Document Image Understanding through Iterative Transductive Learning

Michelangelo Ceci, Corrado Loglisci, Lucrezia Macchia,
Donato Malerba, and Luciano Quercia

Dipartimento di Informatica, Università degli Studi di Bari "Aldo Moro"
{ceci,loglisci}@di.uniba.it,
{donato.malerba,lucrezia.macchia}@uniba.it, luciano.quercia@gmail.com

Abstract. In Document Image Understanding, one of the fundamental tasks is that of recognizing semantically relevant components in the layout extracted from a document image. This process can be automatized by learning classifiers able to automatically label such components. However, the learning process assumes the availability of a huge set of documents whose layout components have been previously manually labeled. Indeed, this contrasts with the more common situation in which we have only few labeled documents and abundance of unlabeled ones. In addition, labeling layout documents introduces further complexity aspects due to multi-modal nature of the components (textual and spatial information may coexist). In this work, we investigate the application of a relational classifier that works in the transductive setting. The relational setting is justified by the multi-modal nature of the data we are dealing with, while transduction is justified by the possibility of exploiting the large amount of information conveyed in the unlabeled layout components. The classifier bootstraps the labeling process in an iterative way: reliable classifications are used in subsequent iterative steps as training examples. The proposed computational solution has been evaluated on document images of scientific literature.

1 Introduction

The recognition of semantically relevant components in the layout extracted from a document image is based on domain-specific knowledge, which is represented in very different forms (e.g. formal grammars or production rules). Several prototypical document image understanding systems have been developed by manually encoding the required knowledge (e.g., DeLoS [14]). However, the layout of documents, even for the same publisher, may change considerably. To prevent obsolescence of the developed systems, it is necessary to continuously update the required knowledge, which is unfeasible if based only on manual encoding.

In order to guarantee *versatility* of Document Image Analysis Systems [1], that is, guarantee competence over a broad and precisely specified class of document images, the application of machine learning methods has been investigated for almost two decades [10][5]. Operatively, a human operator provides a document image analysis system with images of documents and then detects and labels

M. Agosti et al. (Eds.): IRCDL 2012, CCIS 354, pp. 117–128, 2013.

semantically relevant layout components from which document structures are induced. This *supervised* learning approach, though providing some flexibility, still does not ensure the key requirement of versatility. Indeed, to acquire the necessary knowledge on a really broad class of documents, supervised learning methods may require a large set of labeled documents. This contrasts with the common situation in which only few labeled training documents are available due to the significant cost of manual annotation. Therefore, it is important to exploit the large amount of information potentially conveyed by unlabeled documents.

Two main settings have been proposed in the literature to exploit information contained in both labeled and unlabeled data: the *semi-supervised* setting and the *transductive* setting [15]. The former is a type of inductive learning: the learned function is used to make predictions on any possible example. The latter is only interested in making predictions for the given set of unlabeled data. When the set of documents to label is known a priori, the transductive setting is more suitable, since it is an easier problem than (semi-supervised) induction. In this paper, we propose a transductive approach where unlabeled documents are used to reprioritize models learned from labeled documents alone. Indeed, while discriminative learning methods base their decisions on the posterior probability $p(y|x)$, the transductive learning method uses unlabeled documents to improve the estimate of the prior probability $p(x)$, and hence correct the posterior probability $p(y|x)$ by assuming some form of dependence with $p(x)$.

The proposed learning method follows a logic-based approach in which models are represented by a set of rules expressed in relational logic and documents are represented as facts in the same formalism. So, to "understand" the layout structure of an unlabeled document, rules are matched against the relational description of the document layout. The relational representation of document layout and rules is motivated by the fact that layout objects can be related by a number of spatial relationships, such as distance, directional or topological relationships. The study of relational learning in a transductive setting has received little attention (see [4], [11], [13]) while the application of transductive relational learning to bootstrap the labeling process of document image collections remains still unexplored. This work extends the research reported in [6], by introducing an iterative bootstrapping framework and by extending empirical evaluation to additional datasets. In the iterative bootstrapping framework, at each iteration, the algorithm expands the training set by including (originally unlabeled) examples for which the classification is considered to be reliable.

