Discovering and Exploiting Deterministic Label Relationships in Multi-Label Learning

Christina Papagiannopoulou
School of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, Greece
cppapagi@gmail.com

Grigorios Tsoumakas
School of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, Greece
greg@csd.auth.gr

Ioannis Tsamardinos[1,2]
[1]Computer Science Dept., Univ. of Crete, Greece
[2]Institute of Computer Science, FORTH
Voutes Campus, 700 13
Heraklion, Crete, Greece
tsamard.it@gmail.com

ABSTRACT

This work presents a probabilistic method for enforcing adherence of the marginal probabilities of a multi-label model to automatically discovered deterministic relationships among labels. In particular we focus on discovering two kinds of relationships among the labels. The first one concerns pairwise positive entailment: pairs of labels, where the presence of one implies the presence of the other in all instances of a dataset. The second concerns exclusion: sets of labels that do not coexist in the same instances of the dataset. These relationships are represented as a deterministic Bayesian network. Marginal probabilities are entered as soft evidence in the network and through probabilistic inference become consistent with the discovered knowledge. Our approach offers robust improvements in mean average precision compared to the standard binary relevance approach across all 12 datasets involved in our experiments. The discovery process helps interesting implicit knowledge to emerge, which could be useful in itself.

1. INTRODUCTION

Learning from multi-label data has received a lot of attention from the machine learning and data mining communities in recent years. This is partly due to the multitude of practical applications it arises in, and partly due to the interesting research challenges it presents, such as exploiting label dependencies, learning from rare labels and scaling up to large number of labels [33].

In several multi-label learning problems, the labels are organized as a tree or a directed acyclic graph, and there exist approaches that exploit such structure [34, 2]. However, in most multi-label learning problems, flat labels are only provided without any accompanying structure. Yet, it is often the case that implicit deterministic relationships exist among the labels. For example, in the ImageCLEF 2011 photo annotation task [22], which originally motivated the present study, the learning problem involved 99 labels without any accompanying semantic meta-data, among which certain deterministic relationships did exist. In particular, there were several groups of mutually exclusive labels, such as the four seasons *autumn, winter, spring, summer* and the person-related labels *single person, small group, big group, no persons*. There were also several positive entailment (consequence) relationships, such as $river \rightarrow water$ and $car \rightarrow vehicle$. Hierarchies accompanying multi-label data model positive entailment via their is-a edges, but do not model exclusion relationships.

These observations motivated us to consider the automated learning of such deterministic relationships as potentially interesting and useful knowledge, and the exploitation of this knowledge for improving the accuracy of multi-label learning algorithms. While learning and/or exploiting deterministic relationships from multi-label data is not new [24], little progress has been achieved in this direction since then. Past approaches exhibit weaknesses such as being unsuccessful in practice [24], lacking formal theoretical grounding [20, 19] and being limited to existing is-a relationships [2].

Given an unlabeled instance x, multi-label models can output a bipartition of the set of labels into relevant and irrelevant to x, a ranking of all labels according to relevance with x, marginal probabilities of relevance to x for each label or even a joined probability distribution for all labels. The latter is less popular due to the exponential complexity it involves [7]. Among the rest, marginal probabilities are information richer, as they can be cast into rankings after tie breaking and into bipartitions after thresholding. They are also important if optimal decision making is involved in the application at hand, which is often the case.

This work presents a probabilistic method for enforcing adherence of the marginal probabilities of a multi-label model to automatically discovered deterministic label relationships. We focus on two kinds of relationships. The first concerns pairwise *positive entailment*: pairs of labels, where presence of one label implies presence of the other in all instances of a dataset. The second concerns *exclusion*: sets of labels that do not coexist at the same instances of a dataset. These relationships are represented as a deterministic Bayesian network. Marginal probabilities are entered as soft evidence in the network and adjusted through probabilistic inference in order to become consistent with the discovered back-

KDD '15 August 11-14, 2015, Sydney, NSW, Australia
ACM 978-1-4503-3664-2/15/08.
http://dx.doi.org/10.1145/2783258.2783302.

ground knowledge. Our approach offers robust improvement in mean average precision compared to the standard binary relevance approach across all 12 datasets involved in our experiments. The discovery process helps interesting implicit knowledge to emerge, which could be useful in itself.

The rest of this paper is organized as follows. Section 2 introduces our approach. Section 3 presents related work and contrasts it with our approach. Section 4 presents empirical results and Section 5 summarizes the conclusions of this work and suggests future work directions.

2. OUR APPROACH

2.1 Discovering Entailment Relationships

Let A and B be two labels with domain $\{false, true\}$. For simplicity, we will be using the common shortcut notation a, $\neg a$, b and $\neg b$ instead of $A = true$, $A = false$, $B = true$ and $B = false$ respectively. The following four entailment relationships can arise between the two labels:

1. $a \rightarrow b$ and equivalent contrapositive $\neg b \rightarrow \neg a$
2. $b \rightarrow a$ and equivalent contrapositive $\neg a \rightarrow \neg b$
3. $a \rightarrow \neg b$ and equivalent contrapositive $b \rightarrow \neg a$
4. $\neg a \rightarrow b$ and equivalent contrapositive $\neg b \rightarrow a$

The first two express positive entailment, the third one expresses exclusion and the forth one expresses co-exhaustion. Figure 1 presents a contingency table for labels A and B, based on a multi-label dataset with $S + T + U + V$ training examples. Positive entailment corresponds to $T = 0$ or $U = 0$, exclusion to $S = 0$ and coexhaustion to $V = 0$. Furthermore, $S = V = 0$ corresponds to mutually exclusive and completely exhaustive labels, while $T = U = 0$ corresponds to equivalent labels.

In this work, we focus on discovering *pairwise* positive entailment relationships as well as exclusion relationships among *two or more* labels. For a multi-label dataset with q labels, it is easy to extract all four types of pairwise entailment relationships from the corresponding contingency tables in $O(q^2)$ time complexity. For discovering exclusion relationships among more than two labels, we follow the paradigm of the Apriori algorithm [1] in order to find all maximal sets of mutually exclusive labels, such that each of them is not a subset of another. Starting from the pairwise exclusion relationships, we find triplets of mutual exclusive labels, then quads and so on.

