Learning Algorithms of Form Structure for Bayesian Networks

Emilie Philippot, yolande Belaïd, Abdel Belaïd

LEARNING ALGORITHMS OF FORM STRUCTURE FOR BAYESIAN NETWORKS

Emilie Philippot, Yolande Belaïd and Abdel Belaïd

Campus scientifique - BP 239 54506 Vandoeuvre-lès-Nancy Cedex - France

ABSTRACT

In this paper, a new method is presented for the recognition of online forms filled manually by a digital-type clip. This writing process is not very restrictive but it is only sending electronic ink without the pre-printed form, which will require to undertake field recognition without context. To identify the form model of filled fields, we propose a method based on Bayesian networks. The networks use the conditional probabilities between fields in order to infer the real structure. We associate multiple Bayesian networks for different structures levels (i.e. sub-structures) and test different algorithms for form structure learning. The experiments were conducted on the basis of 3200 forms provided by the Actimage compagny, specialist in interactive writing processes. The first results show a recognition rate reaching more than 97%.

1. INTRODUCTION

The work reported in this paper addresses the problem of form classification filled out manually using digital pens. This research is undertaken in collaboration with the Actimage company which is our partner specialist in interactive systems. Actimage is looking for a solution concerning the notetaking by using digital pen with clips. The digital pen is replaced by a transmission device of electronic ink without sending the pre-printed form. The use of this kind of input mode is important for the company because it accelerates the filling procedure. However, the automatic recognition of the form model becomes complex due to the context loss. Figure 1 shows an example of the problem. In the top left, is the completed form. In the upper right is the field transmitted to the system. Below, are the different form models which are candidates for the recognition.

The literature shows, for form classification, a lot of research mainly oriented towards off-line forms where the filled fields are embedded in the form structure. In [1], Ramdane et al. use a method to classify forms by a statistical approach. The document is first segmented into its main rectangular blocks. During the learning phase, each block is matched to each of the blocks in each trained class. Then they total the same number of good matches. Hence, a probability distribution of the block locations in the image surface is trained. During the identification phase, they introduce a penalty for measuring the coefficient of instability of blocks, which modifies the classical expression of the Mahalanobis distance. Xiang ([2]) proposes a new Bayesian networks (BN) model, multi-agent BN. The principle is to share a BN between several entities for security reasons. Each entity knows only a part of the network and performs local inference before sharing its findings with the other entities. In our case, we can apply this model to simplify the BN by limiting the variable number of each network. In [3], the authors propose a comparison of different structure learning algorithms for BN for classification. It shows that the naive networks give the best results.

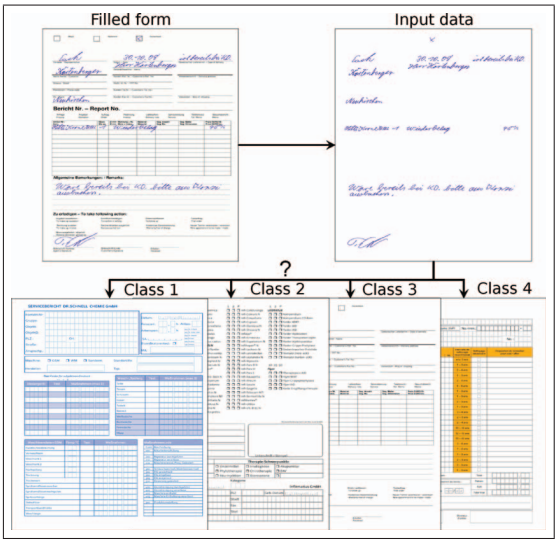

Fig. 1. The research problem

Concerning the online form recognition, the literature mentions only researches related to word recognition without considering the form structure aspect. In [4], Zhang et al. propose an approach for online multi-strokes composite sketchy shape recognition. A classifier using a double-level BN is designed to model the intrinsic temporal orders among the strokes effectively, where a sketchy shape is modeled. The drawing-style tree is then adopted to capture the users' accustomed drawing styles and simplify the training and recognition of BN classifier. In [5], the authors consider the task of structured document classification. They propose a generative model able to handle both structure and con-

tent which is based on BN. They show how to transform this generative model into a discriminant classifier using the Fisher kernel. The model is finally extended for dealing with different types of content information (here text and images).

The paper is organised as follows: first, section 2 describes the proposed approach with the different phases concerning the field extraction, the Bayesian network training and form recognition. In section 3, the first results will be presented before we conclude in the last section.

