Relational Data Mining Techniques for Historical Document Processing

Michelangelo Ceci   Donato Malerba

Dipartimento di Informatica, Università degli Studi
via Orabona, 4 - 70126 Bari - Italy
{ceci, malerba}@di.uniba.it

Abstract. Document image understanding denotes the recognition of semantically relevant components in the layout extracted from a document image. Automatic approaches for document image understanding are highly demanded today by organizations involved in the preservation and valorisation of historical documents that collect more and more document images, whose effective usage critically depends on their fast and accurate indexing and cataloguing. In this context, Data Mining techniques can be profitably applied in order to support the user in the recognition of semantically relevant components in historical document images. However, such application is not straightforward and two important aspects have to be considered: First, extracted models should take into account the inherent spatial nature of the layout of a document image and spatial relations among layout components of interest. Second, low layout quality and standard of such a material introduces a considerable amount of noise in its description. For this reasons, in this paper, we investigate the application of a Statistical Relational Data Mining method, which successfully allows relations between components to be effectively and naturally represented by resorting to the Relational Data Mining framework and guarantees robustness to noise by exploiting statistical methods. Experiments are performed on two historical document corpora from the 20's and 30's.

1 Introduction

Advances in image acquisition and mass storage technology have noticeably increased the capability of collecting large volumes of document images in the form of compressed bitmaps. National archives and public records, libraries, organizations involved in the preservation of historical documents, produce more and more new document images, whose effective usage critically depends on their fast and accurate indexing and cataloguing.

Document image analysis is the subfield of digital image processing that aims at converting document images to symbolic form for modification, storage, retrieval, reuse and transmission [14]. This conversion is a complex process articulated into several stages. Initial processing steps include binarization, skew detection, noise filtering, segmentation. Then document image is decomposed into several constituent items which represent coherent components of the documents (e.g., text lines, half-tone images, line drawings or graphics), without any knowledge of the specific format. This layout analysis step precedes the interpretation or understanding of document images, whose aim is that of recognizing semantically relevant layout components (e.g. title, abstract of a scientific paper, picture of a newspaper).

Domain-specific knowledge appears essential for document image understanding: in the literature, there are no examples of attempts to develop a system that can interpret arbitrary documents [14]. The importance of knowledge representation and acquisition methods in the interpretation of document images has led some distinguished researchers to claim that document image understanding should be considered a branch of artificial intelligence [17]. In many applications presented in the literature, a great effort is made to hand-code the necessary knowledge according to some formalism, such as block grammars [13], geometric trees [7], and frames [19]. However, hand-coding domain knowledge is time-consuming and limits the application of document analysis systems to predefined classes of documents.

To alleviate the burden in developing and customizing document analysis systems, data mining methods can be profitably applied to extract the domain-specific knowledge required for effective document image understanding. In this paper we investigate the induction of classifiers that can be used to automatically recognize semantically relevant layout components. Classifiers are constructed from a set of training documents whose layout structures have already been interpreted by the users and described according to some representation formalism. Therefore, the customization of a document analysis system for a specific class of documents can be performed by extracting the layout structures of a set of training documents, by manually annotating them in order to specify the semantically relevant layout components (logical structures), and then by automatically inducing a set of accurate classifiers to be operationally used on a set of new, previously unseen, documents. In this way, human intervention is limited to annotating layout structures.

In the literature, several methods have been proposed for the construction of classifiers to be used in document image understanding [2, 3, 15, 10, 1]. However, most of them assume that training data are represented in a single table of a relational database, such that each row (or tuple) represents an independent example (a layout component) and columns correspond to properties of the example (e.g., height of the layout component). This single-table assumption [9], however, is too strong for at least three reasons. First, layout components cannot be realistically considered independent observations, because their spatial arrangement is mutually constrained by formatting rules typically used in document editing. Second, spatial relationships between a layout component and a variable number of other components in its neighborhood cannot be properly represented by a fixed number of attributes in a table. Even more so, the representation of properties of the other components in the neighborhood because different layout components may have different properties (e.g., the property “brightness” is appropriate for half-tone images, but not for textual components). Third, logical components, that is, the components of the logical structures, may be related to each other as well. For instance, the logical components ‘title’ and ‘author’ of a printed paper are often interrelated sequentially (the author follows the title). Since the single-table assumption limits the representation of relationships (spatial or non) between examples, it also prevents the discovery of this kind of pattern which can be very useful in document image understanding.

