Bayesian Networks Classifiers applied to Documents

Souad Souafi-Bensafi[1,2], Marc Parizeau[2], Franck Lebourgeois[1], Hubert Emptoz[1]

[1]Reconnaissance de Formes et Vision,
I.N.S.A. de LYON - Bât J. Verne, 20 Av. A. Einstein,
69621 Villeurbanne Cedex FRANCE

[2]Laboratoire de Vision et de Systèmes Numériques,
Université Laval, Département de génie électrique
et de génie informatique, Québec, CANADA, G1K 7P4

Abstract

This paper discusses the use of the bayesian network model for a classification problem related to the document image understanding field. Our application is focused on logical labeling in documents, which consists in assigning logical labels to text blocks. The objective is to map a set of logical tags, composing the document logical structure, to the physical text components. We build a bayesian network model that allows this mapping using supervised learning, and without imposing a priori constraints on the document structure. The learning strategy is based partly on genetic programming tools. A prototype has been implemented, and tested on tables of contents found in periodicals and magazines.

1 Introduction

Bayesian Networks (BN) are probabilistic graphical models that represent a set of random variables for a given problem, and the probabilistic relationships between them [11]. They have been particularly used for problems involving reasoning under uncertainty in artificial intelligence, in different applications including medical diagnostics, classification systems and software debugging. They can be learned from a set of observed data. In this context, we propose a genetic learning method, which does not impose any constraint on the structure of the BN. This paper aims at describing this method for a classification problem, and its application for the first time in the field of document understanding.

The application lies within the logical labeling step of document logical structure recognition. Indeed, a document is considered at two main structuring levels: physical and logical. The physical level represents the layout structure of physical components like characters, words, lines, paragraphs, etc. The logical structure is usually composed of a set of logical functions or labels on the one hand, that need to be assigned to the physical components, and of the relationships between these components on the other hand. Document recognition has been the subject of much research in the last few decades. Most methods are based on syntactical approaches [1], or arise from artificial intelligence such as, rule-based [14] or knowledge-based systems [3]. However, because these approaches were designed to work on very structured documents, they are often inconvenient for documents with irregularities in their organization. Moreover, the complexity of the layouts makes the analysis at the physical level difficult, both for image segmentation and feature extraction of the components. Since logical labeling is often fully dependent on these features, the errors that occur at the physical level cause an instability of the structures that can directly affects the logical level. A typical example of such documents are tables of contents in periodicals or magazines (Figure 1).

Figure 1. A table of contents example.

Our objective in this application, is to develop a generic approach that will allow an adaptive graphical representation of logical components in documents from physical features. We decided to explore a probabilistic approach, namely BN classifiers, expecting model adaptation to eventual irregularities in physical features. Logical labeling has already been tested using a naive bayesian classifier, and in-

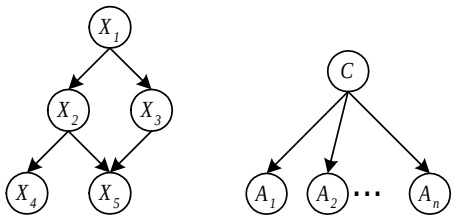

Figure 2. (a):A example of Bayesian Network. (b):Naive Bayesian Classifier.

teresting results were obtained [13]. But the advantage of BN classifiers over naive bayesian classifiers, is that they allow the selection of the most salient features and relationships between features.

The rest of the paper is organized as follows. Section 2 presents a brief introduction on BN, BN classifiers, and their learning and inference processes. Section 3 exposes our specific genetic approach for learning BN classifiers. The system overview of our application framework is then described in section 4, and some results on tables of contents of different periodicals are given in Section 5.

2 Bayesian Networks

We need firstly to give a formal definition of BN in order to expose their learning and inference problems. But since, we use them in a classification problem, we will focus on BN classifiers.

For a set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, a corresponding BN is represented partly by a Direct Acyclic Graph (DAG), in which the nodes represent the variables and the edges express the dependence between variables (Figure 2a). To each variable X_i corresponds the set of its parents Π_{X_i} formed by the variables it depends upon. The second part of the BN is the set of conditional probabilities $P(X_i|\Pi_{X_i})$ according to the graph. The probability among the set $\mathbf{X}$ can thus be decomposed by: $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\Pi_{X_i})$.