As a toy example, consider the label values of a multi-label dataset with 6 labels that are given in Table 1, where to improve legibility we have used a value of 1 to represent *true* and a value of 0 to represent *false*. Our approach would in this case extract the positive entailment relationships $a \rightarrow b$, $a \rightarrow c$, $b \rightarrow c$ and $d \rightarrow c$, and an exclusion relationship for the set of labels $\{A, E, F\}$.

	a	$\neg a$
b	S	T
$\neg b$	U	V

Figure 1: Contingency table for labels A and B

Table 1: A toy multi-label dataset with 10 samples and 6 labels.

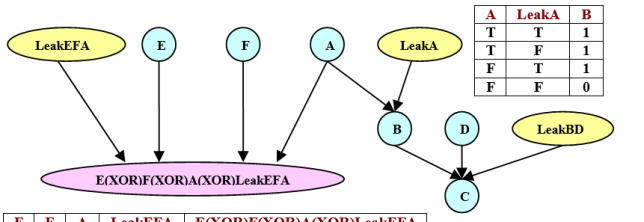
A	B	C	D	E	F
1	1	1	0	0	0
1	1	1	1	0	0
0	0	0	0	1	0
0	1	1	0	1	0
1	1	1	0	0	0
0	1	1	1	0	1
0	0	1	1	1	0
0	0	0	0	0	1
0	0	1	0	0	0
0	0	0	0	0	1

2.2 Exploiting Entailment Relationships

Motivated by the goal of theoretically sound correction of the marginal probabilities $P_x(\lambda)$, obtained by a multi-label model for each label λ given instance x, according to the background knowledge expressed by the discovered relationships, we propose using a deterministic Bayesian network for the representation of these relationships. This network initially contains as many nodes as the labels, with each node representing the conditional probability $P_x(\lambda)$ of corresponding label λ, given instance x, with uniform prior.

We represent the entailment relationship $a \rightarrow b$, among labels A and B, by adding a link from node A to node B. We set the conditional probability table (CPT) associated with node B to contain probabilities $P_x(b|a) = 1$ and $P_x(b|\neg a) = 0$. This is easily generalized in the case of multiple relationships $a_1 \rightarrow b, \ldots, a_k \rightarrow b$, to a CPT with $P_x(b|A_1, \ldots, A_k) = 1$ if $A_1 \vee \ldots \vee A_k = true$ and $P_x(b|A_1, \ldots, A_k) = 0$ otherwise (i.e. when $A_1 \vee \ldots \vee A_k = false$ or equivalently when $\neg A_1 \wedge \ldots \wedge \neg A_k = true$). Such a CPT renders node B deterministic. Our representation assumes that all causes of B have been considered, which is seldom true for typical multi-label datasets. To deal with this discrepancy, we add an additional parent of B as *leak* node, corresponding to a new virtual label whose value is set to: (i) *false* in those training examples where $B = false$, (ii) *true* in those training examples where $B = true$ and all other parents of B are *false*, and (iii) *false* for the rest of the training examples. Note that for the last category of examples where $B = true$ and at least one other parent is *true*, the value of the leak node could also be set to *true* instead of false, but the choice of *false* should lead to semantically simpler virtual labels that are easier to learn. Redundant relationships due to the transitivity property of positive entailment are not represented in the network.

We represent the mutual exclusion relationship among labels $A_1, \ldots, A_k$ by adding a new boolean deterministic node B as common child of all of these labels. We set the CPT of this node to contain probabilities $P_x(b|A_1, \ldots, A_k) = 1$ if one and only one of the parents is true and $P_x(b|A_1, \ldots, A_k) = 0$ otherwise. We consider that this node is true as observed evidence ($B = true$). Our representation assumes that labels $A_1, \ldots, A_k$ cover all training examples, which usually is not the case for typical multi-label datasets. We deal with this discrepancy similarly to the case of positive entailment. In specific, we add an additional parent of B as *leak* node, corresponding to a new virtual label whose value is set to:



A	LeakA	B
T	T	1
T	F	1
F	T	1
F	F	0

E	F	A	LeakEFA	E(XOR)F(XOR)A(XOR)LeakEFA
T	T	T	T	0
T	T	T	F	0
T	T	F	T	0
T	T	F	F	0
T	F	T	T	0
T	F	T	F	0
T	F	F	T	0
T	F	F	F	1
F	T	T	T	0
F	T	T	F	0
F	T	F	T	0
F	T	F	F	1
F	F	T	T	0
F	F	T	F	1
F	F	F	T	1
F	F	F	F	0

B	D	LeakBD	C
T	T	T	1
T	T	F	1
T	F	T	1
T	F	F	1
F	T	T	1
F	T	F	1
F	F	T	1
F	F	F	0

Figure 2: A network that represents labels A, B, C, D, E, F, entailment relationships $a \rightarrow b$, $a \rightarrow c$, $b \rightarrow c$, $d \rightarrow c$ and a mutual exclusion relationship among labels A, E and F.

(i) *true* in those training examples where all other parents of B are *false*, and (ii) *false* in all other training examples.

Continuing the toy example of the previous section, Figure 2 shows the network that our approach would construct to represent the discovered knowledge.

Having constructed the network, we then use any multi-label algorithm that can provide marginal probability estimates to fit the extended (with virtual labels corresponding to leak nodes) training set. For an unlabelled instance x, we first query the multi-label model in order to obtain probability estimates $P_x(\lambda)$ for each of the labels λ, including virtual labels. These are then entered into the network as *soft* (also called *virtual*) evidence [14]. A probabilistic inference algorithm is then used to update the probability estimates for each of the labels, leading to probabilities that are consistent with the discovered relationships. The network is then reset to its prior state in order to process a subsequent unlabelled instance.

Table 2 exemplifies the probability correction process. Each row corresponds to a particular label/node. The second column contains arbitrary probability estimates for a test instance. These probabilities violate the relationships we have discovered. The probability of label C should be larger than that of B, D and $LeakBD$, but only the last constraint holds. In addition, the probability of B should be larger than that of A and $LeakA$, none of which is satisfied. Finally, the probabilities of A, E, F and $LeakEFA$ should add to 1, which does not hold. The third column gives the adjusted probabilities according to our approach, which are now consistent with the discovered relationships.