2. THE SYSTEM OVERVIEW

The approach is partly based on the observation that there are dependencies between fields in a form and between fields and the form. For example, boxes representing "Mrs.", "Mr." and "Miss" in the address area of a form are never checked simultaneously; the presence of a customer identification number implies the absence of field filling concerning the coordinates of the latter.

Figure 2 shows a dependency example that may exist for the address area of a form. Links and probabilities are used to locate and quantify the dependencies that exist between fields themselves and between the fields and the form class. For example, the *zip code* field depends on the field *country*. The table in the high right corner shows that if the *country* is filled, the probability will reach 0.8, meaning that the *zip code* field will also be filled. Conversely, if the *country* field is empty, this means that *zip code* field will be also empty with a probability reaching 0.9.

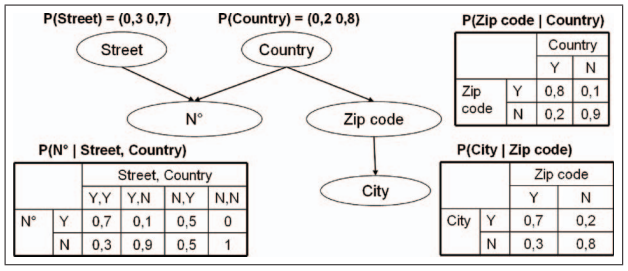

Fig. 2. Example of BN for the address area of a form

Furthermore, working with raw data, produced by a list of handwritten strokes, this may lead to incomplete, ambiguous and overlapped fields. Considering these problems, we opted for BN for their capabilities in qualitative and quantitative dependencies and uncertainty managing. We use a hierarchical approach considering the BN by area of interest, from local blocks (address client, agency information, etc.) until global form. This network hierarchy offers some advantages: 1) the number of trained variables is reduced, 2) several form areas can be represtented by the same network and 3) only modified areas must be re-trained. Once the BN representing the form areas are learned, they are gathered to train a BN for the classification of the entire form. Figure 3 shows the global flowchart of the approach.

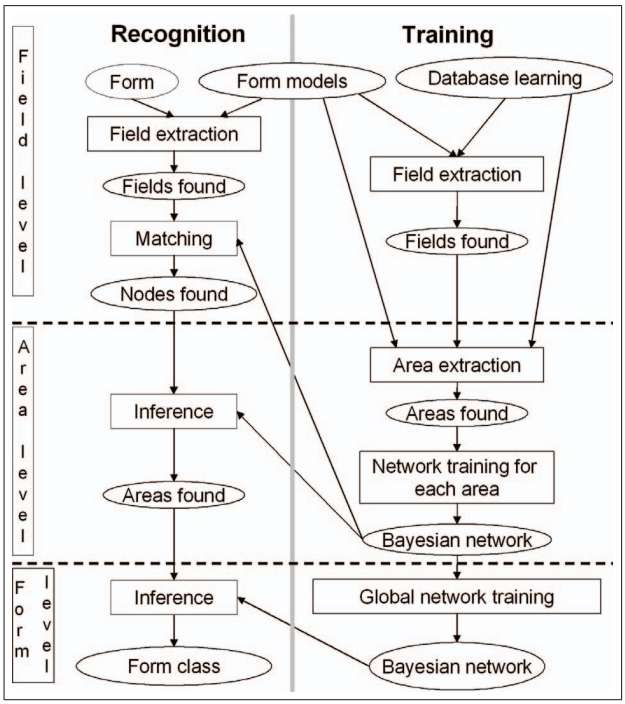

Fig. 3. Global schema of the approach

Each form class is represented by a model. It is a list of fields where each one is represented by its bounding box, its type (checkbox, string, number, etc.) and the area to which it belongs. We use an XML format to describe the model. This model will serve as a basis for the field extraction.

2.1. Field extraction

This is the first step of the system. Fields are written by hand using the digit pen and represented by a list of strokes composed each one by a list of 2D points in the form area. During the writing, the strokes are matched with a model form fields. When a majority of strokes (fixed experimentally to 85%) belongs to the bounding box of a model field, we consider that the field is filled. Once all the strokes are treated, if 20% of them have not been matched, the model candidate is excluded.

Figure 4 shows an example of different possibilities of stroke interpretation for two crosses and the firstname "Louis". In case (1), which corresponds to the reality, each cross is properly associated with a checkbox, and the name to a text box. In case (2), the firstname "Louis" is associated with a checkbox. In case (3), two crosses are combined in a single text box. The challenge will therefore lie to find form initially completed even if the strokes do not correspond fully to the fields of the latter.

