Hierarchical Multidimensional Classification of web documents with MultiWebClass

Francesco Serafino, Gianvito Pio, Michelangelo Ceci, and Donato Malerba

Department of Computer Science, University of Bari Aldo Moro
Via Orabona, 4 - 70125 Bari - Italy

{firstname.surname}@uniba.it

Abstract. Most of works on text categorization have focused on classifying documents into a set of categories with no relationships among them (flat classification). However, due to the intrinsic structure that can be found in many domains, recent works are focusing on more complex tasks, such as multi-label classification, hierarchical classification and multidimensional classification. In this paper, we propose the hierarchical multidimensional classification task, where documents can be classified according to different dimensions/viewpoints (e.g., topic, geographic area, time period, etc.), where in each dimension categories can be organized hierarchically. In particular, we propose the system MultiWebClass, a multidimensional variant of the system WebClassIII, which discovers correlations among categories belonging to different dimensions and exploits them, according to two different strategies, to refine the set of features used during the learning process. Experimental evaluation performed on both synthetic and real datasets confirms that the exploitation of correlations among categories can lead to better results in terms of classification accuracy, possibly reducing specialization error or generalization error, depending on the strategy adopted for the refinement of the feature sets.

Keywords: Structured Output Prediction, Text Categorization, Hierarchical Classification, Multidimensional Classification

1 Introduction

The number of web documents continuously and massively increases every day and their automatic classification is considered an essential task. In recent years, a plethora of classification algorithms has been developed. Some of them work in the single label classification setting, where categories are not organized according to any specific schema. However, (web) documents can naturally be classified into several hierarchically organized categories. For example, a blog article classified as Sport could, at the same time, be classified as Tennis, Roland Garros, and so on. For this reason, recent works have focused on the hierarchical classification task, where class labels are hierarchically organized and each object is associated to more than one class label (according to the hierarchy).

Moreover, documents can be classified according to different classification dimensions. For example, a web page could be classified according to its topic, the referenced geographical information or the publication date. By considering multiple dimensions of classification, each of which possibly hierarchically organized, it is possible to define the task of Hierarchical Multidimensional Classification. More formally, this task represents the combination of Multidimensional Classification and Hierarchical Classification, where: *i)* more than one class attribute is associated to each document, each describing the document according to a different point of view and *ii)* each class attribute is hierarchically organized.

In the literature, several works about multidimensional classification and hierarchical classification have been proposed. Some works consider the first task as a variant of (and, accordingly, convertible to) the latter, whereas other works consider these two tasks separately. For example, in [2] a multidimensional classification method based on Bayesian Network is proposed. The authors perform multidimensional classification on a flat set of labels by organizing class variables (dimensions), feature variables and bridges (from classes to features) as three distinct network subgraphs. In [14], the authors propose the application of multidimensional classification approaches to biomedical texts, in order to extract specific portions of text containing scientific content.

In [16] the authors propose a framework which implements three different classifiers (kNN, naïve Bayes and centroid-based) in order to evaluate three different techniques for multidimensional classification: flat-based, hierarchical-based and multidimensional-based. In the first two cases, they convert the multidimensional model into flat and hierarchical models, respectively, whereas in the last case they consider each dimension separately. Experiments performed on two datasets showed that the multidimensional-based and hierarchical-based approaches outperform the flat-based approach.

Moreover, in a recent work [7], the hierarchical multidimensional classification task is solved by considering it as a multi-label classification task. First, the system builds a set of probabilistic multi-class classifiers (one for each non-leaf node in the hierarchy) which are applied simultaneously to each test instance. Second, a probability is computed for each path in the hierarchy, by combining the output of the classifiers learned for the nodes involved in the path. Finally, the path with the highest probability is the output of the classification.

In this paper, we extend the system WebClassIII [5] (described in Section 2), which offers a hierarchical classification framework, with the more complex task of Hierarchical Multidimensional Classification. Moreover, we exploit the possible multi-dimensionality of the data in order to improve the classification with respect to each single dimension, which are generally classified independently by existing works. In order to exploit such possible dependencies, we propose the identification of correlations among categories belonging to different hierarchies. Such correlations can be exploited to improve the classification accuracy with respect to a given hierarchy, even when the other hierarchies are not the main subject of the classification task. The discovery of correlations is motivated by the reasonable assumption that documents labeled with a given category along

one dimension could be usually labeled with another given category belonging to another dimension. For example, we can consider documents organized according to the *Geographic* dimension and the *Topic* dimension. If many documents labeled with *Rome* in the *Geographic* hierarchy (structured as *Europe* → *Italy* → *Rome*), are also frequently labeled as *Traffic* in the *Topic* hierarchy (structured as *News* → *Accident* → *Traffic*), then it is possible that there is a correlation between the categories *Rome* and *Traffic*, possibly representing the fact that: "Rome is affected by an high number of accidents due to traffic".

