Institute of Mathematics and Computational Sciences

ISSN - 0103-2569

A systematic review on experimental multi-label learning

Newton Spolaôr
Everton Alvares Cherman
Jean Metz
Maria Carolina Monard

N^{o} 392

ICMC TECHNICAL REPORT

São Carlos, SP, Brazil
February/2013

A systematic review on experimental multi-label learning[*] [†]

Newton Spolaôr[1]
Everton Alvares Cherman[1]
Jean Metz[2]
Maria Carolina Monard[1]

[1]Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação
Laboratório de Inteligência Computacional
CEP 13560-970 - São Carlos, SP, Brasil

[2]Universidade Tecnológica Federal do Paraná
CEP 85884-000 - Medianeira, PR, Brasil

February, 2013

[*] This work was supported by the São Paulo Research Council - FAPESP.

[†] Available on-line at: http://www.icmc.usp.br/~biblio/relatorios_tecnicos.php.

Abstract

Multi-label learning deals with the classification problem where each example is associated with a set of labels, which are usually dependent. This research topic has emerged in recent years due to the increasing number of applications where examples are annotated with more than one label. However, there is a lack of reviews focusing on pieces of work which report experimental results for multi-label learning. To this end, the systematic review process can be useful to identify related publications in a wide, rigorous and replicable way. This work uses the systematic review process to answer the following research question: *what are the publications which report experimental results for multi-label learning research?* The systematic review process carried out in this work included the application of 16 selection criteria to narrow the literature review, as we are interested in papers which report specific classifier evaluation measures using datasets publicly available. Moreover, these datasets cannot be preprocessed. In the end, this process enabled us to select 64 relevant publications, as well as identify some interesting facts in the current literature.

Keywords: Systematic Review, Machine Learning, Empirical Research, Supervised Learning

This document was written with the LaTeX text editor. The system of citation of bibliographical references uses the *Apalike* style of the bibTeX system.

Contents

List of Figures

List of Tables

1 Introduction

A research process can be specified as a sequence of activities to collect information about a subject and analyse the information to obtain knowledge. Bibliographical research, part of the research process, can be performed in a distinct way through the systematic literature review process (Kitchenham, 2007), more usually known as Systematic Review (SR).

The SR process enables us to answer Research Questions (RQ) about a subject using previously specified activities to identify, select, evaluate and synthesize publications. To this end, it explores the literature searching for relevant pieces of work in a fair, rigorous and replicable way. Alternatively, a meta-analysis[1] can evaluate the selected pieces of work. However, meta-analysis seems to be unusual in Computer Science, since the reporting protocols often vary in this field (Brereton et al., 2007).

In Computer Science, there have been several applications of the SR process in subjects related to Software Engineering (Zhang and Babar, 2012; Guessi et al., 2011; Kitchenham et al., 2010). Recently, there were some applications of this process in other areas, such as Artificial Intelligence (Spolaôr et al., 2012) and Educational Robotics (Benitti, 2012).

Machine learning, which has significant overlapping with data mining, pattern recognition and parts of statistics, is an important field of Artificial Intelligence. Machine learning deals with the fundamental problem of using a dataset to reproduce the process that generated the data.

Multi-label learning deals with the classification problem where each example (or instance) in the training dataset is associated with a set of labels, *i.e.*, each example can belong to multiple different classes simultaneously. Multi-label learning is an emerging research topic due to the increasing number of applications where examples are annotated with more than one label. Multi-label classification has been used in applications such as bioinformatics, emotion analysis, semantic annotation of media and text categorization (Tsoumakas et al., 2010).

The task of a multi-label classifier is to predict the label set of unseen examples. Thus, multi-label learning is more general than single-label learning, in which each example in the training dataset is associated with only one class, which can assume several values. Whenever there are more than two class values in single-label learning, it is called *multi-class classification*. Case the class value is Yes/No, it is called *binary classification*. In fact, the main difference between multi-label and single-label learning is that classes in multi-label learning are often correlated while the class values in single-label learning are mutually exclusive.

[1] Type of study that synthesizes the results of the review through statistical techniques.

Machine learning research relies to a large extent on experimental observations. Whenever a new learning algorithm is proposed, its performance is compared to existing algorithms. To this end, it is usual to execute the algorithms on several selected datasets from different domains, and the quality of the resulting classifiers are evaluated using appropriate evaluation measures. The final step consists of statistically verifying the hypothesis of improved performance of the new algorithm (Demsar, 2006).

Despite the emergence of multi-label learning research, there are few extensive reviews surveying publications on this topic (Tsoumakas et al., 2010; Carvalho and Freitas, 2009). Moreover, to the best of our knowledge, there is no extensive review focused on papers reporting experimental results for multi-label learning. Thus, this work contributes to reducing this gap by using the systematic review process without meta-analysis.

This work is organized as follows: Section 2 briefly describes the SR process. Section 3 describes the application of this method to identify publications which report experimental results in multi-labeled data, and Section 4 presents the final conclusions.

2 The Systematic Review Process

Some decades ago, the systematic review process emerged in areas such as Medicine. Its popularity has recently increased (Zhang and Babar, 2012; Castro et al., 2002) and guidelines have been proposed in Medicine (Higgins and Green, 2009), Social Science (Petticrew and Roberts, 2006) and Computer Science (Kitchenham, 2007; Biolchini et al., 2005).