The paper is organized as follows. In Section 2, we define the problem to be solved. Sections 3 and 4 are devoted to the presentation of the method. Finally, experimental results are reported in Section 5 and some conclusions are drawn.

2 Motivations and Problem Definition

The recognition of semantically relevant layout components in document images is part of a complex transformation process of document images into a structured symbolic form. This transformation is articulated into several steps. Initial

processing steps include binarization, skew detection, and noise filtering. Then, the document image is segmented into several layout components, such as text lines, half-tone images, line drawings or graphics (this step is called layout analysis). The interpretation or understanding of document images follows layout analysis. It aims to associate a logical label (e.g. title, abstract of a scientific paper, picture of a newspaper) to semantically relevant layout components, as well as to extract relevant relationships between logical components (e.g., reading order). Document image understanding is typically based on layout information, such as the relative positioning of layout components or the size of layout components, as well as on content information (e.g., textual, graphical). This is the case of the work reported in this paper, where the association of logical labels to layout components is based on both layout information and textual information. However, the novelty here is mainly in the strategy applied to learn a classifier which can be used to recognize semantically relevant components.

In this work we investigate this issue and propose a trasductive method for learning classifiers from training data represented in relational formalism. In a formal way, the problem is defined as follows:
Given:

- a database schema SC which consists of a set of h relational tables $\{T_0, \ldots, T_{h-1}\}$, a set PK of primary keys on the tables in SC, and a set FK of foreign key constraints on the tables in SC,
- a target relation $T \in SC$ (that represents layout components) and a target discrete attribute Y in T, different from the primary key of T, whose domain is the finite set $\{C_1, C_2, \ldots, C_L\}$ (Logical label),
- the projection T' of T on all attributes of T except Y,
- a training (working) set that is an instance TS (WS) of the database schema SC with known (unknown) values for Y;

Find: the most accurate classification of Y for examples in WS.

In this work, the classification of Y is based on an approach that exploits both the relational data mining setting and the classical Naïve Bayesian framework.

More precisely, given an object $E \in WS$ to be classified, a classical naïve Bayes classifier assigns E to the class C_i that maximizes the *posterior probability* $P(C_i|E)$. By applying the Bayes theorem, $P(C_i|E)$ is expressed as follows:

$$P(C_i|E) = P(C_i) \cdot P(E|C_i) \ / \ P(E). \tag{1}$$

In fact, the decision on the class that maximizes the posterior probability can be made only on the basis of the numerator, that is $P(C_i) \cdot P(E|C_i)$, since $P(E)$ is independent of the class C_i. The probability $P(C_i|E)$ can then be used to identify examples E for which the classification is reliable. This property can be used to iteratively extend the training data by propagating the most reliable decisions when bootstrapping the labeling process.

In (1), the main problem is in the computation of $P(E|C_i)$. By following the main intuition in [2], it is possible to consider a set $\Re$ of association rules to define a suitable decomposition of the likelihood $P(E|C_i)$ *à la* naive Bayes in

order to simplify the probability estimation problem. In particular, if $\Re(E) \subseteq \Re$ is the set of first order association rules whose antecedent covers E, $P(E|C_i)$ is:

$$P(E|C_i) = P(\bigwedge_{R_j \in \Re(E)} antecedent(R_j)|C_i). \quad (2)$$

The straightforward application of the naïve Bayes independence assumption to all literals in $\bigwedge_{R_j \in \Re(E)} antecedent(R_j)$ is not correct, since it may lead to underestimating $P(E|C_i)$ when several similar clauses in $\Re(E)$ are considered for the class C_i. To prevent this problem the authors resort to the logical notion of factorization. Details are reported in [2].

Although this approach would potentially be used in this application, two main limitations could prevent its actual applicability: *i)* It does not exploit the transductive learning setting. *ii)* As in most associative classifiers, extracted association rules do not permit to adequately characterize classes.