When BNs are used for classification problems, the set of variables is composed of the class C and of attributes (features) $A_1, A_2, \ldots, A_n$, with n being the number of attributes. The naive bayesian classifier is a particular case of BNs, in which the attributes are assumed mutually independent (Figure 2b). Such strong hypothesis, however, is usually unfounded, and the general BN structure is much more powerful.

Bayesian Network learning consists of two parts: learning the graph structure and learning the conditional probabilities. Learning the structure, requires a search procedure with a score function. Two main types of score functions have been used: MDL (Minimum Description Length) [4], and bayesian [2]. However this general learning problem is NP–Hard [6]. Therefore, non-deterministic approaches have been experimented, as for example genetic algorithms [8]. Also, constraints can be imposed among the random variables or on the structure types. For classifiers in particular, different restrictive structures have been studied [4, 10]. The main goal is to assume a minimum of constraints on the structure to be as general as possible, but at the same time to make their manipulation as simple and efficient as possible.

Conditional probabilities can be estimated using *sufficient statistics* which correspond, for the case of discrete random variables, to counting from the training set the number of occurrences of each combination of variable/parents. Given a training set $\mathbf{U}$ containing vectors of values, and $Val(X_i) = \{x_i^1, x_i^2, \ldots, x_i^{r_i}\}$ the value domain of $X_i \in \mathbf{X}$, with r_i the number of the possible values for X_i. Let q_i be the number of distinct values of Π_{X_i} according to the training set $\mathbf{U}$: π_i^j with $j \in \{1, 2, \ldots, q_i\}$. We note N_{ijk} the number of cases in $\mathbf{U}$ for which $X_i = x_i^k$ and $\Pi_{X_i} = \pi_i^j$. Then the conditional probabilities can be estimated from $\mathbf{U}$ and the network structure by simply counting the N_{ijk}: $P[X_i = x_i^k|\Pi_{X_i} = \pi_i^j] = N_{ijk}/\sum_{k=1}^{r_i} N_{ijk}$.

For the general BN model, the inference process consists in determining various probabilities of interest within the model. This problem being NP–Hard, several approximate algorithms have been proposed [6]. But the classifier inference problem is much simpler. It only requires the computation of the class probabilities give the attribute values : $P(C|A_1, A_2, \ldots, A_n) = \alpha \cdot P(C|\Pi_C) \cdot \prod_{i=1}^{n} P(A_i|\Pi_{A_i})$, with α being a normalization factor [4].

3 Learning BNs: proposed approach

We propose a new method for learning BN structure using Genetic Programming [7] (not to be confused with genetic algorithms). GP is an *evolutionary algorithm* that evolves an initial *population* of individuals (programs) and seeks to discover the best breed of programs using three basic genetic operators: *selection* biased toward the fitest individuals, *crossover* to exchange genetic material between individuals, and *mutation* to stem new genetic material. GP typically uses a tree structure to represent programs [5], where the tree nodes are primitives of the solution domain. Branches in the tree represent primitive functions and leave denote the problem parameters. For our BN classifier problem, the programs will correspond to network structures and the *fitness function* will be computed from the score function. The node will correspond to the random variables of the BN, and the father-son relations will define the dependence between variables. Now the main problem is to define a procedure to convert a rooted tree structure into a DAG structure. The solution is to iterate over the tree and insert

each arc into the DAG structure as long as it does not induce a cycle. Arc that produce a cycle are simply discarded. And since the GP framework is constrained so that the tree root is always the class variable, and also since attributes can appear any number of times in the tree, then all DAG structures can stem from the trees. For more details, the reader is referred to [12].

For the fitness function, both the bayesian [2] and the MDL function [10] have been tested. We tried also to combine linearly the two functions, with a positive factor for the bayesian one which has to be maximized, and a negative factor for the MDL function which has to be minimized. Finally, the score function corresponds to this combination weighed using the recognition rate on the training set obtained by applying the learned RB on the training data. To evaluate the fitness of an individual, the following steps are needed. First, the tree structure has to be converted to a graph structure, using the defined conversion procedure. Second, the conditional probabilities of the BN classifier must be estimated from the training set. Finally, the score function can be computed and it will correspond to the fitness measure for the GP selection process.