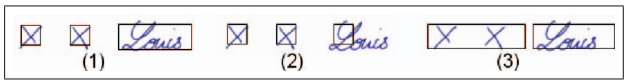

Fig. 4. Interpretation example of strokes

2.2. Bayesian network training

The training takes place in two stages. First, the main areas corresponding to the most important sub-structures of the form are identified manually and presented to the system to initiate the BN accordingly. Then, the training is continued for the entire form. The nodes of the graph represent the form fields as well as its class. Within each node, there is a probability distribution qualifying in a quantitative manner the interaction between nodes. The arcs represent the dependencies between fields. We use three different algorithms for structure training, PC, MWST and naive network [6] in order to test which of the three is best suited to our problem.

The PC algorithm starts with a fully connected graph. We test iteratively conditional dependencies existing between the n variables of the graph of order 0 (dependence between 2 variables) up to order n-1. To test these dependencies, we use the test of the χ^2. If there is conditional independence, the edge of the corresponding nodes are removed. In our case, the advantage of this method is that it allows to highlight the dependencies between fields and class.

The algorithm MWST (Maximum Weight Spanning Tree) seeks to find the covering graph of maximum weight. The starting point is the set of n nodes in the unconnected graph. The graph is built by adding nodes one by one and each addition we seek to maximize the edge weight of the graph. The advantage of this algorithm for our application is that all variables are linked and therefore all involved in the classification.

The naive BN represent a simple form of BN. They start from the assumption that a set of observed variables have an independence on an unobserved variable. The corresponding graph will consist of a single parent node corresponding to the unobserved variable, the class in our case, and several sheets representing the observed variables (e.g. fields). In this structure, the observed variables have no direct interaction between themselves.

Once the BN is trained for all the areas of the form with particular distribution probabilities, the training is enlarged to the entire form by gathering the different BN. We apply the PC algorithm and MWST in order to determine the structure of the global network. We also tested the use of naive BN. The global network is the network that summarizes the relationships between areas representing all form classes. Figure 5 shows an example of a global network obtained from the classification of two form classes. It is observed that certain areas may refer to two separate forms. This is explained by the overlapping of the central area fields in the two forms. During matching, these fields will be filled regarded as forms from class A or T.

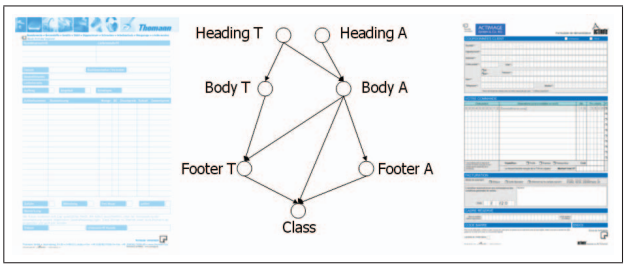

Fig. 5. Example of global BN allowing classification for forms of classes A and T.

2.3. Recognition

Recognition takes place in several stages. First, as in the training phase, fields are first extracted by matching the strokes with the form model fields. Then, for each area (i.e. form substructure), a belonging probability to this area is performed using its BN. Finally, the different probabilities obtained are used to consolidate the form's class with the global network.

3. EXPERIMENTS

We experimented a database provided by the Actimage company. This is real data collected to be treated. It includes four form classes presented in figure 1, containing 800 samples per class. 600 forms are used for training and 200 for recognition using a cross validation method. Each network is trained on 4 different learning bases extracted from 3200 forms in the initial sample. The tests were performed in Matlab using the BNT toolbox [7]. From a global view, the results presented in tables 1, 2, 3 and figure 6 are encouraging.

Class	Heading area		Body area		Footer area		Global	
	R	P	R	P	R	P	R	P
1	99,38	56,68	98,5	98,5	99,25	37,62	96,88	80,2
2	99,7	83,86	66,88	66,88	98,5	74,7	91,8	99,6
3	0,13	25	89,13	89,23	2,25	0,5	98,75	95,98
4	99,02	97,66	99,21	98,88	0.13	0,13	75,63	98,77

Table 1. PC : Recall (R) and precision (P) in %

Class	Heading area		Body area		Footer area		Global	
	R	P	R	P	R	P	R	P
1	61,06	74,5	91,53	89,75	77,63	50,38	98,83	98,62
2	58,02	74,88	87,79	87,25	41,96	50,12	98,63	98,25
3	57,93	74,5	92,47	90,75	53,81	50,5	96,25	95,96
4	61,12	74,5	86,67	85,75	26,42	49,12	97,88	97,53

Table 2. MWST : Recall (R) and precision (P) in %

The overall recognition rate is 97.89 % with the MWST algorithm, 90.76 % with the PC algorithm and 94.97 % with