3. RELATED WORK

We divide the related work according to the uncertainty of relationships considered (discovered and exploited, or simply exploited). We first discuss approaches that similar to us focus on deterministic relationships and we then continue to approaches centered on probabilistic relationships.

Table 2: Marginal probabilities obtained by the multi-label model for each of the labels, including the virtual ones corresponding to leak nodes (before) and updated probabilities after probabilistic inference (after) for the network in Figure 2.

Node	Before	After
A	0.4	0.022
$LeakA$	0.35	0.082
B	0.25	0.096
D	0.6	0.031
$LeakBD$	0.01	0.05
C	0.2	0.345
F	0.3	0.064
E	0.85	0.85
$LeakEFA$	0.3	0.064

3.1 Deterministic Relationships

The idea of discovering and exploiting label relationships from multi-label data was first discussed in [24], where relationships were referred to as *constraints*. An interesting general point of [24] was that label constraints can be exploited either at the learning phase or at a post-processing phase. In addition, it presented four basic types of constraints, which correspond to the four entailment relationships, and noted that more complex types of constraints can be represented by combining these basic constraints with logical connectors. For discovering constraints, it proposed association rule mining, followed by removal of redundant rules that were more general than others. For exploiting constraints, it proposed two post-processing approaches for the label ranking task in multi-label learning. These approaches correct a predicted ranking when it violates the constraints by searching for the nearest ranking that is consistent with the constraints. They only differ in the function used to evaluate the distance between the invalid and a valid ranking. As they focus on label ranking, these approaches cannot be used for correcting the marginal probability estimates of the labels. Results on synthetic data with known constraints showed that constraint exploitation can be helpful, but results on real-world data and automatically discovered constraints did not lead to predictive performance improvements.

An approach for exploiting a set of given label relationships in order to update the marginal probabilities was presented in [20, 19]. For mutually exclusive labels that cover all training examples, the idea was to keep the highest probability and set the rest to zero. For mutually exclusive labels that don't cover all training examples, this rule was used if the highest probability was larger than a threshold. For an entailment relation $a \rightarrow b$, the idea was to set the marginal probability of b to that of a when the marginal probability of a is larger than that of b and larger than 0.5. This post-processing approach, called *concept reasoning*, leads to marginal probabilities that are consistent with the label relationships, but lacks a sound probabilistic framework.

The exploitation of parent-child relationships of a *given* hierarchical taxonomy of labels via a Bayesian network structure was proposed in [2]. In particular, parent labels were conditioned on their child labels, and the output of the binary classifier for each label was conditioned on the corresponding label. Given the classifier outputs as evidence,

a probabilistic inference algorithm is then applied to obtain hierarchically consistent marginal probabilities for each label. Our approach shares the same principle of using a Bayesian network structure for enforcing consistency of marginal probabilities, but: (i) constructs the structure automatically according to relationships discovered from the data, (ii) represents mutual exclusion relationships in addition to positive entailment, and (iii) builds additional binary models for virtual labels corresponding to leak nodes in the deterministic relations represented by the network.

A method for uncovering deterministic causal structures is introduced in [3]. Similarly to our work, it aims at constructing a Bayesian network out of automatically discovered deterministic relationships. Important differences are that it does not consider latent variables, as in our representation of exclusions and our treatment of unaccounted causes of a label via leak nodes. It therefore requires relationships to be supported from the full dataset, which limits its practical usefulness, as rarely such relationships appear in real-world data.

A recent approach that similar to us stresses the aspect of discovering interesting knowledge about the labels in multi-label learning is based on learning rule-based models [17]. In particular, for each label a separate rule model is constructed, but this model uses the rest of the labels as additional features. This can lead to rules that include labels with/without ordinary input features in the preconditions of a rule deriving a particular label. Such rules are a natural representation of entailment relationships among labels (and input features). This approach combines discovery and exploitation of entailment relationships (that are not guaranteed to only include labels) through a specific family of learning algorithms (rule learning). This is very different from obtaining marginal probabilities that adhere to a consistent set of deterministic relationships of the labels, which is the core contribution of our approach.

Another recent approach [8] addresses similar to us the important problem of exploiting deterministic relationships among the labels in a principled probabilistic way. It represents the full joint distribution using a conditional random field (CRF) model, that takes into consideration independent classification scores for the labels, as well as all given *pairwise* exclusion and hierarchy relationships. As the generic CRF model is intractable, it proposed an efficient inference procedure, under the assumption that there are many mutual exclusion relationships, typical of the object classification domain the paper was focusing on.

3.2 Probabilistic Relationships

A Bayesian network structure to encode the relationships among labels as well as between the input attributes and the labels was presented in [36]. The proposed algorithm, called LEAD, starts by building binary relevance models and continues by learning a Bayesian network on the residuals of these models. Then another set of binary models is learned, one for each label, but this time incorporating the parents of each label according to the constructed Bayesian network as additional features. For prediction, these models are queried top-down according to the constructed Bayesian network.

A similar approach that is a bit closer to our work [35] starts by using a Bayesian network structure learning method in order to learn the structure among the labels. Then, maximum likelihood estimation is used to learn the conditional probabilities among the labels. Parent-child relationships with high (low) conditional probabilities are called co-existence (mutual exclusion) relationships. Inference is the used to obtain the joined probability distribution for all labels given prediction by a base multi-label model. A serious limitation of this approach is that the initial structure learning phase is computationally demanding and prohibits its scaling to practical numbers of labels. Indeed, results are only given for 4 datasets with up to 14 labels each.

Both LEAD and [35] discover *probabilistic* parent-child relationships among labels by learning the Bayesian network structure from the data. LEAD does this implicitly (through the residuals of binary models), while [35] does it explicitly. In contrast, our approach discovers *deterministic* parent-child and mutual exclusion relationships among labels from the data, which are then used to define the structure of a corresponding deterministic Bayesian network.

Other approaches to modeling probabilistic relationships include conditional dependency networks [11] and multi-dimensional Bayesian classifiers [4].

