dr hab. inż. Krzysztof Dembczyński
Instytut Informatyki (Institute of Computing Science)
Politechnika Poznańska (Poznań University of Technology)
ul. Piotrowo 2, 60-965 Poznań
tel: (+48) 61 665 2936
kdembczynski@cs.put.poznan.pl

Poznań, September 11, 2019

Review of Piotr Szymański's thesis: A Network Science Perspective on Label Dependencies in Multi-Label Classification

1 Summary of the thesis

The thesis concerns multi-label classification, a machine learning problem in which the task is to construct a classifier which accurately assigns subsets of predefined labels to objects of interest. A classifier is usually built by a learning algorithm on a set of training examples consisting of objects and their known label subsets. The accuracy of a classifier is measured in terms of loss functions (the lower value of a loss the better performance of the classifier) or performance metrics. As several labels can be assigned simultaneously to an example, label dependencies play an important role in designing learning algorithms for multi-label classification. Their proper modeling is crucial in optimizing the performance of a classifier.

The author proposes to model label dependencies using tools from network sciences. The label dependencies in the training set can be expressed by a network or graph. Each node in this graph corresponds to a label. An edge between two nodes indicates that the corresponding labels have been assigned simultaneously to at least one training example. For such graph one can use a community detection algorithm, a popular tool for network analysis, to identify subgroups of strongly related labels. Another solution introduced in the thesis is to embed the nodes of the graph, i.e., labels, in a high-dimensional hidden space representing similarities between labels in terms of typical distances such as Euclidean or cosine. The algorithms used for this goal are similar to the ones for constructing hidden representation of words for natural language processing applications (see, e.g., [3]). The ultimate goal of the thesis is thus design and development of new competitive learning algorithms for multi-label classification, which exploit tools rooted in network sciences to model label dependencies.

Additional contribution of the author is his work on development of a popular Python package, `scikit-multilearn`, which contains a wide spectrum of algorithms for multi-label classification, not only those introduced in this thesis. The author also discusses the problem of appropriate preparation of benchmark datasets for multi-label classification. He notes that for many datasets, the distributions of labels in the training and test set significantly differ from each other. Therefore, he proposes a new algorithm for stratified sampling for multi-label data, which produces training and test sets of higher similarity in terms of the label distribution.

The thesis consists of an introduction and three parts. The introduction contains a general overview of multi-label classification and discusses the main motivations and goals of the work. The contribution of the author to the field is detailed along with a list of his main publications. The outline of the thesis is also presented.

The first part of the thesis contains a problem statement and a description of existing algorithms, benchmark datasets and software packages. Its first chapter presents the formal definition of the

multi-label classification problem. The author discusses in detail the problem of modeling label dependencies and defines popular performance metrics. The second chapter of this part provides a broad description of the most commonly used algorithms. The next chapter describes popular benchmark datasets and commonly accepted scenarios for conducting empirical studies. The author discusses here the problem of splitting data into a training and test set and the related problem of stratified sampling. Chapter 5 of the thesis presents various software packages for multi-label classification, with a particular emphasis on `scikit-multilearn`, being developed by the author. The empirical results presented in this chapter indicate outstanding performance of this package. In most cases the same algorithm implemented in other packages gets worse running times and memory consumption. The last chapter of this part provides a brief summary of its content.

The second part of the thesis contains description of the introduced learning algorithms and stratified sampling. In Chapter 7 of the thesis, the author introduces an algorithm, *Community Detection in Label Network* (CDLN), that is based on partitioning of the set of labels into disjoint subsets using community search algorithms. For each such partition (a subset of labels) a separate multi-label classifier is trained. For a test example, the final assignment is computed as a set-union of responses from each multi-label classifier. The introduced algorithm has been compared to the RAkEL algorithm [10], which uses a random partition of the set of labels. The empirical results clearly indicate the superiority of CDLN. In the next chapter, the author introduces the LNEMLC algorithm. In the first step it builds a hidden representation of labels using network/graph embedding methods such as node2vec [2]. The resulting label embedings are then used as additional features in a multi-label classifier. Since their values are not know during prediction, a multivariate regression algorithm is used to map original feature vectors of examples to the hidden representation of labels. In the experiment, the introduced algorithm obtains some advantages over existing algorithms. In addition, the author estimates the "potential" of the algorithm in case of a perfect multivariate regression model (with zero error). The last chapter of this part deals with the problem of stratified sampling. The author introduces a solution that maximizes a similarity between label-connection graphs of training and test sets.

The last part consists of appendices containing lists of publications, scientific internships and projects of the author, as well as additional empirical results and the bibliography.

