Hierarchical Multi-Label Classification Using Web Reasoning for Large Datasets

Thomas Hassan
Le2i UMR6306, CNRS, Arts et Métiers,
Univ. Bourgogne Franche-Comté
thomas.hassan@u-bourgogne.fr

Christophe Cruz
Le2i UMR6306, CNRS, Arts et Métiers,
Univ. Bourgogne Franche-Comté
christophe.cruz@u-bourgogne.fr

Abstract

Determining valuable data among large volumes of data is one of the main challenges in Big Data. We aim to extract knowledge from these sources using a Hierarchical Multi-Label Classification process called Semantic HMC. This process automatically learns a label hierarchy and classifies items from very large data sources. Five steps compose the Semantic HMC process. This paper focuses on the last two steps where new items are classified according to the label hierarchy. The process is implemented in a scalable and distributed platform to process Big Data. The process is evaluated and compared with multi-label classification algorithms from the state of the art dedicated to the same goal where the Semantic HMC approach outperforms state of the art approaches in some areas.

1. Introduction

The item analysis process requires proper techniques for analysis and representation. In the context of Big Data, this task is even more challenging due to Big Data's characteristics. An increasing number of V's has been used to characterize Big Data [1]: Volume, Velocity, Variety and Value. Volume concerns the large amount of data that is generated and stored through the years by social media, sensor data, etc.[1]. Velocity concerns both the production and the process to meet a demand because Big Data is not only a huge volume of data but it must be processed quickly as new data is generated over time. Variety relates to the various types of data composing the Big Data. These types include semi-structured and unstructured data representing 90% of his content such as audio, video and text. Value measures how valuable the information to a Big Data consumer is. Value is the most important feature of Big Data, because the user expects to make profit out of valuable data. As Big Data analysis can be deemed as the analysis of a special kind of data, many traditional data analysis methods used in Data Mining (algorithms for classification, clustering, regression, among others) may still be utilized for Big Data Analysis [1]. Werner et al. [2] propose a method to semantically enrich an ontology used to describe the domain and classify the news articles. This ontology aims to reduce the gap between the expert's perspective and the classification rules representation. To enrich the ontology and classify the documents they uses an out-of-the-box Description Logics (DL) Web Reasoner. Most of these reasoners are sound and complete to high expressiveness, as OWL2 SROIQ (D) expressiveness, but on the other hand they do not scale: these reasoners cannot handle a large amount of data. Our goal is to extend the work in [2] and to exploit value by analyzing Big Data using a Semantic Hierarchical Multi-Label Classification process (Semantic HMC)[3]. The Semantic HMC is based on an unsupervised ontology learning process using scalable Machine-Learning techniques and Rule-based reasoning. The ontology-described knowledge base (Abox+Tbox) used to represent the knowledge in the classification system is automatically learned from huge volumes of data through highly scalable Machine Learning techniques and Big Data Technologies. Semantic HMC proposes five individually scalable steps to reach the aims of Big Data analytics [4]:

- **Indexation** extracts terms from data items and creates an index of data items.
- **Vectorization** calculates the term-frequency vectors of the indexed items.
- **Hierarchization** creates a label taxonomy (i.e. subsumption hierarchy) from term-frequency vectors.
- **Resolution** creates the reasoning rules to relate data items with the labels based on term-frequency vectors.
- **Realization** first populates the ontology with items and then for each item determines the most specific label and all its subsuming labels.

[3] focuses on the two last steps of the Semantic HMC process. It proposes a new process to hierarchically multi-classify items from huge sets of unstructured texts using DL

ontologies and Rule-based reasoning. This paper is an extention of the work presented in [3] and provides extended experiments with quality evaluation and comparison with some multi-label classification algorithms from the state of the art. The rest of the paper covers five sections. The second section presents background and related work. The third section describes the classification process. The forth section describes the process implementation in a scalable and distributed platform to process Big Data. The fifth section discusses the results. Finally, the last section draws conclusions and suggests further research.

2. Related work

2.1. Ontologies in Classification context

Ontologies are recurrently used in classification systems to describe the classification knowledge (labels, items, classification rules) and to improve the classification process. Galinina et al. [5] used two ontologies to represent a classification system: (1) a Domain ontology that is independent of any classification method and (2) a Method ontology devoted to decision tree classification. Beyond domain description, ontologies can be used to improve the classification process. Elberrichi et al. [6] present a two-steps method for improving classification of medical documents using domain ontologies (MeSH - Medical Subject Headings). Their results prove that document classification in a particular area supported by ontology of its domain increases the classification accuracy.