4 Application to logical labeling

For our logical labeling problem, we use a BN model as a classifier, to recognize relationships between physical description and logical label of given text blocks. The attributes represent the physical features of text blocks and the class variable corresponds to the label that has to be assigned to each block. We applied our method to tables of contents documents in periodical magazines, because of their complexity and variety in form, and of their content structure (Figure 1). The information to extract is organized in different text categories, that must be recognized and stored in a re-usable format. For each magazine, a classifier will be built to model the logical structure of its table of contents.

Text blocks are provided by a segmentation process on document images, using tools that were developed in our laboratory [9]. These tools are adapted to the type of documents we consider. A basic layout structure is composed of a hierarchy of geometric text blocks: characters, words and lines. Typographical information for word level blocks can also be extracted, giving a set of typographical families, each one being composed of words having a single font. In order to describe each block by an attribute vector, several features can be extracted: the typographical family of the block, its left and right neighbors, the alignment and the horizontal and vertical spacing. Exactly 8 discrete features are used. For each block, A_1, A_2, A_3 are typographical families of respectively the block itself and its left and right neighbors; A_4, A_5 are left and right horizontal distances; A_6 corresponds to the alignment of the line that contains the considered block; A_7, A_8 are above and under vertical distances. Finally, C is the class whose values are the different labels according the content of the processed document. We principally use the following labels: *section title, article title, author, page number, summary* and a value associated to the *not labeled* words.

5 Experiments

We implemented the BN's learning process and applied it to document images corresponding to tables of contents pages from four periodical magazines. The constitution of the training and test sets is shown in Table 2. For the genetic learning process, population size was fixed at 500 individuals, tournament selection mode was chosen, and different numbers of generations were used. For each periodical, learning was conducted 8 times with various parameters, leading to 8 corresponding BNs. For each of them, we performed the inference process on the document test set on a per page basis. Recognition results are summarized in Table 1 for the BN that had the best score and for the BN that, on average, provided the best performance. We can observe that the best score BNs are not the most effective. These rates are compared to those obtained using the naïve bayesien classifier (NBC) according to results given in [13] using the same features on the same document base. Different cases can be observed. For example, we can see for the periodical [1], an improvement of the mean rate using the learned BN while the minimum and maximum rates are lower compared with the NBC. These results can be explained by the structures of the learned BNs.

Figure 3 presents the BNs structures giving the best mean recognition rate for the test set and for which recognition rates are given in the Table 1. We can see that the structures produced by the learning process express independence between attribute variables in most cases. This is mainly due to our feature set and the nature of our data. We can deduce that our learning method is efficient because it gave us network structures that better reflect the dependencies in the data, even if these structures ressemble those of naïve bayesian classifiers, but the score function has to be improved.

6 Conclusion

In this paper, a generic probabilistic model is proposed for logical labeling in documents using BNs Classifiers. The goal is to perform this labeling automatically. The model is built with a supervised learning task on the basis of a training set. A prototype has been implemented and applied to periodical magazines. Significant results have been obtained, however for some documents, recognition rates were not satisfying. It is due to instability at the physical and the logical levels, for example, the quality of documents does

	Best score BN (%)			Best mean rate BN (%)			NBC (%)		
	mean	max	min	mean	max	min	mean	max	min
[1]	82.8	92.4	53.8	88.6	98.5	62.1	82.6	99.3	65.9
[2]	91.2	95.8	85.8	94.0	97.9	88.8	94.7	97.9	88.3
[3]	94.3	96.3	86.7	94.4	96.3	86.7	92.8	96.6	84.0
[4]	88.8	94.7	78.1	91.1	96.9	81.3	93.2	98.2	85.6

Table 1. Recognition rates on the test set for the learned BN (best score and best mean) and the naïve bayesian classifier; mean, max and min are computed over all document pages.