Class	Heading area		Body area		Footer area		Global	
	R	P	R	P	R	P	R	P
1	89,5	90,26	85,87	89,02	45,87	57,13	94,67	95,16
2	93	93,81	89,5	90,96	48,13	38,14	95,08	95,42
3	90,5	91,74	83,13	87,48	44,37	56,87	95,29	95,69
4	90,25	91,53	84,88	89,1	45,37	56,31	94,87	95,58

Table 3. Naive BN : Recall (R) and precision (P) in %

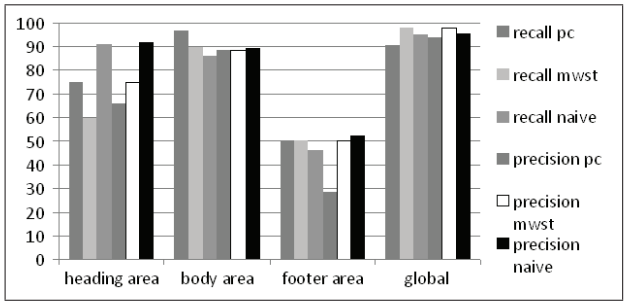

Fig. 6. Average for all classes

the naive BN. Regarding the results on form regions, the algorithm MWST gives more homogeneous results than the PC algorithm. Indeed, the recognition rate and accuracy are constant whatever the basis of learning. Concerning the naive BN, they give better results on regions with a large number of variables. They also get the best results for header pages. However, the PC algorithm can also achieve very good results in certain areas. This can be helpful in cases where a form is completed in several stages with intermediate treatment between each of these steps.

The naive BN give poorer results on the global recognition that the MWST algorithm. This can be explained by the loss interactions between variables complied with. The relationship between the fields are ignored.

Concerning the PC algorithm, we note that the accuracy rate of class 1 which is only 80.2% for the global recognition, is due to the complexity of its structure. Indeed, its fields are short, numerous and very close. The extraction step of the fields is strongly biased by this peculiarity as a text field from another class will cover several fields of class 1. For example, we note that the recall rates of the class footer 4 is only 0.13%. Similarly, we observe that the precision rate of the footer of the class 1 is only 37.62%. This is explained by the overlapping fields in two classes. The fields of class 4 *footer* are completely subsumed by the fields of class 1 *footer*. The matching is biased and the recognition of the form area are distorted. Nevertheless, the recognition rates of classes 1 and 4 are good, since the global network accepts the possibility that a class is defined by an area outside its original model. This problem is significantly mitigated by using the algorithm MWST or the naive BN.

The results show that the MWST algorithm is more efficient. By observing the BN, we can note that the algorithm PC isolates certain variables from the rest of the graph and thus limits the impact of certain fields. The naive BN lose the interaction between variables.

4. CONCLUSION

We have developed and tested a first approach for the classification of online and unconstrained forms by using two levels of BN. The approach exploits the conditional probabilities between area fields and strokes in the fields to find the more close form model. Early results are encouraging and pave the way for many opportunities. In the future, it would be interesting to validate the robustness of our system with a larger number of classes. Then, we plan to test the limits of the system about the direction sheet, and to modify the stroke matching approach by proceeding to a segmentation stage to reduce the matching errors.

Finally, the use of BN on forms could be a way to explore new strategies for filling them and thus allows us the modification of the layout and editing content of forms to adapt them to the writers.

Acknowledgment

This work is conducted under a CIFRE agreement. We would like to thank the Actimage company which collaborated in this work and has provided the necessary database.

5. REFERENCES

[1] S. Ramdane, B. Taconet, A. Zahour, and S. Kebairi, "A statistical method for an automatic detection of form types," *LNCS DAS*, vol. 1655/1999, pp. 84–98, 1999.

[2] Y. Xiang, *Probabilistic reasoning in multiagent systems: a graphical models approach*, Cambridge University Press, 2002.

[3] P. Leray and O. Francois, "Bnt structure learning package: documentation and experiments," Tech. Rep., Laboratoire PSI, INSA Rouen, 2004.

[4] L. Zhang Z. Sun and B. Zhang, "Online composite sketchy shape recognition based on bayesian networks," *LNCS*, vol. 4222/2006, pp. 506–515, 2006.

[5] L. Denoyer and P. Gallinari, "Bayesian network model for semi-structured document classification," *Inf. Process. Manage.*, vol. 40, no. 5, pp. 807–827, 2004.

[6] R. E. Neapolitan, *Learning Bayesian Networks*, Prentice Hall; illustrated edition, 2003.

[7] Kevin Murphy, "The bayes net toolbox for matlab," *Computing Science and Statistics*, vol. 33, 2001.