The rest of the paper is organized as follows. In the next section, the classification framework implemented in WebClassIII, which represents the background of this work, is described. In Section 3 the extension of WebClass to perform hierarchical multidimensional classification is presented. Some experimental results on both synthetic and real datasets are reported and discussed in Section 4. Final conclusions and remarks are reported in Section 5.

2 Background: WebClassIII

WebClass is a classification framework for HTML pages. The last version of WebClass, i.e., WebClassIII, is an extension for the hierarchical text categorization [5] which exploits three different classification approaches: Naïve Bayes [10], centroid-based [6] and SVM [11].

The hierarchical organization of categories is exploited in all the phases of the document classification, namely feature selection, learning of the classification model, and categorization of a new document. Documents are represented as bag-of-words (where each term is associated to its frequency in the document). In general, two alternatives can be considered [1]: *i)* the same feature space is used to represent documents belonging to all categories or *ii)* several specific feature spaces are used to represent documents belonging to different categories. In WebClassIII an intermediate solution is adopted. In particular, for each category, a different document representation is used to decide which subcategory (temporary represented in the same feature space of its parent) is the most appropriate for a given document.

In the learning phase, starting from the root, the system builds a classification model for each category c. When c has only a subcategory, a dummy subcategory is introduced. The training documents associated to the dummy subcategory are those associated only to c (and not to the subcategory). Therefore, the sum of probabilities of all the direct subcategories of c is not necessarily 1.0, since the probability that the document does not belong to any subcategory should be taken into account.

The classification phase is performed in a top-down fashion from the root to the leaves, according to a greedy strategy. When the document reaches an internal category c, it is represented on the basis of the feature set associated to c and the system computes a score for each direct subcategory (dummy categories are not considered during the classification), according to the classification model learned on c. The document is associated to the subcategory with the highest

score above a precomputed threshold (one for each category, see [4] for details). The search proceeds recursively from that subcategory, until no score is greater than the corresponding threshold or a leaf category is reached. The first case mainly happens when the document deals with a general rather than a specific topic or when the document belongs to a specific category which does not appear in the hierarchy. If the search stops at the root, then the document is marked as unclassified.

In the following, we report some details about the identification of an appropriate subset of terms (dictionary) for representing documents belonging to each category, which are then extended in order to exploit correlations among categories.

In particular, documents are initially tokenized, and the set of tokens is filtered in order to remove HTML tags, punctuation marks, numbers and tokens with less than three characters. After tokenization, two standard pre-processing methods are applied, that are stopword removal (based on words in Glimpse [9]) and stemming (using Porter algorithm [12]).

WebClassIII associates each category with a subset of words which best represent documents of that category. In particular, each word $w_{i,c'}$ that appears in at least a document of the category c' (direct subcategory of c), WebClassIII computes a weight $v_{i,c'}$ and builds a dictionary $Dict_{c'}$ containing n_{dict} words with the highest weight. The weight $v_{i,c'}$ is computed as follows:

$$v_{i,c'} = TF_{c'}(w_i) \times DF^2_{c'}(w_i) \times \frac{1}{CF_c(w_i)} \tag{1}$$

where:

- $TF_{c'}(w)$ is the *maximum term frequency* of w over the documents belonging to the category c';
- $DF_{c'}(w)$ is the *document frequency* computed as the percentage of documents of category c' in which w occurs;
- $CF_c(w)$ is the *category frequency* computed as the number of direct subcategories of c having at least a document in which w occurs.

It is noteworthy that positive examples for c' are sufficient for the computation of $TF_{c'}(w)$ and $DF_{c'}(w)$, while the computation of $CF_c(w)$ also requires the negative examples for c'.

The feature set associated to each category c, which is exploited for learning the corresponding classifier, consists of the union of the dictionaries associated to all the subcategories of c (called Hierarchical Feature Set in [5]).

3 Hierarchical Multidimensional Classification

In this Section, we describe MultiWebClass, an extension of WebClassIII which is able to perform Hierarchical Multidimensional Classification. The most straightforward solution would consist in learning a classification model for each dimension independently. However, as discussed in Section 1, this solution is not able to