To overcome these limitations, in this paper, we use Emerging Patterns (EPs) instead of association rules in order to discover a characterization of classes and we use this characterization in a transductive classifier. In fact, emerging patterns discovery is a descriptive data mining task which aims at detecting significant differences between objects of distinct classes. EPs are introduced in [8] as a particular kind of patterns (or multi-variate features) whose support significantly changes from one data class to another: the larger the difference of pattern support, the more interesting the pattern. Change in pattern support is estimated in terms of the support ratio (or *growth rate*). EPs with sharp change in support (high growth rate) can be used to characterize classes.

3 Mining Emerging Patterns with SPADA

Data mining research has provided several solutions (e.g.[8]) for the task of emerging patterns discovery but only one attempt [3] has been done to deal with relational data. In this work, we exploit the system SPADA [12], originally designed for *relational* frequent patterns discovery, for mining emerging patterns.

SPADA represents relational data *à la* Datalog, a logic programming language with no function symbols specifically designed to implement deductive databases. SPADA distinguishes between the set S of *reference* (or target) *objects*, which are the main subject of analysis, and the sets R_k, $1 \leq k \leq m$, of *task-relevant* (or non-target) objects, which are related to the former and can contribute to account for the variation. From a database viewpoint, S corresponds to the target table $T \in SC$ and each R_k corresponds to a different relational table $T_i \in SC$. A unit of analysis corresponds to a tuple in $t \in T$ and to all tuples in the database related to t according to foreign key constraints.

In the following sub-sections, the document description and the learning strategy are described, as it has been modified to mine emerging patterns.

Document Description. In the logic framework adopted by SPADA, a relational database is boiled down into a deductive database where properties of

Table 1. The complete list of used predicates

Layout structure	Locational features	*x_pos_center*/2
		y_pos_center/2
	Geometrical features	*height*/2
		width/2
	Topological features	*on_top*/2
		to_right/2
	Aspatial feature	*type_of*/2
Logical structure	Logical features	application dependent (e.g., *abstract/1*)
Text	Textual features	application dependent (e.g., *text_in_abstract/2*)

both reference objects (which are the main subject of the analysis) and task-relevant objects (which are relevant for the task at hand, but not necessarely the main subjects of the analysis) are represented in the extensional part D_E, while the domain knowledge is expressed as a normal logic program which defines the intensional part D_I. As an example, we report a fragment of the extensional part of a deductive database D which describes multimodal information which can be extracted from any document image:

block(b1). block(b2). ... height(b2,[11..54]). width(b1,[7..82]). ...
on_top(b2,b1). ... on_top(b2,b3). ... part_of(b1,p1). part_of(b2,p1). page_first(p1).
... abstract(b1). title(b2). ... text_in_abstract(b1,'base'). text_in_title(b2,'model')....

In this example, $b1$ and $b2$ are two constants which denote as many distinct layout components (reference objects), while $p1$ denotes a document page (task-relevant object). Predicate *block* defines a layout component, *part_of* associates a block to a document page, *height* and *width* describe geometrical properties of layout components, *on_top* expresses a topological relationship between layout components, $page\_first(p1)$ refers to the position of the page in the document, *abstract* and *title* associate $b1$ and $b2$ with a logical label, *text_in_abstract* and *text_in_title* describe the textual content of the logical components.

The complete list of predicates is reported in Table 1. The aspatial feature *type_of* specifies the content type of a layout component (e.g. image, text, horizontal line). Logical features are used to associate a logical label to a layout object and depend on the specific domain. In the case of scientific papers (considered in this work), possible logical labels are: *affiliation*, *page_number*, *figure*, *caption*, *index_term*, *running_head*, *author*, *title*, *abstract*, *formulae*, *subsection_title*, *section_title*, *biography*, *references*, *paragraph*, *table*. Textual content is represented by means of another class of predicates, which are true when the term reported as second argument occurs in the layout component denoted by the first argument. Terms are automatically extracted by means of a text-processing module[7].