	Training set		Test set	
periodicals	#pages	#words	#pages	#words
[1] Biofutur	3	462	7	1139
[2] Cahiers…	3	489	7	1328
[3] Nature	4	1718	14	6059
[4] NewsWeek	3	641	11	2445

Table 2. Documents training set.

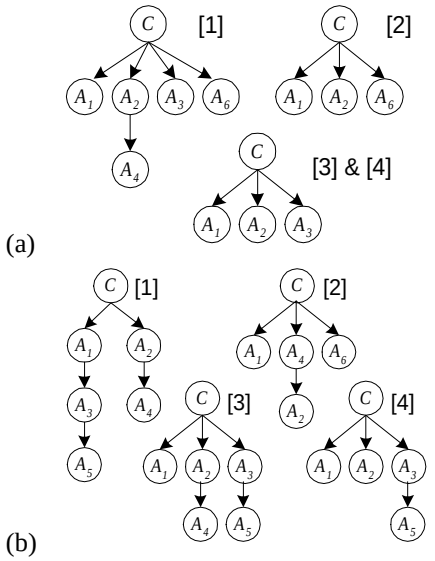

Figure 3. Learned BNs for each periodical: (a) Best mean rate BNs; (b) Best score BNs.

not always allow a perfect segmentation and feature extraction. We compared learned BNs to naïve bayesian classifiers but the difference was not significant because of the nature of our data which is particularly well represented by naïve structures. In our work, we considered two distinct problems: BN's learning and logical labeling for document interpretation. This experience showed, on the one hand, that naïve structures are more adapted to our data, that's the reason why we plan to test genetic learning to select the best features while maintaining a naïve structure. On the other hand, it will be interesting to test our learning method on data in a context that needs to find more complex BNs structures.

References

[1] A. Belaïd. Recognition of table of contents for electronic library consulting. *IJDAR: International Journal on Document Analysis and Recognition*, 4(1):35–45, August 2001.

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

[3] F. Esposito, D. Malerba, and G. Semeraro. Automated Acquisition of Rules for Document Understanding. In 2^{th} *ICDAR*, volume 1, pages 650–654, Tsukuba Science City, Japan, October 1993.

[4] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian Network Classifiers. *Machine Learning*, (29):131–163, 1997.

[5] C. Gagné and M. Parizeau. Introduction à l'outil de programmation génétique Beagle. Rapport technique, Laboratoire de Vision et Systèmes Numériques, RT–01–LVSN, Octobre 2000.

[6] D. Hecherman. A Tutorial on Learning with Bayesian Networks. Technical report, MSR–TR–95–06, Microsoft Research, March 1995.

[7] J. R. Koza. *Genetic Programming : On the Programming of Computers by Means of Natural Selection*. The MIT Press, 1992.

[8] P. Larrañaga, M. Poza, Y. Yurramendi, R. Murga, and C. Kuijpers. Structural Learning of Bayesian Networks by Genetic Algorithms: A performance Analysis of Control Parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):912–926, September 1996.

[9] F. LeBourgeois and H. Emptoz. Document Analysis in Gray Level and Typography Extraction Using Character Pattern Redundancies. In 5^{th} *ICDAR*, pages 177–180, Bangalore, India, September 1999.

[10] S. Monti and G. Cooper. A bayesian Network Classifier that Combines a Finite Mixture Model and a Naïve Bayes Model. In *Proceeding of* 15^{th} *International Conference on Uncertainty in Artificial Intelligence*, 1999.

[11] J. Pearl. *Probabilistic Reasoning in Itelligent Systems*. Morgan Kaufmann, 1988.

[12] S. Souafi-Bensafi. *Contribution à la reconnaissancedes structures des documents écrits: Approche probabiliste*. Thesis, INSA de Lyon, France, 2001.

[13] S. Souafi-Bensafi, M. Parizeau, F. LeBourgeois, and H. Emptoz. Logical Labeling using Bayesian Networks. In 6^{th} *ICDAR*, pages 832–836, Seattle, USA, September 2001.

[14] J. Toyada et al. Study of extracting Japanese newspaper article. In 6^{th} *ICPR*, volume 2, pages 1113–1115, October 1982.