All these issues are ultimately due to the fact that document layout structures are a kind of spatial data and, as such, they are subject to spatial autocorrelation[1]. As already pointed out by Malerba et al. [11], methods investigated in (multi-)relational data mining (MRDM) [9] are the most suitable for spatial data, since they allow spatial relations between layout components to be effectively and naturally represented. MRDM approaches operate on data distributed in a set of tables (not a single one) and look for *relational patterns* that involve multiple tables from a relational database. Spatial relationships can be easily represented by means of a table and integrity constraints.

The limits of some methods reported in the literature on document image understanding and the recent developments in the field of MRDM motivate this work, whose main scope is that of evaluating an approach to classifier construction for document image understanding: namely, a statistical approach based on concepts and principles typical of MRDM. The system that implements the method developed according to this approach is Mr-SBC [6]. Mr-SBC presents some peculiarities such as high efficiency and the computation of a degree of confidence (a posterior probability) in the predicted class, which convey information on the potential uncertainty in classification. Moreover, Mr-SBC is particularly suitable for processing documents whose layout quality is often negatively affected. This is the case of the historical documents that are negatively affected by the presence of stamps, signatures, ink specks, manual annotations, and so on that overlap to those layout components involved in the understanding processes. In fact, it implements a statistical approach that is particularly robust to noise.

The paper is organized as follows. In the next section, the document processing phases are briefly described. In Sections 3, the method implemented Mr-SBC is described with reference to the specific application domain. In Section 4 experimental results on two distinct datasets are shown and conclusions are drawn.

2 Document Images Processing

In order to prove the effectiveness of Mr-SBC, it has been integrated in the document analysis system WISDOM++[2] [4], whose applicability has been investigated in the context of the IST-EU funded project COLLATE[3]. WISDOM++ permits the transformation of document images into XML format by means of several complex steps (see Figure 1):

1) Preprocessing of the raster image of a scanned paper document,
2) Segmentation of the preprocessed raster image into basic layout components
3) Classification of basic layout components according to the type of content (e.g., text, graphics, etc.),
4) Identification of a more abstract representation of the layout (layout analysis),

[1] In statistics, spatial autocorrelation indicates the fact that the effect of an explanatory (independent) or response (dependent) variable at any location may not be limited to the specific location.

[2] http://www.di.uniba.it/~malerba/wisdom++

[3] http://www.collate.de/

5) Classification of the document on the basis of its layout and content,
6) Identification of semantically relevant layout components (document image understanding),
7) Application of OCR only to those textual components of interest,
8) Storage in a relational database and generation of a document in XML format that conveys all information extracted in previous steps, including that on graphical rendering of the document on web browsers.

In the WISDOM++ context, document image understanding is limited to mapping the layout structure of a document into the corresponding logical structure, that is, abstract relationships between layout components are not extracted. The mapping can be performed by means of a set of classification rules, hence the need of automatically inducing some classifiers from data representing the layout of a set of training documents.

3 MRDM Approach to Document Image Understanding

Mr-SBC (Multi-Relational Structural Bayesian Classifier) extends to multi-relational data the naïve Bayesian classifier [8] originally defined for training data represented in a single table. The problem solved by the system can be formalized as follows:

Given:

- a training set of h relational tables $S=\{T_0,T_1,\ldots,T_{h-1}\}$ of a relational database D
- a set PK of primary key constraints on tables in S
- a set FK of foreign key constraints on tables in S
- a target relation $T\in S$
- a target discrete attribute y in T, different from the primary key of T, whose domain is $\{C_1, C_2, \ldots, C_r\}$