The Mining Step. The original algorithm of SPADA mines frequent patterns at multiple levels l of granularity in order to properly deal with hierarchies H_k of objects. When these are available, it is important to take them into account since patterns involving more abstract objects are better supported (although less precise). SPADA operates in two steps for each granularity level: i) pattern generation; *ii)* pattern evaluation. It takes advantage of statistics computed at granularity level l when computing the supports of patterns at the granularity

level $l+1$. To discover emerging patterns, SPADA has been modified to mine patterns which characterize classes by detecting significant differences between the objects of these classes. This problem requires the following formulation:
Given:

- a set S of *reference objects*,
- a label value $y \in Y = \{C_1, C_2, \ldots, C_L\}$ associated to each reference object,
- some sets R_k, $1 \leq k \leq m$, of *task-relevant objects*,
- a background knowledge BK including hierarchies H_k on objects in R_k,
- M granularity levels in the descriptions,
- a set of granularity assignments Ψ_k which associate each object in H_k with a granularity level,
- a couple of sets of thresholds $minSup[l]$, $minGR[l]$ for each granularity level,
- a language bias LB that constrains the search space;

Find: A set of multilevel emerging patterns $\{F | supp_{C_i}(F) \geq minSup[l], GR_{C_i}(F) \geq minGR[l]\}$.

In this formulation, $supp_{C_i}(F)$ represents the support of the pattern F in the subset of reference objects labeled with C_i while the growth rate $GR_{C_i}(F)$ is defined as: $GR_{C_i}(F) = \frac{supp_{C_i}(F)}{supp_{\neg C_i}(F)}$ where $supp_{\neg C_i}(F)$ is the support of the pattern F in the subset of reference objects labeled with $c \in \{C_1, \ldots, C_{i-1}, C_{i+1}, \ldots C_L\}$.

To efficiently mine frequent patterns, SPADA prunes the search space by exploiting the monotonicity of the support. Let F' be a refinement of a pattern F (i.e. F' is more specific that F). If F is an infrequent pattern for the class C_i (i.e. $supp_{C_i}(F) < minSup$), then also $supp_{C_i}(F') < minSup$. This means that F' cannot be an emerging pattern that distinguishes C_i from $\neg C_i$. Hence, SPADA does not refine patterns which are infrequent in C_i.

Unluckily, the monotonicity property does not hold for the growth rate: a refinement of an emerging pattern whose growth rate is lower than the threshold $minGR$ may or may not be an EP. However, also in this case, it is possible to prune the search space. According to [16], we modified the mining algorithm originally developed in SPADA in order to avoid to generate the refinements of a pattern F in the case that $GR_{C_i}(F) = \infty$ (i.e., $supp_{C_i}(F) > 0$ and $supp_{\neg C_i}(F) = 0$). Indeed, due to the monotonicity of support, for each pattern F' obtained as refinement of F: $supp_{C_i}(F) \geq supp_{C_i}(F')$ then $supp_{C_i}(F') = 0$. Thereby, $GR_{C_i}(F') = 0$ in the case that $supp_{C_i}(F') = 0$, while $GR_{C_i}(F') = \infty$ in the case that $supp_{C_i}(F') > 0$. In the former case, F' is not worth to be considered. In the latter case, we prefer F to F' based on the Occams razor principle, according to which all things being equal, the simplest solution tends to be the best one (F has the same discriminating ability than F').

In our application domain, reference objects are all the logical components for which a logical label is specified. Task relevant objects are all the logical components (including undefined components) as well as pages and documents. The BK is used to specify the hierarchy of logical components (Figure 1). The BK also allows us to automatically associate information on page order to layout components, since the presence of some logical components may depend on the page order (e.g. author is in the first page).

article
+ − − heading
| + −− identification
| | + −− (title, author, affiliation)
| + −− synopsis
| + −− (abstract, index_term)
+ − − content
| + −− final components
| | + −− (biography, references)
| + −− body
| + −− (section_title, subsect_title, paragraph, caption, figure, formulae, table)
+ − − page_component
| + −− running_head
| + −− page_number
+ − − undefined

Fig. 1. Hierarchy of logical components

Algorithm 1. The iterative transductive learning algorithm.