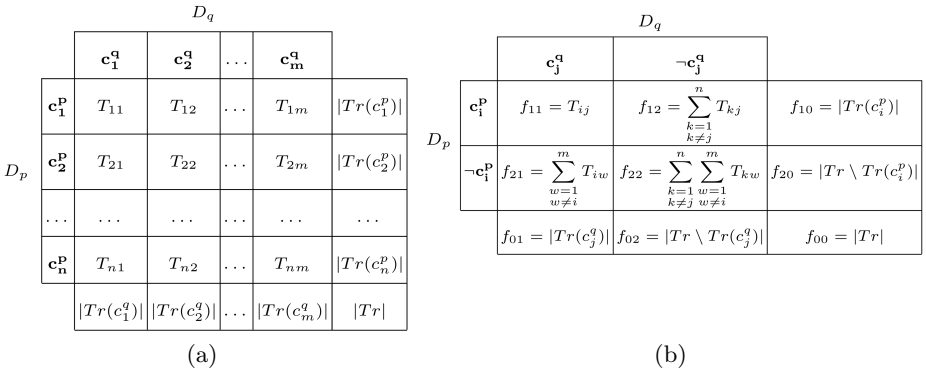

Fig. 1. (a) Contingency matrix between two dimensions D_p and D_q. (b) Contingency matrix between two categories $c_i^p \in D_p$ and $c_j^q \in D_q$ built using artificial categories.

catch possible correlations among categories belonging to different dimensions. In MultiWebClass we adopt a different solution which first identifies possible correlations among categories belonging to different dimensions and then exploit such correlations in order to *extend* the feature sets used in the learning phase so to improve predictive performances. In the following subsections, we describe the proposed approach.

3.1 Discovery of correlations between categories

The identification of correlations between two variables is a common task in statistics which is usually solved by means of a contingency matrix, where rows and columns represent the values of the first and the second variables, respectively. Inspired by this commonly used solution, we use the same strategy for categories. In particular, we build a contingency matrix as shown in Figure 1(a), where:

- the variables on rows and columns represent two classification dimensions $D_p = \{c_1^p, c_2^p, \ldots, c_n^p\}$ and $D_q = \{c_1^q, c_2^q, \ldots, c_m^q\}$, where c_i^p is the i-th category of the p-th dimension;
- each cell value T_{ij} represents the number of documents labeled as both the categories c_i^p and c_j^q in the training set;
- $Tr(c_i^p)$ represents the set of training documents labeled as c_i^p in the hierarchy of the p-th dimension.

Starting from such contingency matrix, we construct a further 2×2 contingency matrix for each pair of categories c_i^p, c_j^q belonging to different dimensions ($p \neq q$) as shown in Figure 1(b). In this matrix, we build two artificial categories $\neg c_i^p, \neg c_j^q$ which consist of all the documents not belonging to c_i^p and c_j^q, respectively. Obviously, we exclude root categories when computing such contingency matrices.

The correlation between two categories can be symmetric or not, on the basis of the considered correlation measure. In this work, we exploit a variant of the *Confidence* measure [15]. Such measure is asymmetric and is defined as $\gamma(c_i^p, c_j^q) = \frac{f_{11}}{f_{10}}$ and $\gamma(c_j^q, c_i^p) = \frac{f_{11}}{f_{01}}$, where f_{11}, f_{10} and f_{01} are the values of the contingency matrix between c_i^p and c_j^q (see Figure 1(b)). This measure is actually a frequency-based estimation of the probability that documents belonging to c_i^p also belong to c_j^q (or vice versa). This measure, however, can lead to unreliable probabilities in the case that very few documents are used for its computation. To overcome this issue, we consider $\frac{f_{11}}{f_{10}}$ (and $\frac{f_{11}}{f_{01}}$) as a proportion in a statistical population and we use the Wilson confidence interval [17] in order to make conservative decisions about the presence of a correlation. We use the Wilson score interval since it directly derives from the Pearson's chi-squared test with two cases (here the two cases for $\frac{f_{11}}{f_{10}}$ are: a document that belongs to c_i^p also belongs to c_j^q or not). Formally, the Wilson score interval for $\frac{f_{11}}{f_{10}}$ is defined as:

$$\left[\frac{f_{11}+\frac{z_0^2}{2}}{f_{10}+z_0^2} - \frac{\sqrt{f_{10}}z_0}{f_{10}+z_0^2}\sqrt{\frac{f_{11}f_{12}}{f_{10}^2} + \frac{z_0^2}{4f_{10}}} \quad , \quad \frac{f_{11}+\frac{z_0^2}{2}}{f_{10}+z_0^2} + \frac{\sqrt{f_{10}}z_0}{f_{10}+z_0^2}\sqrt{\frac{f_{11}f_{12}}{f_{10}^2} + \frac{z_0^2}{4f_{10}}} \right] \tag{2}$$

where z_0 is the Z-score value (according to the normal distribution) for a given confidence $1 - \alpha$.