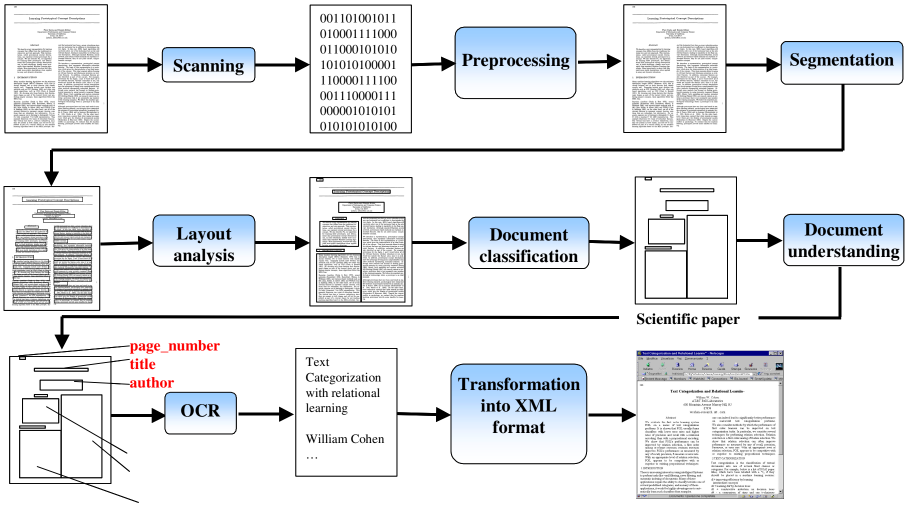

Fig.1 WISDOM++ steps

Find: a multi-relational naive Bayesian classifier which predicts the value of y for some individual represented as a tuple in T (with possibly UNKNOWN value for y) and related tuples in S according to FK.

The solution implemented by Mr-SBC is based on the idea that for an individual I to be classified it is possible to find a set R of first-order definite clauses that classifies I into one of the classes $\{C_1, C_2,\ldots, C_r\}$. The class $f(I)$ associated to I is that maximizing the posterior probability $P(C_i|R)$: $f(I)= arg\ max_i\ P(C_i|R)$
By applying Bayes theorem we have:

$$f(I)= arg\ max_i\ P(C_i|R) = arg\ max_i\ \frac{P(C_i)P(R|C_i)}{P(R)}$$

Since $P(R)$ is independent of the class C_i, it does not affect $f(I)$, that is,

$$f(I)= arg\ max_i\ P(C_i)P(R|C_i)$$

The construction of the set R is based on the notion of foreign key path.

Def. 1. *A foreign key path* is an ordered sequence of tables $\vartheta=(T_{i_1}, T_{i_2},\ldots, T_{i_s})$, s.t.

- ❑ $\forall$j=1, ...,s, $T_{i_j} \in S$
- ❑ $\forall$j=1..s-1, $T_{i_{j+1}}$ has a foreign key to the table T_{i_j}.

All predicates in definite clauses in R are binary and can be of two different types.

Def 2. A binary predicate p is a *structural* predicate associated to a table $T_i \in S$ if a foreign key in T_i exists that references a table $T_{i1} \in S$. The first argument of p represents the primary key of T_{i1} and the second represents the primary key of T_i.

Def 3. A binary predicate p is a *property* predicate associated to a table $T_i \in S$, if the first argument of p represents the primary key of T_i and the second argument represents another attribute value in T_i which is neither the primary key of T_i nor a foreign key in T_i.

Def 4. A first order definite clause associated to the *foreign key path* ϑ is a clause in the form: $p_0(A_1,y)$:- $p_1(A_1,A_2)$, $p_2(A_2,A_3)$, …, $p_{s-1}(A_{s-1},A_s)$, $p_s(A_s,c)$.
or $p_0(A_1,y)$:- $p_1(A_1,A_2)$, $p_2(A_2,A_3)$, …, $p_{s-1}(A_{s-1},A_s)$.
where

1. p_0 is a property predicate associated to the target table T and to the target attribute value y.
2. $\vartheta=(T_{i_1}, T_{i_2},\ldots, T_{i_s})$ is a *foreign key path* such that for each k=1, ..., s-1: p_k is a structural predicate associated to the table T_{i_k}
3. p_s is an optional property predicate associated to the table T_{i_s}.

Mr-SBC searches for all possible definite clauses R_j associated to foreign key paths of a user-defined maximum length and covering the individual I. Then, the probability

$$P(R \mid C_i) = P(\bigcap_{R_j \in R} R_j \mid C_i)$$

is computed by applying the naïve Bayes independence assumption on the minimal factor of the formula $\bigcap_{R_j \in R} R_j$. More details are reported in [6].