3.3 Discussion

We would first like to mention that our approach was developed in parallel and independently of recent approaches [8], [35] and [17]. Indeed, we have published an earlier version of this work at arXiv on April, 2014 [23] .

We would also like to clarify that we do not claim scientific novelty for the approach we use to discover the deterministic relationships among the labels. Indeed, our approach is based on a simple contingency table for positive entailment and follows the paradigm of association rule mining for the exclusion relationships. We argue however, that this is the first paper to explicitly discuss and present actual interpretable examples of deterministic relationships in several multi-label datasets, after many years of multi-label research motivated from exploiting label relationships. From the very long list of such papers, we here point to just a few [10, 34, 7, 26, 18, 37, 16].

Our main novelty is in the approach taken to represent these relationships, i.e. a *deterministic* Bayesian network using virtual leak nodes to enforce consistency of the deterministic CPTs of the network with the data. Learning deterministic Bayesian networks from data is non-trivial [3] and research in this topic is ongoing. One could argue that probabilistic relationships could be more interesting. However, besides being less interpretable, past research on ordinary probabilistic Bayesian networks for multi-label data had either scaling problems [35] or was not directly modeling potentially interesting semantic relationships, such as entailment and exclusion, as it was working on the residuals of binary models corresponding to labels [36].

4. EXPERIMENTS

4.1 Setup

We use the binary relevance (BR) problem transformatiom method for learning multi-label models, which learns one binary model per label. As our approach employs probabilistic inference to *correct* the marginal probability of each label, we consider important to start with good probability estimates. We therefore use Random Forest [6] (10 trees) as the learning algorithm for training each binary model, since it is known to provide relatively accurate probability esti-

mates without calibration [21]. We use the implementations of BR and Random Forest from Mulan [32] and Weka [12] respectively. Our approach is also implemented in Mulan, utilizing the jSMILE library[1] for Bayesian network representation and probabilistic inference. The default clustering algorithm [15] was used for exact probabilistic inference.

We experiment on the 12 multi-label datasets that are shown in Table 3. We adopt a 10-fold cross-validation process and present results in terms of mean average precision (MAP) across all labels, as this was also the measure of choice in the ImageCLEF 2011 challenge that motivated this work. MAP is also the standard evaluation measure in multimedia information retrieval. We also briefly summarize the results for the 27 additional measures of Mulan that due to space limitations are made available online [2].

We set the *minimum support* of discovered relationships to just 2 training examples (avoid single-point generalization). For positive entailment, *support* refers to the positive training examples of the antecedent label, while for exclusion it refers to the sum of the positive examples of all participating labels (S and $T + U$ in Figure 1 respectively). In 3 datasets (Bibtex, Bookmarks, Medical) exclusion discovery did not finish within a week, while in 3 other datasets (Enron, ImageCLEF2011/2012) a large number of exclusion rules was discovered that caused memory outage during network construction in jSMILE. We increased the support exponentially (4, 8, 16, ...) until these issues were resolved. In Section 5, we discuss ideas for automatically selecting appropriate support values towards improving the efficiency and effectiveness of our approach.

Towards repeatability of our experiments and open science, we make available all our source code for the proposed approach, including third-party libraries and instructions on what scripts to execute to replicate our results[4]. We intend to formally release the code of our approach with the next version of Mulan.

4.2 Relationships

This section discusses the relationships discovered by our approach. At each fold of the cross-validation, different relationships can be discovered. Table 3 reports the mean of the discovered relationships across all folds. We only discuss here those appearing in all folds. Tables presenting positive entailment relationships include a column mentioning the relationship support.

4.2.1 Bibtex and Bookmarks.

The labels of the *Bibtex* and *Bookmarks* datasets correspond to tags assigned to publications and bookmarks respectively by users of the social bookmark and publication sharing system Bibsonomy[5].

Table 4 presents the 11 positive entailments that were found in all folds for *Bibtex*. These apparently correspond to a hierarchy relationship between label *statphys23*, an international conference on statistical physics[6] and the 11 topics of this conference. It could be that the user(s) that added publications from this conference renamed label *topic5* to the more descriptive *nonequilibrium*, but did not bother for the rest of the topics. However, on the conference site, topic 5 is listed as *Dynamical systems and turbulence*, while *Nonequilibrium systems* is topic 3 of the conference.

Table 5 presents the 4 positive entailments that were discovered in all folds for *Bookmarks*. The first two most probably belong to a single user who also used the tags *film* and *kultur* whenever he/she used the tag *filmsiveseenrecently*. This is, by the way, an example of an unfortunate choice of tag name, as it involves a time adverb *recently*, whose meaning changes over time. The last two are examples of discovered is-a relationships, as paddling *is-a* (water) sport. We conclude that our approach manages to discover positive entailment relationships of *social* origin.

A minimum support of 128 and 2048 examples led to a mean number of 76.2 and 1 exclusion relationships per fold in *Bibtex* and *Bookmarks* respectively. Due to space limitations we refrain from reporting the 18 exclusions discovered in all folds of *Bibtex*. In *Bookmarks* the relationship involved the following pair of labels: {*computing*; *video*}. No wonder it did not help improve accuracy.

4.2.2 Emotions.

The 6 labels in the *Emotions* dataset concern 3 pairs of opposite emotions of the Tellegen-Watson-Clark model of mood[29]: (*quiet-still*, *amazed-surprised*), (*sad-lonely*, *happy-pleased*) and (*relaxing-calm*, *angry-aggresive*) that correspond to the axes of engangement, pleasantness and negative affect respectively. The Tellegen-Watson-Clark model includes a fourth axis that concerns positive affect. The single discovered exclusion relationship, concerns the pair of opposite labels related to engangement: {*quiet-still*; *amazed-surprised*}.

4.2.3 Enron.

The 53 labels of this dataset are organized into 4 categories: *coarse genre*, *included/forwarded information*, *primary topics*, which is applicable if coarse genre *Company Business, Strategy, etc* is selected, and *emotional tone* if not neutral. There are 13 positive entailment relationships by definition, as there are 13 labels in the *primary topics* category, which are children of label *Company Business, Strategy, etc*.