2.2. Web reasoning in Classification context

Reasoning is used at ontology development or maintenance time as well as at the time ontologies are used for solving application problems [7]. In Classification context, Web reasoning can be used to improve the classification process. In [8] authors presents a document classification method that uses ontology reasoning and similarity measures to classify the documents. In [9] authors introduce a generic, automatic classification method that uses Semantic Web technologies to define the classification requirements, perform the classification and represent the results. The proposed generic classifier is based on an ontology, which gives a description of the entities that need to be discovered. In [2] authors uses out-of-the-box reasoning to classify economical documents but their scalability is limited and cannot be used in large datasets as required in Big Data context.

2.3. Discussion

Most work in the literature focus on describing or improving the classification processes using ontologies but do not take advantage of the reasoning capabilities of web reasoning to automatically multi-classify the items. However as Semantic Web is growing, new high-performance Web Scale Reasoning methods have been proposed [10]. Rule-based reasoning approach allows the parallelization and distribution of work by large clusters of inexpensive machines by programming models for processing and generating large data sets such as Map-reduce[11]. Web Scale Reasoners [10] however, instead of using traditional DL approaches, use entailment rules for reasoning over ontologies. Web-Scale Reasoners based on Map-reduce programming model like WebPie [10] outperforms all other published approaches in an inference test over 100 billion triples [10]. In [12] authors describe a kind of semantic web rule execution mechanism using MapReduce which can be used with OWL-Horst and with SWRL rules. To the extent of our knowledge, a classification process to automatically classify text documents in Big Data context by taking advantage of ontologies and rule-based reasoning to perform the classification is novel.

3. Hierarchichal Multi-Label Classification

In [13], the authors describe in detail the first three steps (Indexation, Vectorization and Hierarchization) of the classification process. Beyond learning the label hierarchy, the process aims to learn a classification model based on a DL ontology presented in [3]. The following subsections describe the last two steps of the process, i.e. how the rules used to classify the items are created and how items are classified using Rule-based Web Reasoning.

3.1. Resolution

The resolution step creates the ontology rules used to relate the labels and the data items, i.e. it establishes the conditions for an $item_i$ to be classified as $label_j$. The rules will define the necessary and sufficient terms of an item $item_i$ be classified as $label_j$. The rules creation process uses thresholds as proposed in [2] to select the necessary and sufficient terms. The main difference with this method is that instead of translating the rules into logical constraints of an ontology captured in Description Logic, these rules are translated in the Semantic Web Rule Language (SWRL). The main interest in using SWRL rules is to reduce the reasoning effort, thus improving the scalability and performance of the system. In the Vectorization step[13], a term co-occurrence frequency matrix $cfm(term_i, term_j)$ is created to represent the co-occurrence of any pair of terms in the collection of items C. Let $P(term_j|term_i)$ be the conditional proportion (number) of the items from collection C common to $term_i$ and $term_j$, in respect to the number of items in $term_j$ such that:

$$P_C(term_i|term_j) = \frac{cfm(term_i, term_j)}{cfm(term_j, term_j)} \tag{1}$$

Two thresholds are defined:

- Alpha threshold (α) such that $\alpha < P_C(term_i|term_j)$, where $term_i \in Label$ and $term_j \in Term$.
- Beta threshold (β) such that $\beta \leq P_C(term_i|term_j) \leq \alpha$, where $term_i \in Label$ and $term_j \in Term$.

These two thresholds are user-defined with a range of $[0, 1]$. Based on these thresholds, two sets of terms are identified (Fig.1):

- Alpha set ($\omega_\alpha^{(term_i)}$) is the set of terms for each label such that:

$$\omega_\alpha^{(term_i)} = \{term_j | \forall term_j \in Term : P_C(term_i|term_j) > \alpha\} \quad (2)$$

  i.e. is the set of terms $term_j$ that co-occur with $term_i \in Label$ with a co-occurrence proportion higher than the threshold α.