The relational nature of the probabilistic classification performed by Mr-SBC, makes the system suitable for the document image understanding domain, where classes C_i are logical labels that can be associated to layout components (individuals to be classified). In addition, tightly-coupling with a Relational DBMS (ORACLE$^{®}$ *10g*) allows Mr-SBC to work, by means of views, on the database used by WISDOM++ to store document data. In this way, Mr-SBC takes advantage of the database schema that provides useful knowledge of data model without asking the user to specify some background knowledge. The logical view that Mr-SBC has on the layout and logical structures of document images is reported in Figure 2.

The application of Mr-SBC to the document image understanding domain is not straightforward and requires some adjustments. First of all, it is necessary to modify the search strategy in order to allow cyclic paths. As observed by Taskar et al. [18], the acyclicity constraint hinders the representation of many important relational dependencies. This is particularly true in the task in hand, where a relation between two logical components is modelled by means of a table. For example, suppose that we need to model the relation *on_top* between two layout components, from a database point of view, this is realized by means of the table “block” and a table “on_top” that contains two foreign keys to the table “block”. In the original formulation of the problem solved by Mr-SBC, first-order classification rules do not consider the same table twice [6], therefore it is not possible to explore the search space by considering first the table “block”, then the table “on_top” and finally, again, the table “block”. To avoid this problem, we modified Mr-SBC, allowing cyclic paths. For this purpose, we considered a new definition of foreign key paths (*Def. 5*) at the cost of increasing the algorithm complexity:

Def. 5. *A foreign key path* is an ordered sequence of tables $\vartheta=(T_{i_1}, T_{i_2}, \ldots, T_{i_s})$, s.t.

$-\forall j=1, \ldots, s,\ T_{i_j} \in S$

$-\forall j=1..s-1,\ T_{i_{j+1}}$ has a foreign key to the table T_{i_j} or T_{i_j} has a foreign key to $T_{i_{j+1}}$

Thus examples of first order definite clause used by Mr-SBC are:

representative(A1,true) :- to_right(A1,A2), on_top(A2,A3)
assessors(A1,true) :- only_middle_row(A1,A2), height(A2, [50.985, +inf])

where only_middle_row(A1,A2) means that A1 and A2 are horizontally aligned when considering the barycentre.

The second adjustment concerns the classification of layout components. In document image understanding, it is possible that the same layout component is associated with two different logical labels. For instance, suppose that the layout analysis is not able to separate the page number and the running head of a scientific paper. Then a single layout component should be double labeled. More in general, the classifier should associate that component with multiple labels. For this reason, a binary classifier is built for each class, such that it discriminates examples assigned to that class from all the others. However, this leads to the problem of “unbalanced datasets”. In fact, data can be characterized by a predominant number of negative examples with respect to the number of positive examples.

Several solutions to the problem of the unbalanced datasets have been proposed in the literature. Some are based on a sampling of examples in order to have a balanced dataset [12]. Others are based on a different idea: a) for each class C_i, examples in the test set are ranked from the most probable member to the least probable member, b) for each test example, a correctly calibrated estimate of the true probability that it belongs to class C_i is computed [20], c) a probability threshold that delimitates the membership and the non-membership of a given test example to the class C_i is computed. This approach fits well our case, since the naive Bayesian classifier for two-class problems tends to rank examples well (even if the classifier does not return a correct probability estimate) [20]. In our solution, the threshold is determined by maximizing the AUC (Area Under the ROC Curve) [16] according to a cost function:

$$cost = P(C_i)\cdot(1\text{-}TP)\cdot c(\neg C_i; C_i) + P(\neg C_i)\cdot FP\cdot c(C_i; \neg C_i)$$

where $P(C_i)$ ($P(\neg C_i)$) is the prior probability that an example does (not) belong to the class C_i, $c(\neg C_i; C_i)$ ($c(C_i; \neg C_i)$) is the cost of classifying a positive (negative) example as negative (positive) for the class C_i, TP is the true positive rate and FP is the false positive rate. In the experiments reported in the next section, different values of $CostRatio = c(\neg C_i; C_i)/\ c(C_i; \neg C_i)$ have been considered.