Table 6 presents the 3 positive entailment relationships that were discovered in all folds. Relationship 1 is among the 13 positive entailments we already knew from the description of the labels, as label *company image – changing / influencing* is a primary topic and therefore a child of label *Company Business, Strategy, etc*. Our approach manages to discover explicit is-a relationships, when these are present in the training data.

A minimum support of 8 examples led to a mean number of 480.7 exclusion relationships per fold. Only one relationship was present in all folds and involved the following interesting pair of concepts {*Company Business, Strategy, etc*; *friendship / affection*}. The conclusion is that there is no room for affection in the business world.

4.2.4 ImageCLEF 2011 and 2012.

Labels in these two datasets correspond to 99 and 94 concepts respectively covering a variety of concepts for image annotation. A difference in the 2012 version of the contest was that concepts being superclasses of other concepts (e.g. *Animal*, *Vehicle* and *Water*) were removed. This is why in

[1] http://genie.sis.pitt.edu/
[2] http://mlkd.csd.auth.gr/kdd2015.xlsx
[3] http://bailando.sims.berkeley.edu/enron_email.html
[4] http://mlkd.csd.auth.gr/kdd2015.zip
[5] http://www.bibsonomy.org/
[6] http://www.statphys23.org/

Table 3: A variety of multi-label datasets and their statistics: number of labels, examples, discrete and continuous features, followed by the mean number of positive entailment and exclusion relationships that were discovered across the 10 training sets of the 10-fold cross-validation process.

				variables		entailment	
dataset	source	labels	examples	disc.	cont.	positive	exclusion
Bibtex	[13]	159	7395	1836	0	11±0	76.2 ± 2.3
Bookmarks	[13]	208	87856	2150	0	4.1 ± 0.3	1 ± 0
Emotions	[31]	6	593	0	72	0	1.1±0.3
Enron	url[3]	53	1702	1001	0	13 ± 15.2	480.7 ± 98.4
ImageCLEF2011	[22]	99	8000	19540	0	27.9 ± 0.9	325.4 ± 31.9
ImageCLEF2012	[30]	94	15000	10000	0	1.2	277.9 ± 43.5
IMDB	[26]	28	120919	0	1001	0	21.6 ± 1.2
Medical	[25]	45	978	1449	0	6.3 ± 1	30.7 ± 7.2
Scene	[5]	6	2407	0	294	0	4 ± 0
Slashdot	[26]	20	3782	0	1079	0	23.2 ± 1.2
TMC2007	[28]	22	28596	49060	0	0	7.5 ± 1.1
Yeast	[9]	14	2417	0	103	3 ± 0	2.3 ± 0.5

Table 4: Positive entailment relationships discovered in the *Bibtex* dataset.

id	relationship			sup	id	relationship			sup	id	relationship			sup
1	nonequilibrium	→	statphys23	68	5	topic4	→	statphys23	62	9	topic9	→	statphys23	82
2	topic1	→	statphys23	86	6	topic6	→	statphys23	63	10	topic10	→	statphys23	130
3	topic2	→	statphys23	75	7	topic7	→	statphys23	129	11	topic11	→	statphys23	143
4	topic3	→	statphys23	151	8	topic8	→	statphys23	73					

the 2011 version of the dataset 27 positive entailments were found in all folds (see Table 7), in contrast with the following single one in the 2012 version: *Spider* → *QuantityNone*, with a support of 16 examples. The consequent of this relationship refers to the number of people that appear in the photo. In other words, no people appear in the 16 spider pictures of that photo collection.

A minimum support of 32 and 64 examples led to a mean number of 325.4 and 277.9 exclusion relationships per fold in the 2011 and 2012 version of *ImageCLEF* respectively. Due to space limitations we refrain from reporting the 24 and 21 exclusion relationships that were discovered in all folds of the 2011 and 2012 version of *ImageCLEF* respectively.

4.2.5 IMDB.

Labels of this dataset correspond to 28 movie genres. Table 8 presents the 15 exclusion relationships that were found in all folds. Labels *Film-Noir*, *Game-Show* and *Talk-Show* and are the most frequent ones in these relationships. We notice some obvious exclusions, such as {*War*, *Reality-TV*}, {Game-Show, Crime} and {*Talk-Show*, *Fantasy*}. On the other hand, such relationships could be a source of inspiration for innovative (or provocative) producers and directors contemplating unattempted combinations of genres.

4.2.6 Medical.

Labels of this dataset correspond to 45 codes/descriptions of the 9th revision of the International Statistical Classification of Diseases (ICD). Table 9 presents the 3 positive entailment relationships that were found in all folds. The support of these relationships is quite weak (up to 4 examples), yet it apparently corresponds to valid, yet already known, medical knowledge. For example, the top page returned by Google for the query "hydronephrosis congenital obstruction of ureteropelvic junction", contains the following excerpt: *Ureteropelvic junction obstruction is the most common pathologic cause of antenatally detected hydronephrosis.* Future work could apply our approach to larger datasets in search of unknown medical knowledge.

A minimum support of 16 examples led to a mean number of 30.7 exclusion relationships per fold. Only one relationship was present in 9 out of the 10 folds (none in all folds) and involved the following 5 concepts: {753.0 Renal agenesis and dysgenesis; 599.0 Urinary tract infection, site not specified; 596.54 Neurogenic bladder NOS; 780.6 Fever and other physiologic disturbances of temperature regulation; 493.90 Asthma,unspecified type, unspecified}.

4.2.7 Scene.

Labels in this dataset correspond to 6 different scenery concepts. The following 2 exclusion relationships were found in all folds: {Sunset, Fall Foliage, Beach} and {Sunset, Fall Foliage, Urban}. These relationships do not really correspond to interesting knowledge, as images with such concept combinations can be found on the Web. However, only a few of the top 30 images returned by Google image search do cover all three concepts of these relationships, a fact that highlights the co-occurence rarity of these concepts. Still, this knowledge should not be generalized beyond the particular limited collection of 2407 images from the stock photo library of Corel.