Input: TS training data, WS working data.
Output: H working objects associated with labels
1: $H \leftarrow \emptyset$; $W' \leftarrow WS$;
2: **while** $WS \neq \emptyset$ **do**
3: Compute the score matrix $\Xi = [score_{TS \cup H}(o_j, C_i)]_{o_j \in W',\ C_i \in Y}$;
4: Sort the objects in $o_j \in W'$ according to $\max_{C_i}(score_{TS \cup H}(o_j, C_i))$;
5: Add all $\langle o_j, \arg\max_{C_i}(score_{TS \cup H}(o_j, C_i)) \rangle$ to H, where o_j is one of the top $\lfloor |WS|/k \rfloor$ objects in W';
6: Remove the top $\lfloor |WS|/k \rfloor$ objects from W';
7: **end while**

4 Transductive Classification

The transductive classifier implemented in our proposal is described in Algorithm 1, where at each iteration of the cycle at line 3, the algorithm labels objects belonging to the working set WS and uses a subset of them of size $\lfloor |WS|/k \rfloor$ as training objects in the subsequent iteration, where k is a user defined parameter[1]. The subset is created according to the function $score_{TS \cup H}(o_j, C_i)$ which represents a membership score of an object o_j to the class C_i. This score is a growth rate based function which is estimated on the current training set $TS \cup H$ and is computed by adapting the EP-based classifier CAEP [9] to the relational setting. The largest score determines the object's class.

In our case, it is computed on the basis of the subset of relational emerging patterns that cover the object to be classified. Formally, let o_j be the description of the object to be classified (an object is represented by a tuple in the target table and all the tuples related to it according to foreign key constraints), $\Re(o_j) = \{F \in \Re | \exists \theta\ F\theta \subseteq o_j\}$ is the set of emerging patterns that cover the object o_j.

[1] This means that there are, at most, $k + 1$ iterations.

The score of o_j on the class C_i is computed as follows:

$$score_{TS\cup H}(o_j, C_i) = \sum_{F\in\Re(o_j)} \frac{GR_{C_i}(F)}{GR_{C_i}(F)+1} sup_{C_i}(F) \tag{3}$$

where $GR_{C_i}(F)$ and $sup_{C_i}(F)$ are computed on the current training set $TS\cup H$.

This measure may result in an inaccurate classifier in the case of unbalanced datasets that is, when training objects are not uniformly distributed over the classes. In order to mitigate this problem the authors in [9] proposed to normalize this score on the basis of the median of the scores obtained from training objects belonging to C_i. This results in the following classification function:

$$class_{TS\cup H}(o_j) = \arg\max_{C_i} \frac{score_{TS\cup H}(o_j, C_i)}{median_{ro\in TS\cup H}(score_{TS\cup H}(ro, C_i))} \tag{4}$$

where $TS \cup H$ represents the training set.

However, in our case, the main problem comes from the different number of EPs that are extracted from different classes. This means that, in our case a different normalization that weights the number of EPs is necessary:

$$score_{TS\cup H}(o_j, C_i) = \frac{1}{|\Re(o_j)|} \sum_{F\in\Re(o_j)} \frac{GR_{C_i}(F)}{GR_{C_i}(F)+1} sup_{C_i}(F) \tag{5}$$

Since $sup_{C_i}(F)$ represents the probability that a reference object belonging to class C_i is covered by F, Equation (5) can be transformed as follows:

$$score_{TS\cup H}(o_j, C_i) = \frac{1}{|\Re(o_j)|} \sum_{F\in\Re(o_j)} \frac{GR_{C_i}(F)}{GR_{C_i}(F)+1} P(F|C_i) \tag{6}$$

By applying the Bayes theorem:

$$score_{TS\cup H}(o_j, C_i) = \frac{1}{|\Re(o_j)|} \sum_{F\in\Re(o_j)} \frac{GR_{C_i}(F)}{GR_{C_i}(F)+1} \frac{P(C_i|F)}{P(C_i)} \times P(F) \tag{7}$$

where $P(C_i|F)$ can be estimated as the percentage of objects covering F in $TS\cup H$ that belong to C_i. $P(C_i)$ can be estimated as the percentage of objects in $TS\cup H$ that belong to C_i. Finally, $P(F)$ is the percentage of objects covering F. According to the transductive learning setting, this factor is estimated by considering the whole set of objects ($TS \cup WS$). This would provide a more reliable estimation of $P(F)$ (since obtained from a larger population of objects potentially coming from the same distribution).