The systematic review process has recently been applied in Computer Science. Several of these applications are related to Software Engineering, which has specific defined guidelines, as well as a systematic review about systematic reviews (Kitchenham et al., 2010). Other examples can be found in Artificial Intelligence (Spolaôr et al., 2010), Human-computer Interaction (Madeo and Peres, 2012) and Intrusion Detection (Pisani and Lorena, 2010). Moreover, we have carried out a previous SR to search for publications related to feature selection in multi-labeled data (Spolaôr et al., 2012).

The systematic review process consists of three steps (Kitchenham, 2007).

Step 1. Planning;

Step 2. Conducting;

Step 3. Reporting.

Step 1 involves specifying the research questions that must be answered and creating a protocol. The activities that integrate this protocol are carried out in Step 2, enabling one to identify a set of publications related to the researched subject. The last step is responsible for reporting the results obtained. These results are usually reported in PhD theses, technical reports, articles or other formats.

Each step consists of several activities described next, which can be executed concomitantly, which could improve themselves.

Step 1. Planning:

- Identification of the need for a review;
- Commissioning a review (optional);
- Specifying the research questions;
- Developing a review protocol;
- Evaluating the review protocol (optional).

Step 2. Conducting:

- Identification of research;
- Selection of publications;
- Study of quality assessment;
- Information extraction;
- Information synthesis.

Step 3. Reporting:

- Specifying the dissemination mechanisms;
- Formatting the main report;
- Evaluating the report (optional).

Specifying the research questions in Step 1 is one of the main activities carried out in the systematic review process, as it guides the development of the criteria contained in the protocol, the scope of the bibliographical review and the activities to be carried out in the other steps. At the end of the SR process, these research questions should be answered, highlighting their importance.

After formulating the research questions, it is possible to develop a review protocol to minimize potential bias during the application of the systematic review process (Kitchenham, 2007). A protocol basically consists of the background on the subject studied and the description of the strategies which are used in Step 2 of the SR process. An advantage of having a protocol is to support the replication of the SR process.

Identifying the publications of the research on a subject is an important activity carried out in Step 2. It is done by using a search strategy, which can be developed based on earlier systematic reviews and previous tests. This strategy requires a Search String (SS) and the resources to be researched[2].

An approach to specify a SS is based on the structure of the research questions (Spolaôr et al., 2012). An alternative approach is based on keywords and their synonyms identified in the RQ. First, these words are registered, such that each keyword and its synonyms belong to a specific group. Afterwards, the words in each group are combined using the boolean operator OR. Finally, all groups are combined into a single search string using the boolean operator AND.

The research question used next exemplifies the alternative approach, which provides support to the generation of the three groups described in Table 1.

RQ What topics (subjects) are taught through robotics in schools? (Benitti, 2012)

teaching	robotic	school
learning	robotics	k-12
teach	robot	
learn	robots	
education	Lego	
educational		

Table 1: Lists of keywords and synonyms related to a research question.

The use of this approach allows us to create the following search string.

SS ((teaching OR learning OR teach OR learn OR education OR educational) AND (robotic OR robotics OR robot OR robots OR Lego) AND (school OR “k-12”))

The selection of publications is another important activity of the systematic review process. It can be performed with the support of inclusion/exclusion criteria and only publications that can answer the research questions should be kept.

An example of exclusion criteria is the deletion of publications that were published before a specific year. Nevertheless, it is interesting to adopt a conservative posture during this activity, as the careless exclusion of a relevant publication implies in a loss of information, which may affect the quality of

[2] Digital libraries, bibliographical databases, specific journals and conference proceedings.

the SR process. To verify if a publication suits the selection criteria, the title, the abstract and/or the other parts of the publication should be read. A publication with a well structured abstract would simplify this activity.

The quality assessment of publications is performed with the use of the quality criteria, usually by checklists, which can be based on some models found in the literature (Fink, 2004). In this context, "quality" means the methodological merit of a study. These criteria contribute, for example, to correlating differences among the results in different publications and among the quality of these results. They also might suggest future trends on the subject of the systematic review process. An example of quality criteria which can be verified is the use of statistical tests.

Checklists provide support to the use of two approaches (Kitchenham, 2007):

1. Specifying more detailed selection criteria;

2. Supporting analysis and synthesis of the information obtained from pieces of work that can answer the research questions.

The first approach requires a separated form to extract information from the new selected publications, while the second one allows us to use a unique form.

The synthesis activity in Step 2, which can be quantitative or qualitative, is useful to summarize and organize the information extracted from the publications. The quantitative synthesis enables researchers to carry out meta-analysis, which is still unusual in Computer Science. Moreover, it considers numerical information, such as size of samples, accuracy and standard deviation, which can highlight differences among publications. On the other hand, qualitative synthesis, based on approaches such as the one suggested by Noblit and Hare (1988), allows researchers to identify similarities among publications.

Other examples of research questions and criteria, as well as a wider introduction to the systematic review process, are described in (Spolaôr et al., 2012).

In the following section, the use of the SR process to search for publications which report experimental results in multi-label learning is described.

3 Systematic Review Application

The SR process described in this work focuses on the identification of publications which report experimental results for multi-label learning research.