- Beta set ($\omega_\beta^{(term_i)}$) is the set of terms for each label such that:

$$\omega_\beta^{(term_i)} = \{term_j | \forall term_j \in Term : \beta \leq P_C(term_i|term_j) \leq \alpha\} \quad (3)$$

  i.e. is the set of terms that co-occur with $term_i \in Label$ with a co-occurrence proportion higher or equal than the threshold β and lower than the threshold α.

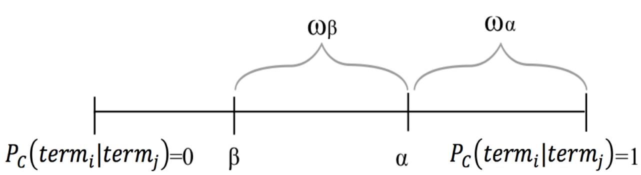

Figure 1. Alpha and Beta sets

.

If the item has at least one term in $\omega_\alpha^{(term_i)}$ it is classified with $term_i$, $term_i \in Label$. For each term that complies with the above rule, a SWRL rule is created.

If the item has at least δ terms in $\omega_\beta^{(term_i)}$, it is classified with $term_i$, $term_i \in Label$. One SWRL rule is generated for each combination of $term_j \in \omega_\beta^{(term_i)}$ where the number of combined terms is at least $\delta = \lceil |\omega_\beta^{(term_i)}| * p \rceil$, and $0 \leq p \leq 0.5$. The set of generated beta rules is the combination C_n^m of m terms of a larger set of n elements. Regarding our approach, n is the number of possible terms $|\omega_\beta^{(term_i)}|$, and m the minimum number of terms δ in each rule (e.g. $C_{20}^{10} = 184756$). In order to limit the number of rules for each label we fix the value of $n \leq 10$. The terms are selected by ranking the terms in $\omega_\beta^{(term_i)}$ using the conditional proportion $P_C(term_i|term_j)$ as the ranking score. Example Alpha and Beta SWRL rules are depicted in Table 1.

Notice that the rules that encompass more than δ terms are not necessary because the combination of any δ terms is sufficient to classify the item.

In non-empty alpha and beta category, beta and alpha rules are both considered. Alpha rules are evaluated as presented in the empty beta category. Beta rules are evaluated as presented in the empty alpha category but with a value $q = p * 2$ because beta rules are, by definition, less relevant than alpha rules. It corresponds to $\delta = \lceil |\omega_\beta^{(term_i)}| * q \rceil$, with $0 \leq q \leq 1$ and $q = p * 2$.

3.2. Realization

The realization step includes two sub-steps: population and classification. The ontology-described knowledge base is populated with new items and their relevant terms at the assertion level (Abox). Each item is described with a set of relevant terms $\omega_\gamma^{(item_i)}$ such that:

$$\omega_\gamma^{(item_i)} = \{term_j | \forall term_j \in Term \wedge \gamma < tfidf_{(item_i, term_j, C)}\} \quad (4)$$

where γ is the relevance threshold, $\gamma < tfidf_{(item_i, term_j, C)}$, $term_j \in Term$, $item_i \in Item$ and $tfidf$ as calculated in the Vectorization step.

The classification sub-step performs the multi-label hierarchical classification of the items. Rule-based reasoning applies exhaustively a set of rules to a set of triples (i.e. the data items) to infer conclusions [14], i.e. the item's classifications.

The rule-based inference engine uses rules to infer the subsumption hierarchy (i.e. concept expression subsumption) of the ontology and the most specific concepts for each data item. This leads to a multi-label classification of the items based in a hierarchical structure of the labels (Hierarchical Multi-label Classification).

4. Implementation

The process is implemented as a combination of available Java libraries that natively support parts of the process. In the first three steps (indexation, vectorization and hierarchization) of the Semantic HMC process, Big Data technologies are used, including MapReduce [11]. The MapReduce algorithms are deployed on a Hadoop cluster https://hadoop.apache.org/. The next subsections describe the implementation details of each step of the classification process.