$$P(F) = \frac{\#\{ro|ro \in TS\cup WS, \exists\theta\ F\theta \subseteq ro\}}{\#\{ro|ro \in TS\cup WS\}} \tag{8}$$

5 Experiments

The proposed approach has been applied to three different real-world datasets consisting of articles published in two international journals, namely IEEE TPAMI and Behavior Genetics (BG), and in the proceedings of the International Conference on Machine Learning (ICML). More precisely, the dataset TPAMI includes twenty-four multi-page papers corresponding to 217 document images from which we consider *abstract, affiliation, author, biografy, caption, figure, formulae, index term, page number, paragraph, references, running head, section title, subsection title, table, title* as possible layout components. The dataset BG includes twenty-four single-page papers from which we consider *abstract, author, index term, page number, paragraph, references, running head, section title, title* as possible layout components. The dataset ICML includes thirty multi-page papers, corresponding to 240 document images from which we consider *abstract, affiliation, author, body, figure, index term, paragraph, section title, subsection title, table, title* as possible layout components.

The iterative transductive classification algorithm is evaluated considering the following experimental setups: 4-fold cross-validation in the case of TPAMI, 6-fold cross-validation in the case of BG and 5-fold cross-validation for ICML. Unlike the standard cross-validation, here one fold at a time is set aside to be used as the *training set* (and not as the *test set*). Small training set sizes allow us to validate the transductive approach, but may result in high error rates.

In the step of mining emerging patterns, three experimental schemes of the thresholds $minGR$, $minSup$ have been set: in the case of TPAMI $minGR = \{1, 2, 8, 64\}$ and $minSup = \{30\%, 40\%, 50\%\}$, in the case of BG $minGR = \{1, 2, 8, 64\}$ and $minSup = \{10\%, 20\%, 30\%\}$, while in the case of ICML $minGR = \{1, 2, 8, 64\}$ and $minSup = \{10\%, 20\%, 30\%\}$. In Table 2 the average number of emerging patterns mined with different parameter values is reported. As expected, by increasing $minSup$ and $minGR$ values, the total number of EPs (sum of the number of EPs in the folds) is reduced. In particular, the number of EPs is more drastically reduced when increasing $minSup$ than when increasing $minGR$. This means that there are several patterns which characterize a class (a specific layout component) and therefore present a high discriminative power with respect to components belonging to other classes.

Another consideration can be done on the number of EPs mined for each specific class (Table 3). We note that the layout components, for which the descriptions are more heterogeneous or which can be misclassified, are characterized by an higher number of EPs. Indeed, the components which present strong regularities (e.g., described with the same set of features) are those which can be more easily identified and which therefore generate a smaller set of EPs for the classification. Differently, the components which present low regularities can be erroneously labeled and therefore require an higher number of EPs to be discriminated from the others[2]. For instance, a figure can be more easily identified than an abstract layout component.

[2] The risk is that in these cases we can have overfitting problems.

Table 2. Total number of emerging patterns mined from TPAMI, BG and ICML

TPAMI	minSup (%)		
minGR	30	40	50
1	528032	344798	254805
2	523274	341534	252355
8	516958	336733	248658
64	513503	334292	246843

BG	minSup (%)		
minGR	10	20	30
1	128327	88684	58603
2	126840	87644	58091
8	122591	84208	55718
64	121363	82980	54490

ICML	minSup (%)		
minGR	10	20	30
1	386996	176407	114492
2	382639	173372	112476
8	376645	169406	109814
64	374736	167742	108595