4 Experimental results

In order to evaluate the effectiveness of Mr-SBC in understanding historical document images, it has been trained on two different datasets. Both datasets have been provided by two distinct European film archives, namely Deutsches Filminstitut (DIF) and Filmarchiv Austria (FAA) in the context of the UE funded project COLLATE (http://www.collate.de/). In both datasets, documents represent rare historic film censorships from the 20's and 30's. In the DIF dataset, documents are generally composed by two pages and the user manually labeled 149 layout

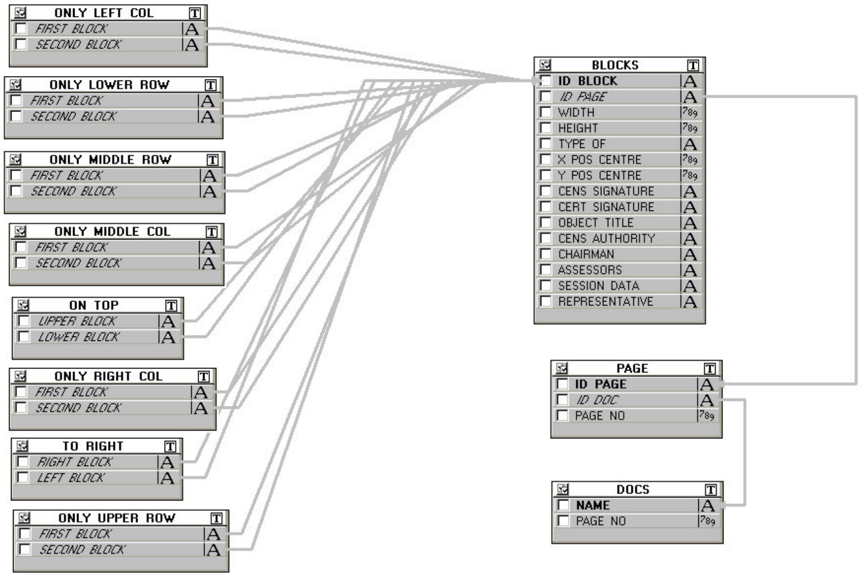

Fig. 2. Logical view of the database input to Mr-SBC.

components out of 950 components in all. The FAA dataset is composed by one page documents and the user manually labeled 140 layout components over 503 components in all. The components that have not been labeled are "irrelevant" for the task in hand or are associated to "noise" blocks: they are automatically considered *undefined*. In both datasets, the percentage of *undefined* components is relatively high. This is mainly due to the presence of ink specks, holes stamps and manual annotations that negatively affect the layout analysis. Table 1 reports logical labels considered in this analysis while Figure 3 represents an example of labeled document image. The performance of the learning tasks is evaluated by means of a 5-fold cross-validation on all datasets, that is, for each dataset, the set of documents is first randomly divided into five folds, and then, for every fold Mr-SBC is trained on the remaining folds and tested on the hold-out fold. In Tables 2 and 3 a brief description of the datasets is reported.

For each learning problem, the number of omission/commission errors is recorded. An omission error occurs when a logical labelling of layout components is missed, while a commission error occurs when a wrong logical labelling is "recommended" by the classifier. In our study we do not consider the standard classification accuracy, because for each learning task, the number of positive and negative examples is strongly unbalanced and, in most cases, the trivial classifier that returns always "undefined" would be the classifier with the best accuracy.

Table 1 Considered logical labels for each dataset.

Source	**Labels**
DIF	*cens_signature, cert_signature, object_title, cens_authority, chairmen, assessors, session_data, representative*
FAA	*dep_signature, adhesive_stamp, stamp, registration_au, date_place, department, applicant, reg_number, film_length, film_producer, film_genre, film_title*

Table 2 DIF Dataset description: Distribution of pages and examples per document grouped by 5 folds.

Fold No.	No. of Documents	No. of pages	No. of labeled components	Total No. of components
1	5	8	28	200
2	5	9	30	196
3	5	9	33	201
4	5	8	25	152
5	5	9	33	201
Total	25	43	149	950

Table 3 FAA Dataset description: Distribution of pages and examples per document grouped by 5 folds.