4.2.8 Slashdot.

Labels in this dataset correspond to 20 topical categories of news articles. The following 4 exclusion relationships were found in all folds: {Interviews, Apache, News, BSD, Idle, AskSlashdot}, {Interviews, Apache, Search, BSD, Idle, AskSlashdot}, {Apache, Search, Science, BookReviews, Linux}

Table 5: Positive entailment relationships discovered in the *Bookmarks* dataset.

id	relationship			sup	id	relationship			sup
1	filmsiveseenrecently	→	film	370	3	paddling	→	sports	379
2	filmsiveseenrecently	→	kultur	370	4	paddling	→	watersports	379

Table 6: Positive entailment relationships discovered in the *Enron* dataset.

id	relationship			sup
1	company image – changing / influencing	→	Company Business, Strategy, etc.	63
2	triumph / gloating	→	Company Business, Strategy, etc.	3
3	triumph / gloating	→	regulations and regulators (includes price caps)	3

and {Apache, Search, BSD, Games, BookRevies}. This knowledge does not seem particularly interesting.

4.2.9 *TMC2007.*

Labels in this dataset correspond to problems that might occur during flights and come from the database of NASA's aviation safety reporting system. Unfortunately, we could not retrieve the actual label semantics.

4.2.10 *Yeast.*

Labels in this dataset correspond to 14 particular gene functional classes from the FunCat hierarchy [27], but unfortunately we do not know the one-to-one match of these functional classes with the variables in the dataset. Personal communication on this issue with the authors of [9] did not resolve the problem, despite their positive response and effort to help[7].

4.3 Results

Tables 10 and 11 present the average MAP of BR across the 10 folds of the cross-validation: (i) in its standard version, and (ii) with the exploitation of *positive entailment* relationships (Table 10) and *exlusion* relationships (Table 11) via our approach. They also present the percentage of improvement brought by our approach, the minimum support and the average number of discovered relationships across the 10 folds of the cross-validation.

In Table 10 we notice that in all datasets where positive entailments were discovered, the exploitation of these relationships led to an increased MAP. Applying the Wilcoxon signed rank test we find a p-value of 0.0156, indicating that the improvements are statistically significant. In some datasets, such as *bookmarks*, improvements are small, while in others, such as *ImageCLEF2011*, improvements are large. The correlation coefficient between the percentage of improvement and the number of relationships divided by the number of labels is 0.913, which supports the argument that the more relationships we discover per label, the higher the improvements in MAP. This was expected to an extend as MAP is a mean of the average precision across *all* labels. Had we focused on just the affected labels, which could be considered as a more appropriate way to evaluate the relationship exploitation part of our approach, we would notice larger improvements.

In the upper part of Table 11 we notice that there are

[7]If you happen to know the actual labels, please communicate them to us.

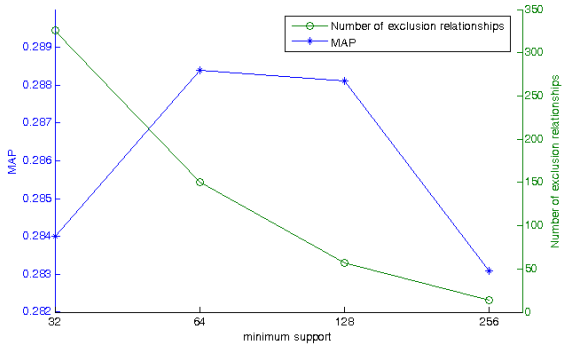


Figure 3: MAP and number of exclusion relationships for the 4 different minimum support values we tried in *ImageCLEF2011*.

8 datasets where our approach leads to improvements in MAP, but another 4 where it leads to reductions. The p-value of the Wilcoxon signed rank test is 0.1099, indicating that the results are statistically insignificant even at the 0.1 level (though marginally). One of the main reasons underlying the negative results of our approach is the large number of spurious exclusion relationships that can be discovered in datasets with a larger number of labels. Indeed, the rarer two labels are, the higher the probability of being considered as mutually exclusive, irrespectively of their actual semantic relationship. We therefore further experiment with exponentially increasing minimum support values in *Bibtex*, *Enron* and the *ImageCLEF* datasets, where a large number of relationships was discovered, until we reach a small set of relationships. The bottom part of Table 11 shows the results achieved through this process. We see that MAP improvements are now achieved for *Bibtex*, *Enron* and *ImageCLEF2012*, while the already improved MAP of *ImageCLEF2011* increases. Note that these are not the best achievable results, both because we only tried a few minimum support values and because our goal was to discover a small number of relationships. Figure 3 shows the MAP and number of exclusion relationships for the minimum support values we tried in *ImageCLEF2011*.

The number of exclusion relationships divided by the number of labels is again positively correlated with the percentage of improvement, with the coefficient being 0.770 in this case (based on the bottom part of Table 11 for the 4 datasets and ignoring *Bookmarks*). The coefficient would probably

Table 7: Positive entailment relationships discovered in the *ImageCLEF2011* dataset.

id	relationship	sup	id	relationship	sup	id	relationship	sup
1	Desert → Outdoor	30	10	Sea → Outdoor	222	19	Fish → Animals	25
2	Desert → Day	30	11	Sea → Water	222	20	Bicycle → NeutralLight	61
3	Spring → NeutralLight	105	12	Cat → Animals	61	21	Bicycle → Vehicle	61
4	Flowers → Plants	367	13	Dog → Animals	211	22	Car → Vehicle	268
5	Trees → Plants	890	14	Horse → Outdoor	28	23	Train → Vehicle	59
6	Clouds → Sky	1104	15	Horse → Day	28	24	Airplane → Vehicle	41
7	Lake → Outdoor	89	16	Horse → Animals	28	25	Skateboard → Vehicle	12
8	Lake → Water	89	17	Bird → Animals	183	26	Ship → Vehicle	79
9	River → Water	130	18	Insect → Animals	91	27	Female → Male	1254

Table 8: Exclusion relationships discovered in the *IMDB* dataset.