Since we are interested in making conservative decisions about the presence of a correlation, we consider the lower bound of this interval as the probability that c_i^p and c_j^q are correlated:

$$\gamma(c_i^p, c_j^q) = \frac{f_{11}+\frac{z_0^2}{2}}{f_{10}+z_0^2} - \frac{\sqrt{f_{10}}z_0}{f_{10}+z_0^2}\sqrt{\frac{f_{11}f_{12}}{f_{10}^2} + \frac{z_0^2}{4f_{10}}} \tag{3}$$

Due to the asymmetry of the considered correlation measure, as shown in Algorithm 1, we pair-wisely search for correlations between two categories belonging to two different dimensions D_p and D_q in both the directions $D_p \rightarrow D_q$ and $D_q \rightarrow D_p$. Note that we are only interested in the discovery of positive correlations (i.e., a given category on a dimension possibly implies a category in another dimension). For this reason, we do not consider the proportions f_{21}/f_{01} and f_{12}/f_{10}. Moreover, since we are only interested to highly correlated pairs of categories, we consider as correlated two categories c_i^p and c_j^q only if $\gamma(c_i^p, c_j^q) > \beta$, where β is a user-defined threshold.

Finally, since we use the correlations to extend the feature sets and we use hierarchical feature sets (this aspect will be clarified in the next subsection), if the correlation $c_i^p \rightarrow c_j^q$ is identified, then it is possible to prove that, for each $c_k^q \in ancestors(c_j^q)$, there exists the correlation $c_i^p \rightarrow c_k^q$. This can be easily proved by observing Equation (3). Indeed, for each $c_k^q \in ancestors(c_j^q)$ we have that $\gamma(c_i^p, c_k^q) \geq \gamma(c_i^p, c_j^q)$. This directly follows from the following observations:

- f_{01} has the same value in both $\gamma(c_i^p, c_k^q)$ and $\gamma(c_i^p, c_j^q)$, since it is the number of documents labeled as c_i^p;

Algorithm 1: `Discovery of correlations among categories`

input : The set of dimensions $\mathcal{D} = \{D_1, D_2, \ldots, D_s\}$;
a correlation measure $\gamma(\cdot, \cdot)$;
a threshold β to consider a discovered correlation as relevant.
output: $correlations = \{\langle c, CorrelatedSet_c \rangle\}_c$, where $CorrelatedSet_c$ is the set of categories d, s.t. exists the correlation $c \rightarrow d$.

```
1  correlations ← ∅;
2  for all pairs of dimensions D_p, D_q do
3  |  correlations ← correlations ∪ findCorrelations(D_p, D_q, β);
4  |  correlations ← correlations ∪ findCorrelations(D_q, D_p, β);
5  return correlations;

6  findCorrelations(D_p, D_q, β)
7  correlations ← ∅;
8  exploredPairs ← ∅;
9  for c_i^p ∈ D_p in pre-order do
10 |  correlatedSet ← ∅;
11 |  for c_j^q ∈ D_q in post-order do
12 |  |  if ⟨c_i^p, c_j^q⟩ ∉ exploredPairs and γ(c_i^p, c_j^q) ≥ β then
13 |  |  |  correlatedSet ← correlatedSet ∪ {c_j^q};
   |  |  |  // Skip the exploration of ancestors of c_j^q
14 |  |  |  for c_k^q ∈ ancestors(c_j^q) and c_k^q ≠ root(D_q) do
15 |  |  |  |  exploredPairs ← exploredPairs ∪ {⟨c_i^p, c_k^q⟩};
16 |  |  |  |  correlatedSet ← correlatedSet ∪ {c_k^q};
17 |  correlations ← correlations ∪ {⟨c_i^p, correlatedSet⟩};
18 return correlations;
```

- f_{12} has the same value in both $\gamma(c_i^p, c_k^q)$ and $\gamma(c_i^p, c_j^q)$, since it is the number of documents which are not labeled as c_i^p;
- f_{11} for $\gamma(c_i^p, c_k^q)$ is greater than or equal to f_{11} for $\gamma(c_i^p, c_j^q)$, since c_k^q contains documents in c_j^q.

In order to take into account this property, we visit the first hierarchy in pre-order and the second hierarchy in post-order. The effect is a reduction of the number of correlations between pairs of categories to be evaluated (see Algorithm 1, lines 14-16).