Fold No.	Document name	No. of pages	No. of labeled components	Total No. of components
1	5	5	34	125
2	5	5	29	115
3	5	5	36	93
4	5	5	26	113
5	5	5	15	57
Total	25	25	140	503

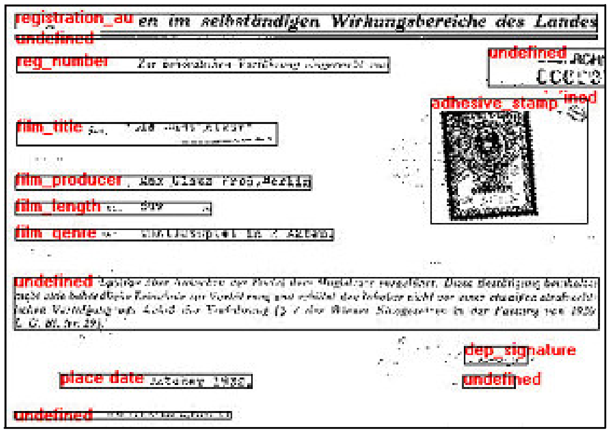
registration_au
en im selbständigen Wirkungsbereiche des Landes
undefined
reg_number
undefined
adhesive_stamp
ined
film_title
film_producer
Max Glass Prod, Berlin
film_length
film_genre
undefined
dep_signature
place date
undefined
undefined

Fig.3 A processed document image: an example of a FAA censorship card.

Henceforth, experimental results are commented for each dataset. They are reported in the order of quality of the extracted layout. In both cases, layout extraction is difficult because of the presence of stamps, signatures, ink specks, manual annotations, and so on, that overlap to those layout components involved in the understanding process. Nonetheless, there is a clear difference between DIF and FAA censorship cards, the former being better structured and less noisy. Therefore, the following ordering of datasets is defined on the basis of layout quality: DIF and FAA.

For DIF dataset, Table 4 reports results of Mr-SBC for different values of *CostRatio* when n=2 and n=3, where n is the number of predicates in the body of a first order definite clause associated to a *foreign key path* (see Def. 4). By increasing *CostRatio*, more importance is given to the cost of false negative $c(\neg C_i; C_i)$ and the ability of the classifier to correctly classify positive examples increases, while the precision of the classification decreases because negative examples are erroneously classified as positive. This behavior is somehow expected and occurs also in the case of FAA documents. What is really surprising is the fact that moving from n=2 to n=3, the number of commission errors noticeably increases. This means that while probability values returned by Mr-SBC for positive and negative examples of each logical components are well separated when n=2, they become closer when n=3, thus causing some problems to the automated threshold determination procedure. In other words, if we rank each positive/negative example of a logical component according to the probability value returned by Mr-SBC, we can observe that Mr-SBC ranks testing examples well (see Figure 4) thus allowing a clear separation among positive and

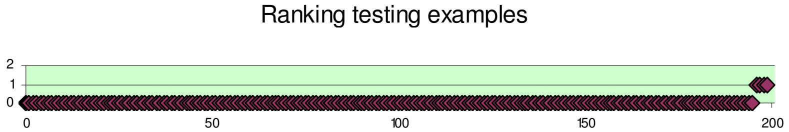

Fig.4 Ranking of the 200 testing examples of logical component *Cens_Signature* for the DIF dataset. Each point represents a testing example E. Examples are ranked on the basis of *P(Cens_Signature = 'true' | E)*. Positive examples are reported above (Y=1) while negative examples are reported below (Y=0). Results are obtained by training Mr-SBC on folds 2,3,4,5 and testing on fold 1 with *CostRatio*=10 and n=2.

negative examples, while the ranking distribution is less skewed when *n*=3.

Mr-SBC with *n*=2 and *CostRatio*∈{1,10} seems to offer a good trade-off between recall and precision for this dataset.

For FAA dataset, (Table 5), results confirm initial observation on the complexity of this learning task because of the poor layout structure of many censorship cards. Mr-SBC presents high percentage of commission errors for *n=3*. This is mainly due to the inherent complexity of the task due to the poor layout structure extracted.