4.1. Resolution

The resolution process creates the ontology rules used to relate the labels and the data items. The rule creation process is divided in a sub-process for each $label_i \in Label$. In

Table 1. Generated Rules Examples

Alpha rules
$Item(?it), Term(?t_1), Label(?t_1), hasTerm(?it, ?t_1) \rightarrow isClassified(?it, ?t_1)$
Beta rule
$Item(?it), Term(?t_1), Term(?t_2), Label(?t_3), hasTerm(?it, ?t_1), hasTerm(?it, ?t_2) \rightarrow isClassified(?it, ?t_3)$

each sub-process $\omega_\alpha^{(label_i)}$ and $\omega_\beta^{(label_i)}$ sets are calculated using the co-occurrence matrix, then classification rules are created for each label. Exploiting a huge co-occurrence matrix to create the ontology rules is a very intensive task, thus this process is also distributed to several machines in the MapReduce paradigm. A MapReduce job creates the rules from the co-occurrence matrix. Following rule generation in MapReduce, the rules are serialized in SWRL language and stored in the ontology-described knowledge base using the OWL-API library. The generated rules along with the label hierarchy are used in the Realization process to classify new items.

4.2. Realization

The realization step populates the ontology and performs the multi-label hierarchical classification of the items. First the ontology is populated with new items and the most relevant terms to describe each document in an assertion level (Abox). To store, manage and query the ontology-described knowledge base (Tbox+Abox) a triple-store is used. Because highly expressive forward chaining description logics reasonners do not scale well and, in our preliminary prototype we decided to adopt the classification at query time approach by using a triple-store with a backward-chaining inference engine. The OWL-API library is used to populate the OWL ontology with new items. A scalable triple-store called Stardog (http://docs.stardog.com) is used to store and query the ontology-described Knowledge Base (Tbox+Abox). Stardog is also used to perform reasoning by backward-chaining inference as well as SWRL rules inference. The rule selector was developed in java, and interacts with Stardog to optimize the query performance.

5. Experiments

In this section, a quality evaluation of the Semantic HMC is depicted. This evaluation focuses on classification accuracy, and complements the performance evaluation depicted in [3]. Finally we discuss the obtained results regarding some algorithms from the state of the art in Hierarchical Multi-Label Classification.

5.1. Quality Evaluation

In this subsection we evaluate the classification performance of the Semantic HMC process for unstructured text classification in a Big Data context. First the dataset, the test environment and the experimental settings used to evaluate the process are described. Then the experimental results are presented and discussed. The evaluation is done using a pre-labeled dataset, composed of training and test data. The training set is used to learn hierarchical relations between the pre-defined labels and classification rules. The test set is used to calculate the classification performance of the algorithm based on standard quality measures. To be able to compare our approach with state-of-the-art, we use a pre-defined set of labels instead of automatically learned labels as it is described in [3].

5.1.1 Delicious dataset

The Delicious dataset[1] is used to perform this evaluation. This dataset is composed of labeled textual data from web pages extracted from the Delicious social bookmarking website[15]. Table 2 shows the dataset specifications. The Delicious dataset was chosed because contains very few features (words) compared to the number of labels, rendering accurate classification difficult [16]. Also, it has been used to evaluate several multi-label classification systems, thus it provides a good baseline to compare our approach.

Table 2. Delicious dataset specifications

$\lvert Train \rvert$	$\lvert Test \rvert$	$\lvert Labels \rvert$	$\lvert Terms \rvert$
12,910	3,181	983	500

5.1.2 Measures

Specific evaluation metrics to multi-label learning are proposed in literature and generally categorized into two groups : example-based metrics and label-based metrics[17]. We use a label-based metric to evaluate the Semantic HMC. In label-based metrics the micro-averaged as well as the macro-averaged precision and recall are used : the learning system's performance is evaluated on each class label separately, and then the macro/micro-averaged value across all class labels is returned. These measures are calculated as in [18].

5.1.3 Results

The Hierarchization phase of the Semantic HMC process automatically generates a hierarchical relations between labels. This hierarchy, along with the classification rules cre-

[1] http://mulan.sourceforge.net/datasets-mlc.html

ated in the Resolution step are used to perform hierarchical multi-label classification. Figure 2 shows a sample of the hierarchichal relations ($skos : hasBroaderRelation$) between labels automatically created for the Delicious dataset. The set of parameters used to create the hierarchy and classification rules is described in table 3. This parameters can have a high impact in the quality of the results. The Top and Bottom Tresholds are used to calculate the hierarchichal relations between labels as defined in [13].