3.2 Exploiting discovered correlations

As shown in Section 2, in WebClassIII, the feature set associated to each category is the union of the dictionaries of its subcategories. In this work, we exploit the discovered correlations to extend the feature set of some categories. In particular, given a discovered correlation $c_i^p \rightarrow c_j'^q$ (where $c_j'^q$ is a subcategory of c_j^q), we

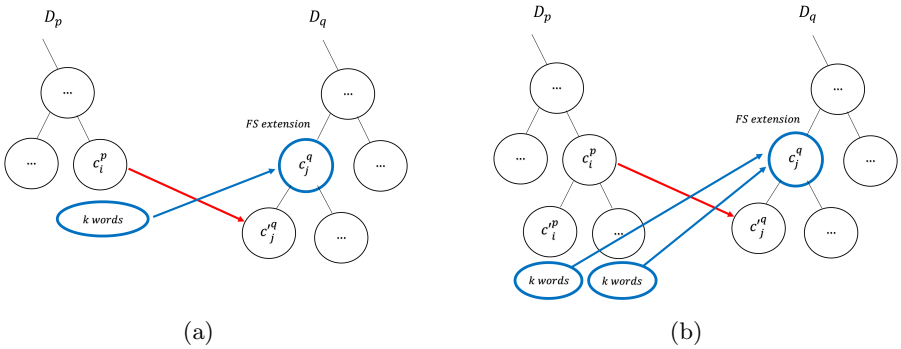

Fig. 2. Feature set extension FSE I, when (a) c_i^p is a leaf and when (b) c_i^p is not a leaf. Red arrows indicate correlations, while blue arrows indicate feature set extension.

extend the feature sets of the involved categories according to two different strategies:

- **FSE I: Category Dictionaries.** This strategy extends the feature set of the category c_j^q by including features in the dictionaries of categories in the dimension D_p:
  - When c_i^p is a leaf, then the feature set of the category c_j^q is extended by including the top-k ($k < n_{dict}$) terms of the dictionary associated c_i^p (see Figure 2(a));
  - When c_i^p is not a leaf, the feature set of the category c_j^q is extended by including the top-k ($k < n_{dict}$) terms from the dictionaries of the subcategories of c_i^p (see Figure 2(b)).

  The top-k terms are selected according to Equation 1. The rationale behind this strategy is to use, in the classifier associated to c_j^q, some of the features that the classifier associated to c_i^p uses to discriminate among its child categories.
- **FSE II: Correlated Words.** This strategy works similarly to FSE I, but adds a different set of terms to the feature set of c_j^q. In particular, the top-k ($k < n_{dict}$) words, according to Equation 1, are those that appear in the feature sets of both c_i^p and $c_j'^q$, but not to c_j^q. Note that the feature selection algorithm implemented in WebClassIII could have pruned some features of $c_j'^q$ when building the feature set of c_j^q. This strategy could restore such features because of the presence of the correlation.

FSE II is more conservative than FSE I since it uses, for feature extension, features extracted from the same dimension. On the contrary, since FSE I incorporates in one dimension features coming from a completely different dimension, it is more "daring", but can also incur into errors (possibly) coming from unrelated features.

4 Experiments

The experimental evaluation has been performed using, in WebClass, the Naïve Bayes classification algorithm which was proved to perform the best among those implemented in the framework [5]. For evaluation purposes, we consider both real and synthetically generated datasets. Since we are interested in the evaluation of the contribution given by the exploitation of the correlations, in this work the results obtained with both the proposed feature set extension strategies introduced in MultiWebClass (i.e., FSE I and FSE II) are compared only with the results obtained by WebClass III. This because WebClass III does not exploit correlations among different hierarchies and represents the non-multidimensional counterpart of the approach we propose.

4.1 Datasets

Real dataset. The evaluation on real data has been performed on the dataset Reuters Corpora Volume 1 (RCV1) [8]. It contains more than 800,000 stories collected by the international news agency Reuters from 20th August 1996 to 19th August 1997. News are manually labeled according to the following three classification dimensions:

- *Topic*: the main subject of a story (hierarchically organized);
- *Industry*: the type of business discussed in a story (hierarchically organized);
- *Region*: a geographic location or an economic/political group (not hierarchically organized).

The training and testing sets are obtained according to [8], i.e., by selecting documents published from 20th August 1996 to 31st August 1996 as training set, and documents published from 1st September 1996 to 19th August 1997 as testing set. Coherently with [5], we considered only documents associated to a single category. The result is a set of 4,517 training documents and 146,248 testing documents.

Synthetic datasets. The evaluation on synthetic data has been performed by means of the 5 fold cross validation approach on a set of datasets generated by simulating the presence of correlations among categories belonging to different dimensions. In particular, the generation of the datasets takes into account the following aspects:

- *Dimensions.* The set of dimensions $\mathcal{D}$ is created using as parameters: the number of dimensions s, the tree *depth* and the *degree* of nodes. Each dimension is represented as a full and perfectly-balanced tree.
- *Dictionaries.* Each leaf category is associated with a fixed number of terms (50 in our experiments) randomly selected from the Ispell American English[1] dictionary, whereas each internal category is associated with the dictionary obtained as the union of the terms selected for its subcategories.