This process was carried out by the four authors of this work during a five month period (August - December of 2012) at the Institute of Mathematics and Computer Science, University of São Paulo, and the Aristotle University of Thessaloniki[3]. In what follows, the three steps of the systematic review process are described.

3.1 Planning

As already mentioned, there are few extensive reviews surveying multi-label learning publications. It should be noted that previous SR on publications which report experimental results in this topic have not been found. However, there is a SR about feature selection to support multi-label learning (Spolaôr et al., 2012), from which the current protocol, described next, was based on.

The following research question was defined by us.

RQ *What are the publications which report experimental results for multi-label learning research?*

Based on the background described in (Cherman et al., 2012; Metz et al., 2012), the following groups of keywords and synonyms were considered to specify the search string. More details about them are described in Appendix A.1.

- set of labels: keywords related to the type of dataset from which the results are gathered.
- multi-label learning: keywords related to the research area.
- experimental: keywords related to the sources of the results.

The online bibliographical databases selected to find the publications were: ACM Portal[4], CiteSeer^{X}[5], IEEE Xplore[6], ScienceDirect[7], Scirus[8], Scopus[9], SpringerLink[10], Wiley Interscience[11] and Web of Science[12]. In some cases, the search string was adapted to suit database limitations, such as the maximum number of topics.

The adaptations include decomposition of the search string into smaller ones and the posterior union of the results. Furthermore, the scope of the

[3] http://www.auth.gr/en
[4] http://portal.acm.org
[5] http://citeseerx.ist.psu.edu
[6] http://ieeexplore.ieee.org
[7] http://www.sciencedirect.com
[8] http://scirus.com
[9] http://www.scopus.com
[10] http://link.springer.com/
[11] http://onlinelibrary.wiley.com
[12] http://isiknowledge.com

search string was limited to the title, abstract and keywords, whenever the source supported this requirement.

Some retrieved pieces of work can be duplicated, as some sources, *i.e.*, journals, proceedings and others, are indexed by more than one bibliographical database. Thus, cases with duplicated title were automatically or manually removed, keeping only one copy of the publication.

We then divided the remaining pieces of work among ourselves, such that each one of them was reviewed by one person. Whenever a piece of work fulfilled one or more exclusion criteria, it was removed. If there were doubts about removing a publication, a second reviewer verified the doubtful papers.

In what follows, we specified the 16 Selection Criteria (SC) used in this work. It should be noted that all of them are exclusion criteria.

SC 1 Publications that do not suit the RQ;

SC 2 Duplicated publications by the same authors, *i.e.*, similar title, abstract, results or text. In this case, only one is kept;

SC 3 Publications that also perform label selection;

SC 4 Publications that do not address explicitly multi-labeled data;

SC 5 Tutorial slides;

SC 6 Publications composed of only one page (abstract papers), posters, presentations, proceedings and program of scientific events;

SC 7 Publications hosted in web pages which are not accessed through the account of the University of São Paulo or the Aristotle University of Thessaloniki;

SC 8 Publications written in a language different than English;

SC 9 Publications that do not evaluate any multi-label algorithm;

SC 10 Publications that do not consider any multi-label publicly available dataset in attribute-value format;

SC 11 Publications that do not use any example-based or label-based multi-label evaluation measures;

SC 12 Publications that do not consider supervised learning algorithms;

SC 13 Publications that do not tabulate any experimental results from multi-label algorithms;

SC 14 Publications that do not perform flat multi-label learning;

SC 15 Publications that only perform active multi-label learning;

SC 16 Publications that perform data preprocessing in every dataset.

After selecting the publications, we applied the four Quality Criteria (QC) described next according to the second approach, as explained at the end of Section 2.

QC 1 Does the publication compare multi-label learning algorithms?

QC 2 Does the publication use more than one dataset?

QC 3 Does the publication use more than one evaluation measure?

QC 4 Does the publication report standard deviation from any evaluation measures?

It should be emphasised that the quality criteria are only applied in the selected papers, *i.e.*, the ones that do not fulfill any exclusion criteria. For example, if a paper uses two datasets, but only one is publicly available in attribute-value format (SC 10), it will not fulfill the QC 2 in this work.

The form subsequently constructed with the information extracted from the selected papers consists of a LibreOffice[13] electronic spreadsheet with 42 columns, which are described in Appendix A.2. As most of the information extraction has to be carried out manually, this process was double checked. This tool enables us to easily verify the quality criteria. For example, counting the data sets used in each publication can be done using the standard spreadsheet functions.

A relational database was built to appropriately record the 42 columns, modelling each sheet as a database table. In this database, each sheet column is a table attribute and each sheet line is an instance. Figure 1 shows the corresponding entity-relationship model, which is composed of four tables: *main*, *dataset*, *measure* and *paper*, as well as some relationships between them.

The *main* table records the experimental settings and results published in the papers which are able to answer the research question, and some foreign keys which link results to a paper and a dataset. Furthermore, the *dataset* table also records usual statistics from multi-label datasets, such as label cardinality and label density, and the *measure* table manages the name and type of the recorded multi-label evaluation measures, enabling us to register measures recently proposed. The *paper* table records the selected publications.

We performed a short qualitative synthesis of the information described in the form, as is the case in most of the systematic reviews carried out in the Software Engineering area (Brereton et al., 2007).