Table 3. Execution Settings for Delicious Dataset

Parameter	Step	Value
Top Threshold	Hierarchization	50
Bottom Threshold	Hierarchization	40
Alpha Threshold	Resolution	20
Beta Threshold	Resolution	10
Term ranking (n)	Resolution	5
p	Resolution	0.25
Term Threshold (γ)	Realization	2

Table 4 shows the results obtained by the Semantic HMC process on the Delicious dataset.

Table 4. Quality results for the Delicious Dataset

	Precision	Recall	F1-measure
Micro	0.284	0.74	0.410
Macro	0.0676	0.178	0.0979

5.2. Comparison with the state of the art

Table 5 shows the Macro-F1 measure and Micro-F1 measure obtained on the Delicious dataset. The results of the proposed process (SHMC) are compared with several state-of-the-art approaches results with the same dataset[18][19][16]. In Table 5, it is observed that the Semantic HMC approach outperforms state-of-the-art approaches in micro F1-measure, while the macro F1-measure is comparable to most other approaches. These results show that the classification performance of our ontology-based approach is comparable to the performance of the selected algorithms from the state-of-the-art in machine learning.

Table 5. Performance of various algorithms on the Delicious dataset

Algorithm	Macro F1	Micro F1
SHMC	0.0979	**0.410**
CGS_p	0.10378	0.29740
TNBCC	0.0880	N/A
Path-BCC	0.084	N/A
BR	0.096	0.234
CC	0.100	0.236
HOMER	0.103	0.339
ML-kNN	0.051	0.175
RFML-C4.5	**0.142**	0.269
RF-PCT	0.083	0.248

6. Conclusions

This paper describes an unsupervised hierarchical multi-label classification process from unstructured text in the scope of Big Data. Following the performance evaluation depicted in [3], a quality evaluation is depicted, comparing the classification accuracy of the Semantic HMC with several approaches from the state-of-the-art. The experiment shows that the classification performance of the Semantic HMC process that uses ontologies and rule-based reasoning to classify unstructured text documents is comparable to the performance of algorithms from the state-of-the-art in machine learning field. Also, unlike most approaches from the data-mining field, the ontology-based approach provides human-readable explanations of the classifications, that can be used to monitor the classification process by experts. Our current work is twofold: (1) the application of the process to domain-specific data and (2) the maintenance of the classification model regarding a stream of data in a Big Data Context.

References

[1] "Big data: A survey", *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, Jan. 2014, ISSN: 1383469X. DOI: 10.1007/s11036-013-0489-0 (cit. on p. 1).

[2] D. Werner, N. Silva, C. Cruz, and A. Bertaux, "Using DL-reasoner for hierarchical multilabel classification applied to economical e-news", in *Proceedings of 2014 Science and Information Conference, SAI 2014*, 2014, pp. 313–320, ISBN: 9780989319317. DOI: 10.1109/SAI.2014.6918205 (cit. on pp. 1, 2).

[3] R. Peixoto, T. Hassan, C. Cruz, A. Bertaux, and N. Silva, "An unsupervised classification process for large datasets using web reasoning", in *SBD'16: Semantic Big Data Proceedings*, ACM, Ed., San Francisco (CA), USA, 2016 (cit. on pp. 1, 2, 4, 5).

[4] T. Hassan, R. Peixoto, C. Cruz, A. Bertaux, and N. Silva, "Semantic HMC for big data analysis", in *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, 2015, pp. 26–28, ISBN: 9781479956654. DOI: 10.1109/BigData.2014.7004482. arXiv: 1412.0854 (cit. on p. 1).

[5] A. Galinina and A. Borisov, "Knowledge modelling for ontology-based multiattribute classification system", *Applied Information and Communication ...*, pp. 103–109, 2013 (cit. on p. 2).

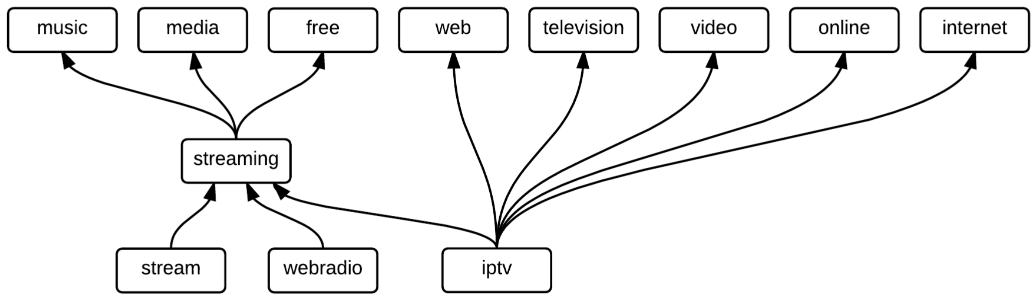

Figure 2. Automatically generated hierarchy from Delicious dataset (sample)

[6] Z. Elberrichi, B. Amel, and T. Malika, "Medical Documents Classification Based on the Domain Ontology MeSH", *ArXiv preprint arXiv:1207.0446*, 2012 (cit. on p. 2).