[1] http://fmg-www.cs.ucla.edu/geoff/ispell-dictionaries.html

- *Document Generation.* For each category c, a set of d_c documents is generated. Each document consists of a set of t terms, randomly selected from the dictionary associated to c, and a set of t_g general terms randomly taken from the Ispell American English dictionary. The complete set of documents belonging to the category c is then obtained as the union of the d_c documents (properly) belonging to c and the set of documents belonging to its subcategories.
- *Correlation Injection.* Correlations are injected between pairs of categories belonging to different dimensions. Injection is performed by labeling documents belonging to a category of a given dimension also with a given category of another dimension. The system takes an input parameter n_{corr} which represents the number of correlations to inject into leaf categories, whereas correlations between internal categories are injected in a bottom-up fashion. In particular, for each correlation $c_i^p \rightarrow c_j^q$ injected at l-th level, we introduce a correlation $father(c_i^p) \rightarrow father(c_j^q)$ at $(l-1)$-th level. In order to avoid the injection of several and redundant correlations involving the same internal categories, only the most frequent correlations are preserved. The percentage of correlations to preserve is given by the user-defined parameter $CorrRatio$.

We generated all the synthetic datasets with the following parameters: $s = 3$ (i.e., 3 hierarchically organized classification dimensions), $degree = 2$ (i.e., each internal node has 2 children), $n_{corr} = 20$ (i.e., for each pair of dimensions, we inject a correlation for 20 randomly selected pairs of leaf categories) and $CorrRatio = 0.40$ (i.e., we preserve the top-40% most frequent correlations among parent categories).

For each category, all the documents are generated by randomly selecting 80 terms from the dictionary of the category (i.e., $t = 80$) and 20 terms from the Ispell American English dictionary (i.e., $t_g = 20$). The number of hierarchical levels (i.e., the parameter $depth$) and the number of documents d_c for each category are set to different values in order to generate different synthetic datasets. In particular, we generated five different datasets, that are:

- **4-20**, which has 4 hierarchical levels and 20 documents for each category;
- **4-30**, which has 4 hierarchical levels and 30 documents for each category;
- **4-40**, which has 4 hierarchical levels and 40 documents for each category;
- **4-50**, which has 4 hierarchical levels and 50 documents for each category;
- **3-40**, which has 3 hierarchical levels and 40 documents for each category.

While the results obtained on the datasets 4-20, 4-30, 4-40 and 4-50 are compared in order to analyze the performance by varying the number of documents for each category, results obtained on the datasets 3-40 and 4-40 are compared in order to obtain a preliminary analysis on the sensitiveness of the algorithm to the complexity of the hierarchical structure of each classification dimension.

4.2 Experimental Setting

In the following, we report some details about the parameter setting of MultiWebClass. In particular, we set the confidence value $1-\alpha$ to 0.95. Consequently, the Z-score value is $z_0 = 1.96$. The threshold value β has been set to 0.3. n_{dict} (size of dictionaries) is set to 25, which provided good results in WebClassIII. The number of terms k, propagated by the proposed strategies for feature set extension is set to 20 (coherently with n_{dict}).

As regards the classification dimensions, we considered the *Topic* dimension for RCV1 (which is typically used for classification purposes [13]), exploiting the correlations with the other two dimensions (i.e., *Industry* and *Region*) and the first dimension for the synthetic datasets, exploiting the correlations with the other two dimensions[2].

The comparison between WebClassIII and MultiWebClass has been performed according to five evaluation measures, that are:

- *Accuracy*, which is the percentage of correctly classified documents;
- *Generalization Error*, which is the percentage of documents classified as a super-category of the correct category;
- *Specialization Error*, which is the percentage of documents classified into a subcategory of the correct category;
- *Misclassification Error*, which is the percentage of documents classified as a category which is in a different path with respect to the correct category in the hierarchy;
- *Unknown Ratio*, which is the percentage of documents that are not classified (actually classified in the root category of the hierarchy).

Intuitively, the sum of all the considered measures is always equal to 1.

4.3 Results

According to the experimental setting, in this section we report the results obtained with all the considered datasets and perform three different analyses on:

- synthetic datasets with a fixed depth of the hierarchies (i.e., $depth = 4$) and a different number of documents for each category (i.e., $d_c = \{20, 30, 40, 50\}$);
- synthetic datasets with a fixed number of documents for each category (i.e., $d_c = 40$) and different depths of the hierarchies (i.e., $depth = \{30, 40\}$);
- real data, i.e., on the RCV1 dataset.

Synthetic datasets with fixed depth. Results for this analysis are reported in Table 1. As it can be observed from the table, all the considered approaches were able to make at least a decision in the root node, i.e., the Unknown Ratio is 0. Moreover, by observing the Misclassification Error, we can see that all the systems were almost always able to consider the correct path in the hierarchy.