[13] http://www.libreoffice.org

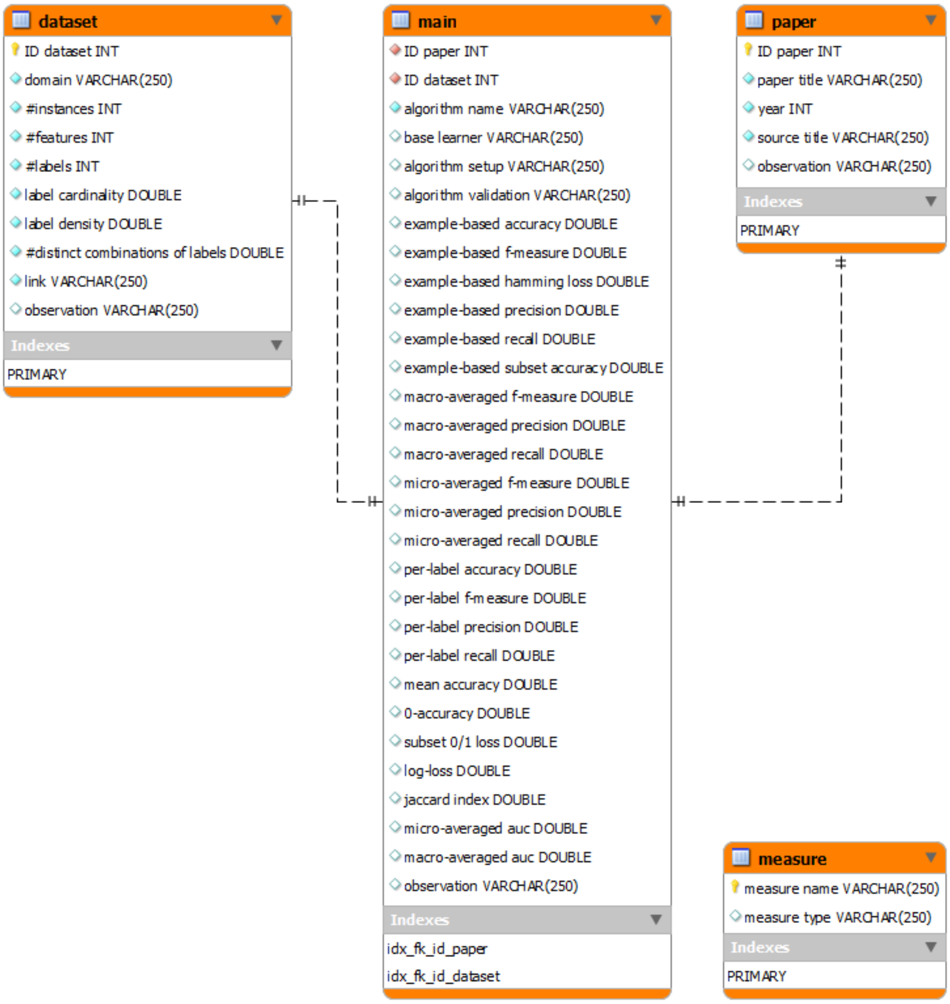
dataset
ID dataset INT
domain VARCHAR(250)
#instances INT
#features INT
#labels INT
label cardinality DOUBLE
label density DOUBLE
#distinct combinations of labels DOUBLE
link VARCHAR(250)
observation VARCHAR(250)
Indexes
PRIMARY
main
ID paper INT
ID dataset INT
algorithm name VARCHAR(250)
base learner VARCHAR(250)
algorithm setup VARCHAR(250)
algorithm validation VARCHAR(250)
example-based accuracy DOUBLE
example-based f-measure DOUBLE
example-based hamming loss DOUBLE
example-based precision DOUBLE
example-based recall DOUBLE
example-based subset accuracy DOUBLE
macro-averaged f-measure DOUBLE
macro-averaged precision DOUBLE
macro-averaged recall DOUBLE
micro-averaged f-measure DOUBLE
micro-averaged precision DOUBLE
micro-averaged recall DOUBLE
per-label accuracy DOUBLE
per-label f-measure DOUBLE
per-label precision DOUBLE
per-label recall DOUBLE
mean accuracy DOUBLE
0-accuracy DOUBLE
subset 0/1 loss DOUBLE
log-loss DOUBLE
jaccard index DOUBLE
micro-averaged auc DOUBLE
macro-averaged auc DOUBLE
observation VARCHAR(250)
Indexes
idx_fk_id_paper
idx_fk_id_dataset
paper
ID paper INT
paper title VARCHAR(250)
year INT
source title VARCHAR(250)
observation VARCHAR(250)
Indexes
PRIMARY
measure
measure name VARCHAR(250)
measure type VARCHAR(250)
Indexes
PRIMARY

Figure 1: Entity-relationship diagram modelling the information extracted from the selected papers.

3.2 Conducting

Using the search string in the bibliographical databases selected allowed us to identify 1543 publications. As mentioned before, in many cases the same publication can be retrieved by more than one database. Therefore, automatic[14] and manual removal of publications with the same title was carried out. Afterwards, the exclusion criteria described in Section 3.1 were applied, resulting in a dramatic reduction of the number of publications. Figure 2 summarizes the results of the procedures carried out to select the final 64 papers.