2 General Comments and Remarks

The thesis is based on six articles co-authored by the PhD candidate. Two of them have been published in top journals, for example, in *Journal of Machine Learning Research* which is the most prestigious journal in the field of machine learning. Three articles have been presented at conferences or scientific workshops. The article describing the LNEMLC algorithm has not yet been published in a peer-reviewed venue or journal, but is available from the Arxiv repository. The exact bibliographic details of these articles are given below:

- Piotr Szymański and Tomasz Kajdanowicz. MLG: Enchancing multi-label classification with modularity-based label grouping. In Jeng-Shyang Pan, Marios M. Polycarpou, Michał Woźniak, André C. P. L. F. de Carvalho, Héctor Quintián, and Emilio Corchado, editors, *Hybrid Artificial Intelligent Systems*, pages 431–440, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg
- Piotr Szymański, Tomasz Kajdanowicz, and Kristian Kersting. How is a data-driven approach better than random choice in label space division for multi-label classification? *Entropy*, 18(8):282, 2016
- Piotr Szymański and Tomasz Kajdanowicz. Is a data-driven approach still better than random choice with naive Bayes classifiers? In Ngoc Thanh Nguyen, Satoshi Tojo, Le Minh Nguyen, and Bogdan Trawiński, editors, *Intelligent Information and Database Systems*, pages 792–801. Springer International Publishing, 2017
- Piotr Szymański and Tomasz Kajdanowicz. A network perspective on stratification of multi-label data. In Luís Torgo, Bartosz Krawczyk, Paula Branco, and Nuno Moniz, editors, *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 74 of *Proceedings of Machine Learning Research*, pages 22–35, ECML-PKDD, Skopje, Macedonia, 22 Sep 2017. PMLR

- Piotr Szymański and Tomasz Kajdanowicz. scikit-multilearn: A Python library for multi-label classification. *Journal of Machine Learning Research*, 20(6):1–22, 2019
- Piotr Szymanski, Tomasz Kajdanowicz, and Nitesh Chawla. LNEMLC: label network embeddings for multi-label classifiation. *CoRR*, abs/1812.02956, 2018

This is certainly an interesting and well-written thesis. The first chapters introduce a reader to the world of multi-label classification in a formal but simple way. One can see from the writing that the author has a wide mathematical background. Unfortunately, the later chapters of the thesis lack a deeper theoretical analysis of the introduced algorithms. This makes some of the drawn conclusions to be incomplete or incorrect. To avoid this the author should confront his knowledge with existing literature on the theoretical foundations of multi-label classification. An appropriate analysis of the algorithms would significantly improve the thesis. Nevertheless, the proposed algorithms are promising and provide a basis for further research. The text of the thesis, tables, figures, and charts have been carefully formatted. Unfortunately, the text contains several typographic mistakes such as unfinished sentences, missing words, or a wrong type of citation. I suppose that these typos have been accidentally included during the final edits of the thesis.

It should be emphasized that the PhD candidate is an author of `scikit-multilearn`, a very popular software package for multi-label classification. He has also managed to establish extensive international collaboration by visiting such research centers as Stanford University, Josef Stefan Institute in Ljubljana, University of Technology Sydney, University of Notre Dame, and Technical University of Dortmund.

3 Detailed comments and remarks

As already mentioned above, the main disadvantage of the thesis is the lack of the theoretical analysis of the introduced algorithms. For example, the author does not specify for what performance measure the CDLN algorithm has been designed. The experiments concern a wide spectrum of measures, but without theoretical analysis it is difficult to draw conclusions from the reported results. Let me observe that if the partition of labels obtained by a community detection algorithm factorized the conditional joint distribution into independent subsets of labels, then minimization of the subset 0/1 loss for each subset of labels would imply minimization of the subset 0/1 loss for all labels. The above reasoning is similar to [1]. In this article, however, a partition of labels has been obtained by using statistical conditional independence tests and Markov boundary discovery algorithms. From the thesis, we unfortunately do not learn what kind of relation between labels is modeled by the network approach. Note that for multi-class data, for which the label dependencies are strong (since labels are exclusive), the resulting graph will consist of disjoint nodes without any edges. Should this be a correct partition of labels in this case?

Similar remarks apply to the LNEMLC algorithm. There is no theoretical analysis trying to answer when the algorithm performs well. Furthermore, there is another quite controversial element concerning this algorithm. The author estimates the "potential" of the algorithm in case of a perfect multivariate regression model. One has to remark that the resulting mapping of labels to the hidden representation is an injection. In other words, there is 1-to-1 correspondence between labels and their hidden representations. The same concerns the representation of label subsets. This means that a perfect multivariate regression model should lead to a perfect multi-label classifier. When estimating the "potential" of the algorithm, one can easily obtain perfect results by random projection of labels to a high-dimensional space and using a sufficiently complex multi-label classifier which gets zero training error. I suppose that this is not a desired result. Summarizing, this part of experiments says nothing about the introduced method. It is possible that label embeddings used as additional features improve the performance. Unfortunately, the thesis does not contain convincing theoretical arguments supporting this approach. Nevertheless, some of the empirical results presented by the author are promising.