Finally, Mr-SBC time complexity strongly depends on the value of *n* (see Table 6) this depends on the number of classification queries (i.e. probabilities) performed (estimated) (see Table 7). The high number of probabilities to be estimated leads to a leveling out of probabilities. This phenomenon can be a cause of the low performance of Mr-SBC observed when *n*=3. Indeed, as observed by Bellman [5], the number of examples should increase exponentially with the number of features to maintain a given level of accuracy ("curse of dimensionality"). A possible solution to this problem can be in pruning clauses that are not useful in discriminating among classes.

5 Conclusions

In this paper, we presented the application of the (multi-)relational statistical data mining approach Mr-SBC in the task of document image understanding from a particular kind of documents, namely historical documents. By resorting to a statistical relational data mining approach it is possible to face two important sources of complexity coming from the task in hand. First, it is possible to take into account the inherent spatial nature of the layout of a document image and spatial relations among layout components of interest. This is obtained by exploiting the relational data mining framework that allow spatial relations between layout components to be effectively and naturally represented. Second, it is possible to work with noisy data, where noise is due to the low layout quality and standard. This is obtained by exploiting robustness of statistical methods to noise.

Experiments have been performed on real world datasets provided by two different European film archives, namely Deutsches Filminstitut (DIF) and Filmarchiv Austria (FAA). Results show good performances of Mr-SBC on both datasets when the complexity of the extracted model is quite low (*n*=2). When Mr-SBC is asked to consider more complex properties, the performance degenerates due to problems associated to the high dimensionality of the search space (*n*=3). Thanks to the tight-coupling of the data mining system with an OR-DBMS, learning times are kept under control. As future work, we intend to propose a method fro pruning involved rules, to extend features by including textual features associated to layout components and we intend to compare our Multi-relational approach with an approach based on a propositionalization process [9] that transforms multiple relational tables in a single table thus allowing the application of traditional data mining methods.

Acknowledgments

The work presented in this paper is partial fulfillment of the research objective set by the project Ateneo2006 'Metodi di scoperta di conoscenza per ubiquitous computing'.

Table 4 Average number of omission errors over positive examples and commission errors over negative examples on DIF dataset. MrSBC results are obtained varying *CostRatio* in {1,10,20,30}and *n* in {2,3}.

	Folds	**No. Ex**	**Mr-SBC (*n*=2)**				**Mr-SBC (*n*=3)**			
			1	10	20	30	1	10	20	30
Omission Errors	1	28	11	4	2	2	12	7	6	5
	2	30	9	6	6	4	9	5	5	5
	3	33	13	4	4	4	11	7	3	2
	4	25	6	5	4	4	15	8	7	6
	5	33	14	5	5	5	13	11	8	7
			35.02%	16.31%	14.08%	12.75%	41.12%	25.64%	19.89%	17.16%
Commission Errors	1	1372	3	15	23	23	392	419	421	425
	2	1342	13	28	28	52	581	613	620	620
	3	1347	4	13	13	13	588	599	615	627
	4	1039	6	18	28	42	162	204	221	243
	5	1374	6	18	18	18	399	437	443	447
			0.50%	1.43%	1.74%	2.37%	31.86%	34.25%	35.03%	35.75%

Table 5 Average number of omission errors over positive examples and commission errors over negative examples on FAA dataset. Results are obtained with *CostRatio* in {1,10,20,30}and *n* in {2,3}.

	Folds	**No. Ex**	**Mr-SBC (*n*=2)**				**Mr-SBC (*n*=3)**			
			1	10	20	30	1	10	20	30
Omission Errors	1	34	15	7	6	5	3	1	0	0
	2	29	13	9	9	8	2	0	0	0
	3	36	16	10	9	9	0	0	0	0
	4	26	14	9	8	8	1	1	1	1
	5	15	6	6	5	5	3	3	3	3
			45.45%	30.80%	27.56%	26.28%	7.91%	5.36%	4.77%	4.77%
Commission Errors	1	1466	249	265	275	293	1103	1107	1108	1108
	2	1351	236	259	264	282	1018	1020	1020	1020
	3	1080	181	193	209	212	901	901	901	901
	4	1330	5	30	39	79	998	999	999	999
	5	669	63	85	88	104	401	419	419	427
			12.20%	14.02%	14.75%	16.39%	73.80%	74.44%	74.45%	74.69%

Table 6 Average learning times. Results are expressed in seconds. Results are obtained with *CostRatio*=10.

