, NJ: Prentice-Hall.

Elidan, G., N. Lotner, N. Friedman, and D. Koller. 2001. Discovering hidden variables: A structure-based approach. *Advances in Neural Information Processing Systems* 13:479–485.

Friedman, N. 1997. Learning belief networks in the presence of missing values and hidden variables. In *Proc. of the 14th Int. Conf. on Machine Learning*, 125–133. San Francisco: Morgan Kaufmann.

Friedman, N., D. Geiger, and M. Goldszmidt. 1997. Bayesian network classifiers. *Machine Learning* 29:131–163.

Ghahramani, Z., and M. I. Jordan. 1994. Learning from incomplete data. In Technical report 108, MIT Center for Biological and Computational Learning, Massachusetts.

Greiner, R., X. Su, B. Shen, and W. Zhou. 2005. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. *Machine Learning* 59:297–322.

Grossman, D., and P. Domingos. 2004. Learning Bayesian network classifiers by maximazing conditional likelihood. In *Proc. of the 21th Int. Conf. on Machine Learning*, 361–368. Banff, Canada: ACM Press.

Gurwicz, Y., and B. Lerner. 2006. Bayesian class-matched multinet classifier. *Lecture Notes in Computer Science* 4109:145–153.

Heckerman, D. 1995. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95–06, Microsoft Research.

Heckerman, D., D. Geiger, and D. M. Chickering. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20:197–243.

Herskovits, E. 1991. Computer-based probabilistic-network construction. Ph.D. dissertation, Department of Computer Science, Stanford University.

Huang, C., and A. Darwiche. 1994. Inference in belief networks: A procedural guide. *Int. Journal of Approximate Reasoning* 15:225–263.

Keogh, E. J., and M. J. Pazzani. 2002. Learning the structure of augmented Bayesian classifiers. *Int. Journal on Artificial Intelligence Tools* 11:587–601.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220 (4598): 671–680.

Kontkanen, P., P. Myllymäki, T. Sliander, and H. Tirri. 1999. On supervised selection of Bayesian networks. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence*, ed. K. Laskey and H. Prade, 334–342. San Francisco: Morgan Kaufmann.

Langley, P., and S. Sage. 1994. Induction of selective Bayesian classifiers. In *Proc. of the 10th Conf. on Uncertainty in Artificial Intelligence*, ed. R. Lopez de Mantaras and D. Poole, 399–406. Seatle, WA: Morgan Kaufmann.

Lauritzen, S. L., and D. J. Spiegelhalter. 1988. Local computation with probabilities on graphical structures and their application to expert systems. *Journal of Royal Statistics B* 50:157–224.

Lerner, B. 2004. Bayesian fluorescence in-situ hybridization signal classification. *Artificial Intelligence in Medicine* 30:301–316.

Lerner, B., W. F. Clocksin, S. Dhanjal, M. A. Hultén, and C. M. Bishop. 2001. Feature representation and signal classification in fluorescence in-situ hybridization image analysis. *IEEE Trans. on SMC A* 31:655–665.

Lerner, B., L. Koushnir, and J. Yeshaya. 2007. Segmentation and classification of dot and non-dot-like fluorescence in-situ hybridization signals for automated detection of cytogenetic numerical abnormalities. *IEEE Trans. on Information Technology in Biomedicine* 11:443–449.

Lerner, B., J. Yeshaya, and L. Koushnir. 2007. On the classification of a small imbalanced cytogenetic image database. *IEEE Trans. on Computational Biology and Bioinformatics* 4:204–215.

Luo, J., and M. Boutell. 2005. Automatic image orientation detection via confidence-based integration of low-level and semantic cues. *IEEE Trans. on PAMI* 27:715–726.

Madden, M. G. 2003. The performance of Bayesian network classifiers constructed using different techniques. In *Working Notes of the ECML/PKDD-03 Workshop on Probuhilistic Graphical Models for Classification*, 59–70.

Malka, R., and B. Lerner. 2004. Classification of fluorescence in-situ hybridization images using belief networks. *Pattern Recognition Letters* 25:1777–1785.

Murphy, K. 2001. The Bayes Net Toolbox for Matlab. *Computing Science and Statistics*, 33:331–350.

Nath, J., and K. L. Johnson. 2000. A review of fluorescence in situ hybridization (FISH): Current status and future prospects. *Biotechnic Histochemistry* 75:54–78.

Pazzani, M. J. 1996. Searching for dependencies in Bayesian classifiers. In *Learning from data: AI and statistics V*, ed. D. Fisher and H. J. Lenz, 239–248. New York: Springer Verlag.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco: Morgan Kaufmann.

Pearl, J., and T. S. Verma. 1991. A statistical semantics for causation. *Statistics and Computing* 2:91–95.

Pena, J. M., J. A. Lozano, and P. Larranaga. 1999. Learning Bayesian networks for clustering by means of constructive induction. *Pattern Recognition Letters* 20:1219–1230.

Silva, R., R. Scheines, C. Glymour, and P. Spirtes. 2006. Learning the structure of linear latent variable models. *Journal of Machine Learning Research* 7:191–246.

Singh, M., and M. Valtorta. 1995. Construction of Bayesian network structures from data: A brief survey and an efficient algorithm. *International Journal of Approximate Reasoning* 12:111–131.

Spirtes, P., C. Glymour, and R. Scheines. 2000. *Causality, prediction and search*. 2nd ed. Cambridge, MA: MIT Press.

Yehezkel, R., and B. Lerner. 2009. Bayesian network structure learning by recursive autonomy identification. *Journal of Machine Learning Research* 10:1527–1570.

Zhang, Y., and Q. Ji. 2005. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. on PAMI* 27:699–714.