{Film-Noir, Game-Show, Adult, News}, {Film-Noir, Adult, Family}, {Game-Show, Horror},
{Film-Noir, Game-Show, Western}, {Film-Noir, Documentary}, {Game-Show, Thriller}
{Film-Noir, Talk-Show, Western}, {Talk-Show, Adventure}, {Game-Show, Crime}
{Game-Show, Adult, Biography}, {Film-Noir, Comedy}, {War, Reality-TV}
{Film-Noir, Western, Reality-TV}, {Talk-Show, Mystery}, {Talk-Show, Action}

Table 9: Positive entailment relationships discovered in the *Medical* dataset.

id	relationship			sup
1	753.21 Congenital obstruction of ureteropelvic junction	→	591 Hydronephrosis	4
2	786.05 Shortness of breath	→	753.0 Renal agenesis and dysgenesis	4
3	787.03 Vomiting alone	→	753.0 Renal agenesis and dysgenesis	3

Table 10: Mean MAP of BR with and without exploitation of positive entailments via our approach, along with the percentage of improvement, the minimum support and the average number of discovered relationships.

dataset	standard BR	minsup	positive entailment	impr%	#relations
Bibtex	0.2152 ± 0.0114	2	0.2168 ± 0.0109	0.279	11 ± 0
Bookmarks	0.1474 ± 0.0041	2	0.1475 ± 0.0041	0.068	4.1 ± 0.3
Enron	0.2810 ± 0.0476	2	0.2821 ± 0.0480	0.391	3.8 ± 0.9
ImageCLEF2011	0.2788 ± 0.0113	2	0.2871 ± 0.0100	2.977	27.9 ± 0.9
ImageCLEF2012	0.2376 ± 0.0089	2	0.2380 ± 0.0084	0.168	1.2 ± 0.9
Medical	0.5997 ± 0.0768	2	0.6134 ± 0.0661	2.284	6.3 ± 1
Yeast	0.4545 ± 0.0145	2	0.4617 ± 0.0141	1.584	3 ± 0

Table 11: Mean MAP of BR with and without exploitation of exlusions via our approach, along with the percentage of improvement, the minimum support and the average number of discovered relationships.

dataset	standard BR	minsup	exclusion	impr%	#relations
Bibtex	0.2152 ± 0.0114	128	0.2117 ± 0.0126	-1.626	76.2 ± 2.3
Bookmarks	0.1474 ± 0.0041	2048	0.1473 ± 0.0040	-0.068	1 ± 0
Emotions	0.7163 ± 0.0339	2	0.7265 ± 0.0368	1.424	1.1 ± 0.3
Enron	0.2810 ± 0.0476	8	0.2573 ± 0.0476	-8.434	480.7 ± 98.4
ImageCLEF2011	0.2788 ± 0.0113	32	0.2840 ± 0.0132	1.865	325.4 ± 31.9
ImageCLEF2012	0.2376 ± 0.0089	64	0.2308 ± 0.0090	-2.862	277.9 ± 43.5
IMDB	0.0900 ± 0.0030	2	0.0938 ± 0.0026	4.222	21.6 ± 1.2
Medical	0.5997 ± 0.0768	16	0.6223 ± 0.0597	3.769	30.7 ± 7.2
Scene	0.8139 ± 0.0146	2	0.8385 ± 0.0140	3.023	4 ± 0
Slashdot	0.3982 ± 0.0323	2	0.4452 ± 0.0422	11.803	23.2 ± 1.2
TMC2007	0.3276 ± 0.0069	2	0.3474 ± 0.0075	6.044	7.5 ± 1.1
Yeast	0.4545 ± 0.0145	2	0.4625 ± 0.0163	1.760	2.3 ± 0.5
Bibtex	0.2152 ± 0.0114	256	0.2165 ± 0.0114	0.604	2.8 ± 0.4
Enron	0.2810 ± 0.0476	32	0.2816 ± 0.0477	0.214	22.2 ± 1.8
ImageCLEF2011	0.2788 ± 0.0113	128	0.2881 ± 0.0129	3.336	56.6 ± 3.3
ImageCLEF2012	0.2376 ± 0.0089	256	0.2391 ± 0.0087	0.631	39.8 ± 1.1

be higher had we calculated the actual number of labels involved in the exclusion relationships.

The average improvement offered by exclusion (3.4%) is larger than those offered by positive entailment (1.1%), and so is the average number of relationships (19.3 vs 8.2). Exclusions typically involve more than two labels, while positive entailments are pairwise and some of them redundant due to the transitivity property of positive entailment. A deeper analysis should look at the number of labels involved in each type of relationship, which we leave as future work.

Table 12 shows results of utilizing both types of relationships. We notice that in *Bibtex*, *Yeast*, *Medical* and *ImageCLEF2012* the combination of both types of relationships leads to larger improvement than their individual improvement. The combined improvement is smaller than the sum of individual improvements, with the exception of *Bibtex*. This could be due to labels appearing in both types of relationship. For *Enron* and *ImageCLEF2011* we notice that the combined improvement is smaller than the largest of the two individual improvements, that of positive entailment. This is another indication of spurious exclusions existence.

Similar improvements are noticed for most bipartition-based measures, such as Hamming loss, subset accuracy, example-based F-measure and micro-averaged F-measure, particularly with the exclusion relationships. No clear conclusions can be reached for label ranking measures, as for some datasets there are benefits, while for others losses.

5. SUMMARY AND FUTURE WORK

This work introduced an approach that discovers entailment relationships among labels within multi-label datasets and exploits them using a probabilistic technique that enforces the adherence of the marginal probability estimates of multi-label learning with the discovered knowledge.

Our approach can be extended in a number of directions. An important issue concerns the statistical validity of the extracted relationships, especially when based on infrequent labels. We are working on automatically selecting the minimum support per relationship based on suitable statistical significance tests in order to separate chance artifacts from confident findings. We expect this to both improve accuracy results and reduce the complexity of the discovery process. On the opposite direction, it would be also interesting to investigate whether approximate relations, where the contingency table frequencies are not necessarily zero due to noise, can lead to improved results. Another important direction is the generalization of our approach to be able to discover all types of entailment among any number of labels.

On the empirical part of this work we intend to work harder towards assessing the relative performance of our approach compared to the recent related work [35, 8]. This is non-trivial, as they don't offer their full experiment code for replication. Moreover, results of the latter were given only on a single domain of object classification, while the former's computational complexity with respect to the number of labels prohibits its application to more realistic multi-label datasets. We also intend to investigate how our approach performs when used in conjunction with a base multi-label classifier that can output marginal probabilities and already exploits label dependencies. Preliminary experiments with Classifier Chains [26] showed larger benefits for our approach. Finally, we intend to investigate the effect that the quality of predicted probabilities has on our approach.