	MrSBC (*n*=2)	MrSBC (*n*=3)
DIF	94.7	774.4
FAA	108.3	736.1

Table 7 Complexity of the induced model. Results are obtained with *CostRatio*=10.

	DIF	FAA
Mr-SBC - No. classification queries (n=2)	1,265	1,287
Mr-SBC - No. classification queries (n=3)	16,093	8,025
No. classes	8	12

References

[1] Aiello M., Monz C., Todoran L., Worring M. (2002): Document Understanding for a Broad Class of Documents. International Journal of Document Analysis and Recognition IJDAR 5(1), 1-16.

[2] Akindele O.T., Belaïd A. (1995). Construction of generic models of document structures using inference of tree grammars. Proceedings of the 3rd ICDAR. 206-209.

[3] Allen J.F. 1983. Maintaining knowledge about temporal intervals. Communications of the ACM, 26(11). 832-843.

[4] Altamura O., Esposito F., Malerba D. (2001). Transforming paper documents into XML format with WISDOM++. International Journal on Document Analysis and Recognition IJDAR 4(1), 2-17.

[5] Bellman R.E. (1961). Adaptive Control Processes. Princeton University Press.

[6] Ceci M., Appice A., Malerba D. (2003). Mr-SBC: a Multi-Relational Naive Bayes Classifier. Principles and Practice of Knowledge Discovery in Databases, European Conference, PKDD vol 2838 of LNAI Springer-Verlag, 95–106.

[7] Dengel A., Bleisinger R., Hoch R., Fein F., Hones F.. (1992). From Paper to Office Document Standard Representation. Computer, vol. 25, no. 7. 63-67.

[8] Domingos P., Pazzani M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. Machine Learning 29(2-3). 103–130.

[9] Dzeroski S., Lavrac N. (2001). Relational Data Mining. Springer-Verlag, Berlin Germany.

[10] Le Bourgeois F., Souafi-Bensafi S., Duong J., Parizeau M., Coté M., Emptoz H. 2001. Using statistical models in document images understanding. Workshop on Document Layout Interpretation and its Applications, DLIA.

[11] Malerba D. and Lisi F.A. (2001). Discovering Associations Between Spatial Objects: An ILP Application, in C. Rouveirol & M. Sebag (Eds.), Inductive Logic Programming, Lecture Notes in Artificial Intelligence, 2157, Springer, Berlin, Germany. 156-163.

[12] Mladenic D., Grobelnik M. (1999). Feature selection for unbalanced class distribution and naive bayes. Proc. of the 16th International Conference on Machine Learning. 258–267.

[13] Nagy, G., Seth, S.C., Stoddard, S.D. (1992). A Prototype Document Image Analysis System for Technical Journals. IEEE Computer, 25(7). 10-22.

[14] Nagy G. 2000. Twenty Years of Document Image Analysis in PAMI, IEEE Trans. PAMI-22, 1, invited contribution to Anniversary Issue. 38-62.

[15] Palmero G.I.S., Dimitriadis Y.A. (1999). Structured Document Labeling and Rule Extraction using a New Recurrent Fuzzy-neural System. International Journal of Document Analysis and Recognition IJDAR. 181-184.

[16] Provost F., Fawcett T. 2001. Robust classification for imprecise environments. Machine Learning 42(3). 203–231.

[17] Tang Y.Y., Yan C. D., Suen C. Y. (1994). Document Processing for Automatic Knowledge Acquisition, in IEEE Trans. on Knowledge and Data Engineering, 6(1). 3-21.

[18] Taskar B., Abbeel P., Koller D. 2002. Discriminative probabilistic models for relational data. Proc. of Int. Conf. on Uncertainty in Artificial Intelligence. 485-492.

[19] Wenzel C., Maus H. (2001). Leveraging corporate context within knowledge-based document analysis and understanding. International Journal on Document Analysis and Recognition, Vol 3 Issue 4. 248-260.

[20] Zadrozny B. Elkan C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. Proc. of the 18th International Conference on Machine Learning ICML. 609–616.