[2] In the case of synthetic datasets, results do not depend on the specific dimension.

Dataset 4-20					
System	**Accuracy**	**Gen. Error**	**Spec. Error**	**Misclass. Error**	**Unknown Ratio**
MWC - FSE I	0.759	**0.168**	0.073	**0.000**	**0.000**
MWC - FSE II	**0.764**	0.217	**0.019**	**0.000**	**0.000**
WebClass III	0.685	0.246	0.069	**0.000**	**0.000**
Dataset 4-30					
System	**Accuracy**	**Gen. Error**	**Spec. Error**	**Misclass. Error**	**Unknown Ratio**
MWC - FSE I	**0.795**	**0.153**	**0.052**	**0.000**	**0.000**
MWC - FSE II	0.751	0.162	0.087	**0.000**	**0.000**
WebClass III	0.698	0.245	0.057	**0.000**	**0.000**
Dataset 4-40					
System	**Accuracy**	**Gen. Error**	**Spec. Error**	**Misclass. Error**	**Unknown Ratio**
MWC - FSE I	**0.731**	0.261	**0.008**	**0.000**	**0.000**
MWC - FSE II	0.711	**0.229**	0.030	0.030	**0.000**
WebClass III	0.676	0.245	0.079	**0.000**	**0.000**
Dataset 4-50					
System	**Accuracy**	**Gen. Error**	**Spec. Error**	**Misclass. Error**	**Unknown Ratio**
MWC - FSE I	**0.829**	**0.099**	0.072	**0.000**	**0.000**
MWC - FSE II	0.715	0.147	0.138	**0.000**	**0.000**
WebClass III	0.727	0.247	**0.026**	**0.000**	**0.000**

Table 1. Classification results of synthetic dataset 4-20, 4-30, 4-40 and 4-50 on the first dimension.

The main differences can be observed in the other three measures. In particular, FSE I always leads to better accuracy values when the number of documents per category increases (i.e., $d_c > 20$). This approach also obtains good results in terms of Generalization Error, sometimes at the cost of a slightly higher Specialization Error, which confirms that FSE I generally leads to less conservative decisions, if compared to FSE II. Overall, by comparing the results obtained by FSE I and FSE II with those obtained by WebClassIII, it is possible to see that the exploitation of the discovered correlations leads to better results.

Synthetic datasets with different depth. This analysis aims at evaluating the performance with respect to the depth of the hierarchy. Results are reported in Table 2. As expected, the higher the complexity of the classification hierarchy, the lower the classification accuracy. However, the proposed approaches always

lead to better results. A more detailed analysis reveals that, as expected, moving from 3 to 4 levels in the hierarchy leads to reduce the advantage of the MultiWebClass with respect to WebClassIII (percentage gain in accuracy changes from 9.3% to 8.1% in the case of FSE I). As regards the specific error measures, we can observe that all the systems generally prefer to make Generalization Errors with respect to Specialization Errors when the complexity of the hierarchy increases.

Dataset 3-40					
System	**Accuracy**	**Gen. Error**	**Spec. Error**	**Misclass. Error**	**Unknown Ratio**
MWC - FSE I	**0.834**	**0.129**	0.037	**0.000**	**0.000**
MWC - FSE II	0.833	0.131	**0.036**	**0.000**	**0.000**
WebClass III	0.763	0.169	0.068	**0.000**	**0.000**
Dataset 4-40					
System	**Accuracy**	**Gen. Error**	**Spec. Error**	**Misclass. Error**	**Unknown Ratio**
MWC - FSE I	**0.731**	0.261	**0.008**	**0.000**	**0.000**
MWC - FSE II	0.711	**0.229**	0.030	0.030	**0.000**
WebClass III	0.676	0.245	0.079	**0.000**	**0.000**

Table 2. Classification results of synthetic dataset 3-40 and 4-40 on the first dimension.

Real Data. Finally, we compare the results obtained with MultiWebClass and WebClassIII on RCV1. Results are reported in Table 3.

From the results we can see that, differently from what we observed for synthetic datasets, Unknown Ratio is nonzero. This because, contrary to the synthetic datasets, testing set contains documents that do not come from the same data distribution of training documents. However, despite higher Unknown Ratio, FSE II obtains better results in terms of Accuracy, Generalization Error, Specialization Error and Misclassification Error. This confirms the more conservative nature of FSE II, which makes more accurate predictions and avoids wrong decisions when the degree of uncertainty is high. This general behavior can suggest us the use of FSE II when we would like to obtain a more accurate classification, at the price of some unclassified instances, whereas FSE I (and, in some cases, the original WebClass III) is more appropriate when we want to force classification (reducing Unknown Ratio), at the price of a higher Specialization Error. This observation does not hold for synthetic datasets since the unknown ratio is always zero due to the considerations about data distribution reported before.