First, most of the identified publications (55%) were automatically removed, as they have duplicated titles. We then manually reviewed this procedure, as some publications have mistyped titles in the databases, eliminating more publications (5%). Finally, we carefully removed publications fulfilling one or

[14] A simple computational framework was implemented to remove publications with identical titles.

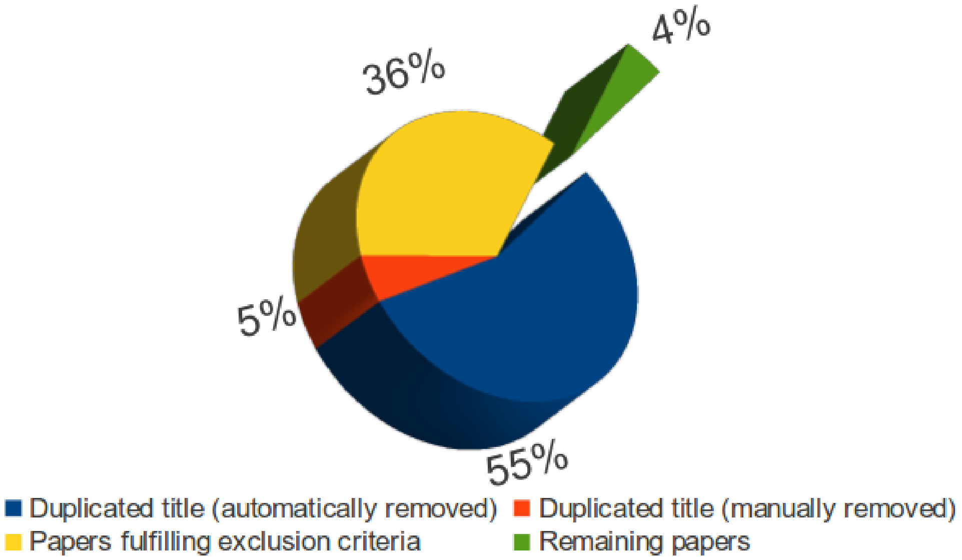

Figure 2: Summary of the procedures to select relevant papers from a total of 1543 publications.

more exclusion criteria (36%)[15]. To carry out this procedure, the abstract and, eventually, the whole publication had to be read. As Figure 2 shows, only 4% of the papers identified were kept.

The next activity performed was the quality assessment of the 64 final publications. This activity was carried out using the information extracted from these papers, already structured in the electronic spreadsheet. The next section describes the synthesis activity based on the quality criteria from all the 64 publications, as well as the way we are using to reporting systematic review results.

3.3 Reporting

The following dissemination mechanisms were selected to report the results:

1. Submitting a paper to a conference. This paper uses the information extracted in the current report for comparison against the multi-label baseline classifier proposed by Metz et al. (2012).

2. Hosting the report in a web site at the University of São Paulo[16];

3. Disseminating the site to the community, with special attention to our research collaborators from UFABC[17], UNIOESTE[18] and Aristotle University

[15] Nine papers have potential to fulfill exclusion criteria, although we could not confirm it. These papers were kept separated from the 64 selected publications.

[16] http://www.icmc.usp.br/~biblio/relatorios_tecnicos.php

[17] http://dgp.cnpq.br/buscaoperacional/detalhegrupo.jsp?grupo=IWU4103700AHR2

[18] http://www.foz.unioeste.br/labi

of Thessaloniki[19]. The electronic spreadsheet form with the information extracted from the selected publications, as well as additional information about the selected publications, can be obtained by request from the authors of this systematic review.

Before synthesizing the selected papers according to the quality criteria, some graphical summaries of the 64 selected publications are presented. Machine learning is empirical (Dietterich, 1990). To this end, the community usually carries out experimental studies to evaluate the performance of learning algorithms (Langley, 2000). Datasets are useful in these studies, motivating us to verify how often benchmark multi-label datasets are used in the selected papers. Figure 3 shows the number of papers using each multi-label dataset, highlighting the high frequency of the yeast dataset (51 papers, or 80% of the 64 final papers).