6. ACKNOWLEDGEMENTS

IT was partially funded by the EPILOGEAS GSRT ARISTEIA II project, No 3446 and ERC Consolidator Grant CAUSALPATH, No617393.

7. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[2] Z. Barutcuoglu, R. E. Schapire, and O. G. Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

[3] M. Baumgartner. Uncovering deterministic causal structures: a boolean approach. *Synthese*, 170(1):71–96, 2009.

[4] C. Bielza, G. Li, and P. LarraÃśaga. Multi-dimensional classification with bayesian networks. *International Journal of Approximate Reasoning*, 52(6):705 – 727, 2011.

[5] M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[6] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

[7] K. Dembczynski, W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

[8] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 48–64, 2014.

[9] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems 14*, 2002.

[10] J. Fürnkranz, E. Hüllermeier, E. L. Mencia, and K. Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.

[11] Y. Guo and S. Gu. Multi-label classification using conditional dependency networks. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1300–1305. AAAI Press, 2011.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11, 2009.

[13] I. Katakis, G. Tsoumakas, and I. Vlahavas. Multilabel text classification for automated tag suggestion. In *Proceedings of the ECML/PKDD 2008 Discovery Challenge*, Antwerp, Belgium, 2008.

[14] K. Korb and A. E. Nicholson. *Bayesian Artificial Intelligence*. CRC Press, Inc., Boca Raton, FL, USA, 2003.

[15] S. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical

Table 12: Mean MAP of BR with and without exploitation of both types of relationships via our approach, along with the percentage of improvement.

dataset	standard BR	minimum support positive	minimum support exclusion	our approach	impr%
Bibtex	0.2152 ± 0.0114	2	256	0.2181 ± 0.0103	1.347
Bookmarks	0.1474 ± 0.0041	2	2048	0.1474 ± 0.0040	0
Enron	0.2810 ± 0.0476	2	32	0.2816 ± 0.0477	0.214
ImageCLEF2011	0.2788 ± 0.0113	2	32	0.2846 ± 0.0129	2.080
ImageCLEF2012	0.2376 ± 0.0089	2	256	0.2393 ± 0.0085	0.716
Medical	0.5997 ± 0.0768	2	16	0.6263 ± 0.0566	4.436
Yeast	0.4545 ± 0.0145	2	2	0.4677 ± 0.0141	2.904

structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society series B*, 50:157–224, 1988.

[16] X. Li and Y. Guo. Bi-directional representation learning for multi-label classification. In *Machine Learning and Knowledge Discovery in Databases*, volume 8725 of *Lecture Notes in Computer Science*, pages 209–224. Springer Berlin Heidelberg, 2014.

[17] E. Loza Mencía and F. Janssen. Stacking label features for learning multilabel rules. In *Discovery Science - 17th International Conference, DS 2014, Bled, Slovenia, October 8-10, 2014, Proceedings*, volume 8777 of *Lecture Notes in Computer Science*, pages 192–203. Springer, 2014.

[18] G. Madjarov, D. Gjorgjevikj, and S. Dzeroski. Two stage architecture for multi-label learning. *Pattern Recognition*, 45(3):1019–1034, 2012.

[19] E. Mbanya, S. Gerke, C. Hentschel, and P. Ndjiki-nya. Sample selection, category specific features and reasoning. In *Working Notes of CLEF 2011*, 2011.

[20] E. Mbanya, C. Hentschel, S. Gerke, M. Liu, A. Nürnberger, and P. Ndjiki-nya. Augmenting bag-of-words - category specific features and concept reasoning. In *Working Notes of CLEF 2010*, 2010.

[21] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 625–632, New York, NY, USA, 2005. ACM.

[22] S. Nowak, K. Nagel, and J. Liebetrau. The clef 2011 photo annotation and concept-based retrieval tasks. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[23] C. Papagianopoulou, G. Tsoumakas, and I. Tsamardinos. Discovering and exploiting entailment relationships in multi-label learning. *arXiv preprint arXiv:1404.4038 [cs.LG]*, 2014.

[24] S.-H. Park and J. Fürnkranz. Multi-label classification with label constraints. In *ECML PKDD 2008 Workshop on Preference Learning*, 2008.

[25] J. P. Pestian, C. Brew, P. Matykiewicz, D. J. Hovermale, N. Johnson, K. B. Cohen, and W. Duch. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, pages 97–104, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[26] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.

[27] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Güldener, G. Mannhaupt, M. Münsterkötter, and H. W. Mewes. The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*, 32(18):5539–5545, 2004.

[28] A. Srivastava and B. Zane-Ulman. Discovering recurring anomalies in text reports regarding complex space systems. In *Proc. 2005 IEEE Aerospace Conference*, pages 3853 –3862, 2005.

[29] A. Tellegen, D. Watson, and L. A. Clark. On the dimensional and hierarchical structure of affect. *Psychological Science*, 10(4):297–303, 1999.

[30] B. Thomee and A. Popescu. Overview of the imageclef 2012 flickr photo annotation and retrieval task. In P. Forner, J. Karlgren, and C. Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[31] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multilabel classification of music into emotions. In *Proc. 9th International Conference on Music Information Retrieval (ISMIR 2008), Philadelphia, PA, USA, 2008*, 2008.

[32] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research (JMLR)*, 12:2411–2414, July 12 2011.

[33] G. Tsoumakas, M.-L. Zhang, and Z.-H. Zhou. Introduction to the special issue on learning from multi-label data. *Machine Learning*, 88(1-2):1–4, 2012.

[34] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008.

[35] S. Wang, J. Wang, Z. Wang, and Q. Ji. Enhancing multi-label classification by modeling dependencies among labels. *Pattern Recognition*, 47(10):3405 – 3413, 2014.

[36] M.-L. Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 999–1008, New York, NY, USA, 2010. ACM.

[37] Y. Zhang and D.-Y. Yeung. Multilabel relationship learning. *ACM Trans. Knowl. Discov. Data*, 7(2):7:1–7:30, Aug. 2013.