In designing multi-label classifiers one should distinguish two subproblems. The first one is optimization of classifier's responses with respect to a considered performance metric. The second one is estimation of parameters required by the optimization subproblem. This formulation simplifies construction and theoretical analysis of multi-label classifiers. It also enables straight-forward interpretation of the empirical results.

When discussing probabilistic classifier chains, the author does not cite the original article in which the algorithm has been introduced. Moreover, the author should properly distinguish inference algorithms, which for a given test example seek the best label subset with respect to a given performance measure, from algorithms that seek the best label order to be used during training or inference.

In the main text, the empirical results are mainly presented in terms of rankings of methods and the change of positions in the rankings. The quantitative results are only given in the appendix. The author should present both types of results in the main text to avoid any confusions. The quantitative results allow to precisely verify whether the tested methods perform as expected from the theoretical analysis.

4 Minor remarks

Below I list some minor remarks concerning the text:

- Chapter 2.3 should contain references to articles discussing label dependencies.
- Chapter 3.1.1: Please note that both information gain and entropy boil down to the same splitting criterion.
- Chapter 3.1.3: It should be clarified whether the author uses row or column vectors.
- Chapter 3.2.1.2: This statement $P(y_3|h_2(x), h_1(x), x) \neq P(y_3|h_1(x), h_2(x), x)$ is mathematically incorrect.
- Page 33 and 34: The paragraph containing description of Homer lacks the name of the algorithm.
- Page 34: In the sentence "CLEMS (...) obtains E" it is not clear what E is.
- Page 64: "It's complexity" $\Leftarrow$ "Its complexity".
- Page 92: Instead "<<" please use "$\ll$" (`\ll`).
- Page 107: What are "second-order label dependencies"?
- An example of an unfinished sentence (page 29): "(...): uses the MAP principle to assing labels to a new"
- An example of a wrong citation type (page 33): "Random k-labelsets RAkEL Tsoumakas et at. [2011a]" $\Rightarrow$ "Random k-labelsets (RAkEL) [Tsoumakas et at., 2011a]"
- An example of missing words (page 62): "Modularity optimizing approaches (...) measure the between two probabilistic settings"

5 Final conclusion

The critical remarks and comments presented above should be treated as a scientific discussion which aims at directing the author to more in-depth research in the field of multi-label classification. The introduced algorithms are interesting and obtain promising empirical results. Complementing the dissertation with their theoretical analysis would undoubtedly be a valuable result. Nevertheless, I assess the current content of the thesis very well. Summing up the above review, it should be stated that the goal of the thesis has been achieved. Considering all the above comments and remarks, I am asking for admission of Piotr Szymański to further stages of the doctoral examination.

dr hab. inż. Krzysztof Dembczyński

References

[1] Maxime Gasse, Alex Aussem, and Haytham Elghazel. On the optimality of multi-label classification under subset zero-one loss for distributions satisfying the composition property. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2531–2539. JMLR.org, 2015.

[2] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864. ACM, 2016.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119. Curran Associates Inc., 2013.

[4] Piotr Szymański and Tomasz Kajdanowicz. MLG: Enchancing multi-label classification with modularity-based label grouping. In Jeng-Shyang Pan, Marios M. Polycarpou, Michał Woźniak, André C. P. L. F. de Carvalho, Héctor Quintián, and Emilio Corchado, editors, *Hybrid Artificial Intelligent Systems*, pages 431–440, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[5] Piotr Szymański and Tomasz Kajdanowicz. Is a data-driven approach still better than random choice with naive Bayes classifiers? In Ngoc Thanh Nguyen, Satoshi Tojo, Le Minh Nguyen, and Bogdan Trawiński, editors, *Intelligent Information and Database Systems*, pages 792–801. Springer International Publishing, 2017.

[6] Piotr Szymański and Tomasz Kajdanowicz. scikit-multilearn: A Python library for multi-label classification. *Journal of Machine Learning Research*, 20(6):1–22, 2019.

[7] Piotr Szymanski, Tomasz Kajdanowicz, and Nitesh Chawla. LNEMLC: label network embeddings for multi-label classifiation. *CoRR*, abs/1812.02956, 2018.

[8] Piotr Szymański and Tomasz Kajdanowicz. A network perspective on stratification of multi-label data. In Luís Torgo, Bartosz Krawczyk, Paula Branco, and Nuno Moniz, editors, *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 74 of *Proceedings of Machine Learning Research*, pages 22–35, ECML-PKDD, Skopje, Macedonia, 22 Sep 2017. PMLR.

[9] Piotr Szymański, Tomasz Kajdanowicz, and Kristian Kersting. How is a data-driven approach better than random choice in label space division for multi-label classification? *Entropy*, 18(8):282, 2016.

[10] Grigorios Tsoumakas and Ioannis Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *Proceedings of the 18th European Conference on Machine Learning*, ECML '07, pages 406–417. Springer-Verlag, 2007.