In Table 4, we show some correlations discovered by our approach and used for feature set extension. It is noteworthy that most of them appear reasonable. For example, some regions which are usually subject to political issues are correlated to the topic *Government/Social*. Moreover, some regions whose economy is based on some specific business activities are correlated to the topic *Corporate/Industrial*.

Reuters RCV1					
System	**Accuracy**	**Gen. Error**	**Spec. Error**	**Misclass. Error**	**Unknown Ratio**
MWC - FSE I	0.559	0.141	0.003	0.087	0.210
MWC - FSE II	**0.566**	**0.129**	**0.002**	**0.078**	0.225
WebClass III	0.561	0.188	0.003	0.101	**0.147**

Table 3. Classification results of RCV1 on the dimension *Topic*.

Source category	**Target category**	**Correlation Strength**
Region.Cyprus	Topic.Government/Social	0.51841
Region.EuropeanUnion	Topic.EuropeanCommunity	0.51322
Region.Macedonia	Topic.Government/Social	0.46769
Industry.PortsAndShippingServices	Topic.Corporate/Industrial	0.42569
Region.Ghana	Topic.EquityMarkets	0.42438
Region.Malawi	Topic.Government/Social	0.35930
Region.Syria	Topic.Government/Social	0.35479
Region.Bahrain	Topic.Government/Social	0.35027
Region.Jamaica	Topic.Corporate/Industrial	0.34238
Region.Malta	Topic.Government/Social	0.32404
Industry.PortsAndShippingServices	Topic.Capacity/Facilities	0.31651
Region.CzechRepublic	Topic.Markets	0.30804
Region.UnitedArabEmirates	Topic.Government/Social	0.30070

Table 4. Correlations discovered on RCV1 with correlation strength greater than 0.3.

5 Conclusions and Future Work

In this paper we tackled the Hierarchical Multidimensional Classification task and presented, at this purpose, the system MultiWebClass. In particular, MultiWebClass discovers correlations between categories belonging to different hierarchies and exploits them by extending (according to two different strategies) the feature sets used for learning classifiers.

Results on both synthetic and real datasets show that the exploitation of the discovered correlations, which appear reasonable after a quick qualitative analysis, can lead to better classification performances in terms of accuracy. Moreover, the different strategies proposed for feature set extension appear appropriate for different goals (i.e., higher accuracy *vs* higher number of classified instances), since they have a different degree of conservativeness when making decisions.

For future work, we intend to deeply analyze the sensitiveness of MultiWebClass to different parameter settings. We will also consider additional strategies for exploiting the discovered correlations, possibly including negative correlations. Moreover, inspired by the work in [3], we will explore the task of multidimensional hierarchical classification in the transductive setting. Finally, we intend to perform experiments on additional real-world datasets, also related to different application domains (e.g., biological data).

Acknowledgements

We would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

References

1. C. Apté, F. Damerau, and S. M. Weiss. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*, 12(3):233–251, 1994.
2. C. Bielza, G. Li, and P. Larranaga. Multi-dimensional classification with bayesian networks. *International Journal of Approximate Reasoning*, 52(6):705–727, 2011.
3. M. Ceci. Hierarchical text categorization in a transductive setting. In *Workshops Proc. of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pages 184–191. IEEE Computer Society, 2008.
4. M. Ceci and D. Malerba. Hierarchical Classification of HTML Documents with WebClassII. In F. Sebastiani, editor, *ECIR 2003, Pisa, Italy, April 14-16, 2003, Proceedings*, volume 2633 of *LNCS*, pages 57–72. Springer, 2003.
5. M. Ceci and D. Malerba. Classifying web documents in a hierarchy of categories: a comprehensive study. *Journal of Intelligent Information Systems*, 28(1):37–78, 2007.
6. E.-H. S. Han and G. Karypis. *Centroid-based document classification: Analysis and experimental results*. Springer, 2000.
7. J. Hernández, L. E. Sucar, and E. F. Morales. Multidimensional hierarchical classification. *Expert Systems with Applications*, 41(17):7671–7677, 2014.
8. D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.
9. U. Manber, S. Wu, et al. Glimpse: A tool to search through entire file systems. In *Usenix Winter*, pages 23–32, 1994.
10. T. M. Mitchell. *Machine learning*. McGraw Hill series in computer science. McGraw-Hill, 1997.
11. J. Platt et al. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methodssupport vector learning*, 3, 1999.
12. M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
13. R. E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.
14. H. Shatkay, F. Pan, A. Rzhetsky, and W. J. Wilbur. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093, 2008.
15. P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41. ACM, 2002.
16. T. Theeramunkong and V. Lertnattee. Multi-dimensional text classification. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.
17. E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.