[7] R. Moller and V. Haarslev, *Tableau-Based Reasoning*, 2009. DOI: 10.1007/978-3-540-92673-3{_}23 (cit. on p. 2).

[8] J. Fang, L. Guo, and Y. Niu, "Documents classification by using ontology reasoning and similarity measure", in *Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010*, vol. 4, 2010, pp. 1535–1539, ISBN: 9781424459346. DOI: 10.1109/FSKD.2010.5569338 (cit. on p. 2).

[9] D. Ben-David, T. Domany, and A. Tarem, "Enterprise data classification using semantic web technologies", in *The Semantic Web-ISWC 2010*, ser. ISWC'10, Berlin, Heidelberg: Springer-Verlag, 2010, pp. 66–81, ISBN: 3-642-17748-4, 978-3-642-17748-4 (cit. on p. 2).

[10] J. Urbani, "Three Laws Learned from Web-scale Reasoning", in *2013 AAAI Fall Symposium Series*, 2013, pp. 76–79 (cit. on p. 2).

[11] J. Dean and S. Ghemawat, "MapReduce : Simplified Data Processing on Large Clusters", *Communications of the ACM*, SIGMOD '07, vol. 51, no. 1, L. P. Daniel, Ed., pp. 1–13, 2008, ISSN: 00010782. DOI: 10.1145/1327452.1327492. arXiv: 10.1.1.163.5292 (cit. on pp. 2, 3).

[12] H. Wu, J. Liu, D. Ye, H. Zhong, and J. Wei, "A distributed rule execution mechanism based on MapReduce in sematic web reasoning", *Proceedings of the 5th Asia-Pacific Symposium on Internetware - Internetware '13*, pp. 1–7, 2013. DOI: 10.1145/2532443.2532457 (cit. on p. 2).

[13] R. Peixoto, T. Hassan, C. Cruz, A. Bertaux, and N. Silva, "Semantic HMC: A Predictive Model using Multi-Label Classification For Big Data", in *The 9th IEEE International Conference on Big Data Science and Engineering (IEEE BigDataSE-15)*, 2015, ISBN: 978-1-4673-7952-6. DOI: 10.1109/Trustcom.2015.578 (cit. on pp. 2, 5).

[14] J. Urbani, F. Van Harmelen, S. Schlobach, and H. Bal, "QueryPIE: Backward reasoning for OWL horst over very large knowledge bases", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, ser. ISWC'11, vol. 7031 LNCS, Berlin, Heidelberg: Springer-Verlag, 2011, pp. 730–745, ISBN: 9783642250729. DOI: 10.1007/978-3-642-25073-6{_}46 (cit. on p. 3).

[15] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and efficient multilabel classification in domains with large number of labels", in *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*, 2008, pp. 30–44 (cit. on p. 4).

[16] Y. Papanikolaou, T. N. Rubin, and G. Tsoumakas, "Improving gibbs sampling predictions on unseen data for latent dirichlet allocation", *ArXiv preprint arXiv:1505.02065*, 2015 (cit. on pp. 4, 5).

[17] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms", *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014, ISSN: 10414347. DOI: 10.1109/TKDE.2013.39 (cit. on p. 4).

[18] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning", *Pattern Recognition*, vol. 45, no. 9, pp. 3084–3104, 2012 (cit. on pp. 4, 5).

[19] L. E. Sucar, C. Bielza, E. F. Morales, P. Hernandez-Leal, J. H. Zaragoza, and P. Larranaga, "Multi-label classification with bayesian network-based chain classifiers", *Pattern Recognition Letters*, vol. 41, pp. 14–22, 2014, Supervised and Unsupervised Classification Techniques and their Applications, ISSN: 0167-8655 (cit. on p. 5).