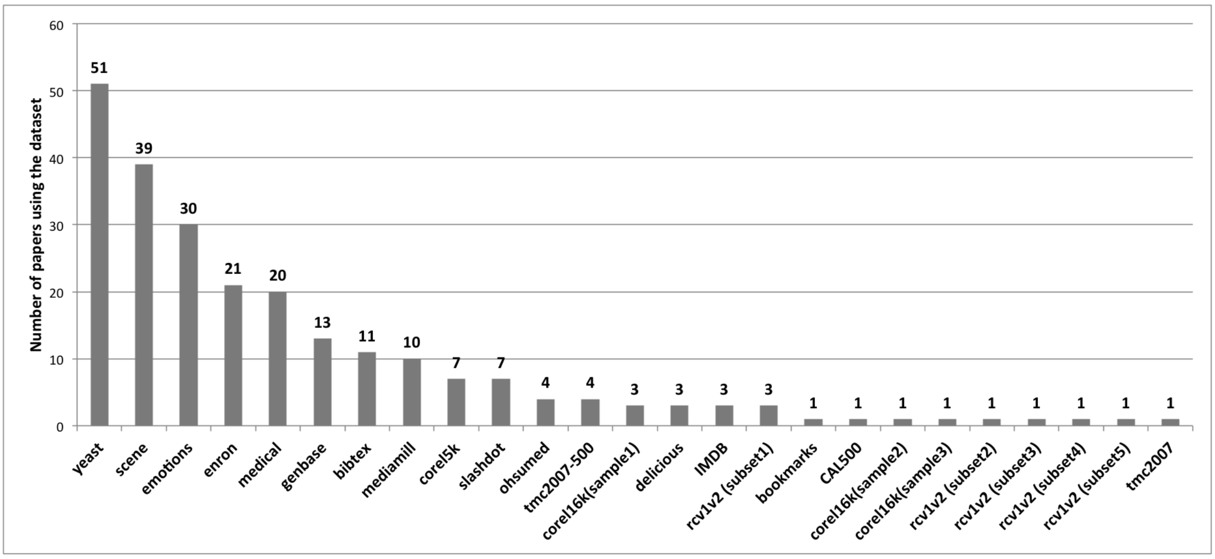

Figure 3: Number of papers using each multi-label dataset (total: 64 papers).

As already mentioned, we do not consider experimental results from preprocessed datasets (SC 16, Page 8). This constraint left out a great number of publications which report experimental results for multi-label learning research, such as text categorization, which is a typical multi-label problem. Thus, usual text datasets, such as Reuters, were not considered.

Figure 4 shows the percentage of the 64 papers selected published per year. It is worth noting that the most usual source (nearly 10%) of the 64 publications was Machine Learning[20].

The synthesis of the 64 selected publications according to the quality criteria (Page 8) is summarized in Table 2. Once more, it should be emphasised

[19] http://mlkd.csd.auth.gr
[20] http://www.springer.com/computer/ai/journal/10994

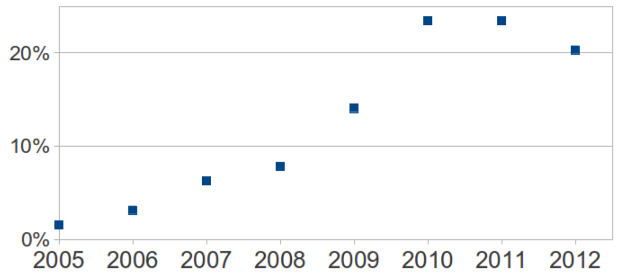

Figure 4: Percentage of the 64 papers published per year.

that the quality criteria are only verified in the selected results which do not fulfill any exclusion criteria.

QC 1	94%
QC 2	69%
QC 3	64%
QC 4	34%
QC 1, QC 2, QC 3, QC 4 fulfilled	14%
QC 1, QC 2, QC 3 fulfilled	52%
No QC fulfilled	0%

Table 2: Percentage of the 64 papers fulfilling quality criteria.

The quality criteria enable us to highlight some features from the 64 publications found by the SR process carried out in this work.

- As expected, most publications (94%) compare multi-label learning algorithms. Furthermore, 52% of the papers perform this comparison, using more than one publicly available and not preprocessed dataset, according to more than one evaluation measure. In fact, using several evaluation measures is important, as multi-label evaluation measures focus on different aspects of multi-label classifiers.
- Only 22 papers (34%) publish the standard deviations of the evaluation measures (QC 4). In addition, only 9 papers (14%) fulfill this criterion together with the other ones.
- Every paper fulfills at least one quality criterion, reinforcing that the criteria chosen can be useful to synthesize the selected papers.

4 Conclusion and Future Work

In this report we described the use of the systematic review process, which enabled us to find papers reporting experimental results for multi-label learning. A brief introduction about the SR process, including examples for some of its activities, as well as some summaries about the papers selected by the process, were presented.

The systematic review process is a useful method for bibliographical research that allows a wide, rigorous and reproducible literature exploration. In fact, the multi-label evaluation measure results published in the 64 papers were useful for comparison against a multi-label baseline classifier proposed in (Metz et al., 2012). These advantages compensate the additional effort required to carry out this process.

As is the case in some well known systematic literature reviews (Kitchenham et al., 2010), potentially relevant papers might not have been identified by us, as we only used nine electronic databases to search for publications. Nevertheless, the results found can be used as basis to validate future systematic reviews related to multi-label learning research.

The protocol proposed in this work, including the search string, the selection criteria and the quality criteria, could be used in forthcoming surveys on experimental multi-labeled learning. Furthermore, the application of the SR process in Artificial Intelligence and related areas could also use portions of this report as an initial support.

Acknowledgements

This research was supported by the Brazilian Research Council FAPESP.

References

Benitti, F. B. V. (2012). Exploring the educational potential of robotics in schools: A systematic review. *Computers & Education*, pages 978–988. Cited in page 1 and 4.

Biolchini, J., Mian, P. G., Natali, A. C. C., and Travassos, G. H. (2005). Systematic review in software engineering. Technical report, Alberto Luiz Coimbra Institute - Graduate School and Research in Engineering/Federal University of Rio de Janeiro - Brazil. Available at: http://www.cin.ufpe.br/~in1037/leitura/systematicReviewSE-COPPE.pdf. Cited in page 2.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., and Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, 80(4):571–583. Cited in page 1 and 8.

Carvalho, A. C. P. L. F. D. and Freitas, A. A. (2009). *A tutorial on multi-label classification techniques*, volume 5 of *Studies in Computational Intelligence 205*, pages 177–195. Springer. Cited in page 2.