Table 3. Minimum and maximum number of emerging patterns mined per class

TPAMI	**minSup (%)**		
minGR	30	40	50
1	min:11470(references) max:89008(index_term)	min:5450(figure) max:43422(abstract)	min:3319(figure) max:37475(abstract)
2	min:11394(references) max:88158(index_term)	min:5436(figure) max:42908(abstract)	min:3310(figure) max:37035(abstract)
8	min:11309(references) max:87124(index_term)	min:5364(figure) max:42085(abstract)	min:3271(figure) max:36304(abstract)
64	min:11276(references) max:86426(index_term)	min:5321(figure) max:41880(abstract)	min:3240(figure) max:36112(abstract)
BG	**minSup (%)**		
minGR	10	20	30
1	min:4380(references) max:45671(abstract)	min:4380(references) max:27342(author)	min:4380(references) max:15923(abstract)
2	min:4380(references) max:45179(abstract)	min:4380(references) max:26820(author)	min:4380(references) max:15825(abstract)
8	min:4218(references) max:43555(abstract)	min:4218(references) max:25713(author)	min:4218(references) max:15148(abstract)
64	min:4075(references) max:43171(abstract)	min:4075(references) max:25437(author)	min:4075(references) max:14764(abstract)
ICML	**minSup (%)**		
minGR	10	20	30
1	min:13923(body) max:169787(author)	min:5131(body) max:39728(abstract)	min:2780(body) max:27849(abstract)
2	min:13905(body) max:168468(author)	min:5120(body) max:38886(abstract)	min:2769(body) max:27213(abstract)
8	min:13843(body) max:166879(author)	min:5089(body) max:37828(abstract)	min:2756(body) max:26453(abstract)
64	min:13814(body) max:166671(author)	min:5065(body) max:37408(abstract)	min:2741(body) max:26152(abstract)

In Table 4, the macro average F1-score values are reported. Results are collected for different values of *minGR* and *minSup*. As we can see, better results are obtained when increasing *minGR* and/or when decreasing *minSup*. Indeed, higher values of *minGR* lead to exclude EPs with low discriminative capabilities and consider those with higher growth rate values with the result of (slightly) higher accuracy. While, when increasing *minSup* the number of EPs decreases and this leads to exclude models which, being infrequent, can characterize each class, with the result that the system has no enough information to discriminate among classes.

In Figure 2, precision, recall and $F1$ values are plotted by varying the value of k (which regulates the number of iterations). While, by increasing the number of iterations there is no improvement in terms of precision, results in terms of recall show benefits coming from the iterative transductive (bootstrapping) approach. This means that, with the iterative transduction, the system is able to associate

Table 4. Macro average $F1$-score values on TPAMI, BG and ICML with $k = 1$ and by varying $minGR$ and $minSup$

TPAMI	minSup (%)		
$minGR$	30	40	50
1	0.2906	0.2949	0.2555
2	0.3258	0.2694	0.2509
8	0.3264	0.2689	0.2511
64	0.3072	0.2684	0.2502

BG	minSup (%)		
$minGR$	10	20	30
1	0.6359	0.6323	0.6199
2	0.6548	0.6287	0.6091
8	0.6566	0.6341	0.6135
64	0.6411	0.6295	0.6142

ICML	minSup (%)		
$minGR$	10	20	30
1	0.3247	0.2791	0.2493
2	0.3118	0.2762	0.2686
8	0.3052	0.2988	0.1987
64	0.4028	0.2969	0.1976

to the correct class components that, otherwise, would remain unclassified. An exception is represented by TPAMI, where the system, due to the high number of components and to highly unbalanced data, is not able to reach good values of precision/recall. Obviously, a bad initial classification, negatively affects results of the iterative transductive approach.

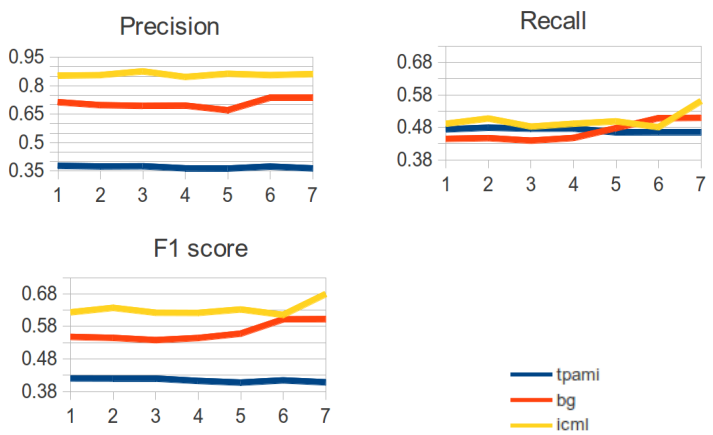

Fig. 2. Macro average precision, recall and $F1$-score on TPAMI, BG and ICML by varying the value of k. Results for TPAMI are obtained with $minGR = 8$ and $minSup = 30$ while results for BG and ICML are obtained with $minGR = 8$ and $minSup = 10$.