Castro, A. A., Saconato, H., Guidugli, F., and Clark, O. A. C. (2002). A systematic review and meta-analysis course (in portuguese). Available at: http://www.virtual.epm.br/cursos/metanalise. Cited in page 2.

Cherman, E. A., Metz, J., and Monard, M. C. (2012). Incorporating label dependency into the binary relevance framework for multi-label classification. *Expert Systems with Applications*, 39(2):1647–1655. Cited in page 6.

Demsar, J. (2006). Statistical comparison of classifiers over multiple data sets. *Journal of Machine Learning Research*, pages 1–30. Cited in page 2.

Dietterich, T. G. (1990). Exploratory research in machine learning. *Machine Learning*, 5(1):5–10. Cited in page 11.

Fink, A. G. (2004). *Conducting Research Literature Reviews - From the Internet to Paper*. Sage Publication. Cited in page 5.

Guessi, M., de Oliveira, L. B. R., and Nakagawa, E. Y. (2011). Current state of representation of reference architectures. Technical report, University of São Paulo - Brazil. Available at: http://www.icmc.usp.br/~biblio/relatorios_tecnicos.php. Cited in page 1.

Higgins, J. P. T. and Green, S. (2009). Cochrane handbook for systematic reviews of interventions. The Cochrane Collaboration. Available at: http://www.cochrane-handbook.org/. Cited in page 2.

Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., Niazi, M., and Linkman, S. (2010). Systematic literature reviews in software engineering - a tertiary study. *Information and Software Technology*, 52(8):792–805. Cited in page 1, 2, and 13.

Kitchenham, B. A. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical report, Evidence-based Software Engineering - United Kingdom. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.1446&rep=rep1&type=pdf. Cited in page 1, 2, 3, and 5.

Langley, P. (2000). Crafting papers on machine learning. In *International Conference on Machine Learning*, pages 1207–1212. Cited in page 11.

Madeo, R. C. B. and Peres, S. M. (2012). Automated gesture analysis: a systematic review considering temporal aspects (in portuguese). EACH Technical Report PPgSI-002/2012. 18 pg. University of São Paulo. Cited in page 2.

Metz, J., Abreu, L. F. D., Cherman, E. A., and Monard, M. C. (2012). On the estimation of predictive evaluation measure baselines for multi-label learning. In Pavón, J., Duque-Méndez, N. D., and Fuentes-Fernández, R., editors, *Advances in Artificial Intelligence - IBERAMIA 2012*, volume 7637 of *Lecture Notes in Artificial Intelligence*, pages 189–198. Springer Berlin Heidelberg. Cited in page 6, 10, and 13.

Noblit, G. W. and Hare, R. D. (1988). *Meta-Ethnography: Synthesizing Qualitative Studies*. Sage Publications. Cited in page 5.

Petticrew, M. and Roberts, H. (2006). *Systematic Review in the Social Sciences: A Practical Guide*. Wiley-Blackwell. Cited in page 2.

Pisani, P. H. and Lorena, A. C. (2010). Intrusion detection with keystroke dynamics: a systematic review (in portuguese). Federal University of ABC Technical Report 06/2011. 28 pg. Available at: http://cmcc.ufabc.edu.br/index.php/relatorios-de-pesquisa/19-2011/187-rel-06.html. Federal University of ABC. Cited in page 2.

Spolaôr, N., Lorena, A. C., and Lee, H. D. (2010). A systematic review on multiobjective metaheuristics for feature selection (in portuguese). Federal University of ABC Technical Report 03/2010. 32 pg. Available at: http://cmcc.ufabc.edu.br/index.php/relatorios-de-pesquisa/14-2010/117-technical-report-032010.html. Federal University of ABC. Cited in page 2.

Spolaôr, N., Monard, M. C., and Lee, H. D. (2012). A systematic review to identify feature selection publications in multi-labeled data. ICMC Technical Report No 374. 31 pg. Available at: http://www2.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_374.pdf. University of São Paulo. Cited in page 1, 2, 4, 5, and 6.

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2010). Mining multi-label data. *Data Mining and Knowledge Discovery*, pages 1–19. Cited in page 1 and 2.

Zhang, H. and Babar, M. A. (2012). Systematic reviews in software engineering: An empirical investigation (in press). *Information and Software Technology*, (0):1–14. Cited in page 1 and 2.

A Appendix

A.1 Groups of Keywords and Search String

In what follows, groups of keywords and search string used to carry out the SR process are described.

set of labels ”multi-label”, ”multi label”, ”multilabel”, ”multiple label”, ”multiple labels”, ”label correlation”, ”label correlations”, ”correlation of label”, ”correlations of label”, ”correlation of labels”, ”correlations of labels”, ”label set”, ”label sets”, ”set of label”, ”sets of label”, ”set of labels”, ”sets of labels”, ”label relationship”, ”relationship of label”, ”relationships of label”, ”relationship of labels”, ”relationships of labels”, ”label dependence”, ”label dependencies”, ”dependence of label”, ”dependencies of label”, ”dependence of labels”, ”dependencies of labels”, ”label co-occurrence”, ”label co-occurrences”, ”co-occurrence of label”, ”co-occurrences of label”, ”co-occurrence of labels”, ”co-occurrences of labels”, ”label cooccurrence”, ”label cooccurrences”, ”cooccurrence of label”, ”cooccurrences of label”, ”cooccurrence of labels”, ”cooccurrences of labels”, ”label combination”, ”label combinations”, ”combination of label”, ”combinations of label”, ”combination of labels”, ”combinations of labels”.

multi-label learning ”multi-label algorithms”, ”multi-label algorithm”, ”multi-label learning”, ”multi-label classifiers”, ”multi-label classifier”, ”multi label algorithms”, ”multi label algorithm”, ”multi label learning”, ”multi label classifiers”, ”multi label classifier”, ”multilabel algorithms”, ”multilabel algorithm”, ”multilabel learning”, ”multilabel classifiers”, ”multilabel classifier”, ”machine learning algorithms”, ”machine learning algorithm”, ”supervised learning”, classifier, classifiers, classification, ”multi-label classification”, ”multilabel classification”, ”multi label classification”

experimental experimental, experiments, experiment, ”empirical evaluation”, ”empirical evaluations”, ”evaluation measures”, ”evaluation measure”, ”empirical research”, ”empirical researches”, baseline, baselines

SS (”multi-label” OR ”multi label” OR ”multilabel” OR ”multiple label” OR ”multiple labels” OR ”label correlation” OR ”label correlations” OR ”correlation of label” OR ”correlations of label” OR ”correlation of labels” OR ”correlations of labels” OR ”label set” OR ”label sets” OR ”set of label” OR ”sets of label” OR ”set of labels” OR ”sets of labels” OR ”label relationship” OR ”relationship of label” OR ”relationships of label” OR ”relationship of labels” OR ”relationships of labels” OR ”label dependence” OR ”label dependencies” OR ”dependence of label” OR ”dependencies of label” OR ”dependence of labels” OR ”dependencies of labels” OR ”label co-occurrence” OR ”label co-occurrences” OR ”co-occurrence of label” OR ”co-occurrences of label” OR ”co-occurrence of labels” OR ”co-occurrences of labels” OR ”label cooccurrence” OR ”label cooccurrences” OR ”cooccurrence of label” OR ”cooccurrences of label” OR ”cooccurrence of labels” OR ”cooccurrences of labels” OR ”label combination” OR ”label combinations” OR ”combination of label” OR ”combinations of label” OR ”combination of labels” OR ”combinations of labels”**) AND (**”multi-label algorithms” OR ”multi-label algorithm” OR ”multi-label learning”

OR ”multi-label classifiers” OR ”multi-label classifier” OR ”multi label algorithms” OR ”multi label algorithm” OR ”multi label learning” OR ”multi label classifiers” OR ”multi label classifier” OR ”multilabel algorithms” OR ”multilabel algorithm” OR ”multilabel learning” OR ”multilabel classifiers” OR ”multilabel classifier” OR ”machine learning algorithms” OR ”machine learning algorithm” OR ”supervised learning” OR classifier OR classifiers OR classification OR ”multi-label classification” OR ”multilabel classification” OR ”multi label classification”**) AND (**experimental OR experiments OR experiment OR ”empirical evaluation” OR ”empirical evaluations” OR ”evaluation measures” OR ”evaluation measure” OR ”empirical research” OR ”empirical researches” OR baseline OR baselines**)**

A.2 Information to be Extracted

In what follows, the 42 columns in the electronic spreadsheet are described. The corresponding entity-relationship model is described in Figure 1, Page 9.

1. Publication ID;
2. Publication title;
3. Publication year;
4. Source title;
5. Dataset name;
6. Dataset domain;
7. Dataset number of instances;
8. Dataset number of features;
9. Dataset number of labels;
10. Dataset label cardinality[21];
11. Dataset label density[22];
12. Dataset number of distinct combinations of labels;
13. Link to download the dataset;
14. Algorithm name;
15. Algorithm setup;
16. Base learner used by the algorithm;
17. Validation used in the experiments;
18. Result of the *example-based Accuracy* evaluation measure;
19. Result of the *example-based F-measure* evaluation measure;

[21] Average number of labels associated with each instance.

[22] Normalized label cardinality.

20. Result of the *Hamming Loss* evaluation measure;
21. Result of the *example-based Precision* evaluation measure;
22. Result of the *example-based Recall* evaluation measure;
23. Result of the *Subset Accuracy* evaluation measure;
24. Result of the *Macro-averaged F-measure* evaluation measure;
25. Result of the *Macro-averaged Precision* evaluation measure;
26. Result of the *Macro-averaged Recall* evaluation measure;
27. Result of the *Micro-averaged F-measure* evaluation measure;
28. Result of the *Micro-averaged Precision* evaluation measure;
29. Result of the *Micro-averaged Recall* evaluation measure;
30. Result of the *per label Accuracy* evaluation measure[23];
31. Result of the *per label F-measure* evaluation measure;
32. Result of the *per label Precision* evaluation measure;
33. Result of the *per label Recall* evaluation measure;
34. Result of the *Mean Accuracy* evaluation measure;
35. Result of the $\alpha = 0$ *Accuracy* evaluation measure;
36. Result of the *Subset 0/1 Loss* evaluation measure;
37. Result of the *Log Loss* evaluation measure;
38. Result of the *Jaccard Index* evaluation measure;
39. Result of the *Macro-averaged AUC* evaluation measure;
40. Result of the *Micro-averaged AUC* evaluation measure;
41. Observations.

[23] Measure applied separately on each label.