6 Conclusions

In this work, the induction of a classifier for the automated recognition of relevant layout components has been investigated. In particular, we have investigated the combination of transductive inference with principled relational classification in order to face the challenges posed by the application domain, characterized by complex and heterogeneous data, which are naturally modeled as several tables of a relational database, and characterized by the availability of a small (large) set of labeled (unlabeled) data. On the basis of an iterative bootstrapping approach, we exploit reliable classifications to classify other working examples in subsequent iterative steps. Interesting results on three real-world datasets are reported. They show that the iterative bootstrapping approach is able to increase recall of the obtained classifications.

References

1. Baird, H.S., Casey, M.R.: Towards Versatile Document Analysis Systems. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 280–290. Springer, Heidelberg (2006)
2. Ceci, M., Appice, A.: Spatial associative classification: Propositional vs structural approach. Journal of Intelligent Information Systems 27(3), 191–213 (2006)
3. Ceci, M., Appice, A., Malerba, D.: Discovering Emerging Patterns in Spatial Databases: A Multi-relational Approach. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 390–397. Springer, Heidelberg (2007)
4. Ceci, M., Appice, A., Malerba, D.: Transductive Learning for Spatial Data Classification. In: Koronacki, J., Raś, Z.W., Wierzchoń, S.T., Kacprzyk, J. (eds.) Advances in Machine Learning I. SCI, vol. 262, pp. 189–207. Springer, Heidelberg (2010)
5. Ceci, M., Berardi, M., Malerba, D.: Relational Data Mining and ILP for Document Image Understanding. Applied Artificial Intelligence 21(4-5), 317–342 (2007)
6. Ceci, M., Loglisci, C., Malerba, D.: Transductive Learning of Logical Structures from Document Images. In: Biba, M., Xhafa, F. (eds.) Learning Structure and Schemas from Documents. SCI, vol. 375, pp. 121–142. Springer, Heidelberg (2011)
7. Ceci, M., Malerba, D.: Classifying web documents in a hierarchy of categories: a comprehensive study. Journal of Intelligent Information Systems 28(1), 37–78 (2007)
8. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: International Conference on Knowledge Discovery and Data Mining, pp. 43–52. ACM Press (1999)
9. Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: Classification by Aggregating Emerging Patterns. In: Arikawa, S., Nakata, I. (eds.) DS 1999. LNCS (LNAI), vol. 1721, pp. 30–42. Springer, Heidelberg (1999)
10. Esposito, F., Malerba, D., Semeraro, G.: Multistrategy learning for document recognition. Applied Artificial Intelligence 8(1), 33–84 (1994)
11. Krogel, M.-A., Scheffer, T.: Multi-relational learning, text mining, and semi-supervised learning for functional genomics. Mach. Lear. 57(1-2), 61–81 (2004)
12. Lisi, F.A., Malerba, D.: Inducing multi-level association rules from multiple relations. Machine Learning 55(2), 175–210 (2004)
13. Malerba, D., Ceci, M., Appice, A.: A relational approach to probabilistic classification in a transductive setting. Engineering Applications of Artificial Intelligence 22(1), 109–116 (2009)
14. Niyogi, D., Srihari, S.N.: Knowledge-based derivation of document logical structure. In: ICDAR 1995: Proceedings of the Third International Conference on Document Analysis and Recognition, vol. 1, p. 472. IEEE Computer Society, Washington, DC (1995)
15. Seeger, M.: Learning with labeled and unlabeled data. Technical report, Institute for Adaptive and Neural Computation. University of Edinburgh (2001)
16. Zhang, X., Dong, G., Ramamohanarao, K.: Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. In: Knowledge Discovery and Data Mining, pp. 310–314 (2000)