Report on the 1st International Workshop on Learning to Quantify (LQ 2021)

Juan José del Coz and Pablo González
Artificial Intelligence Center
University of Oviedo
33204 Gijón, Spain
{juanjo,gonzalezgpablo}@uniovi.es

Alejandro Moreo and Fabrizio Sebastiani
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, Italy
{alejandro.moreo,fabrizio.sebastiani}@isti.cnr.it

ABSTRACT

The 1st International Workshop on Learning to Quantify (LQ 2021 – `https://cikmlq2021.github.io/`), organized as a satellite event of the 30th ACM International Conference on Knowledge Management (CIKM 2021), took place on two separate days, November 1 and 5, 2021. As the main CIKM 2021 conference, the workshop was held entirely online, due to the COVID-19 pandemic. This report presents a summary of each keynote speech and contributed paper presented in this event, and discusses the issues that were raised during the workshop.

1. LEARNING TO QUANTIFY

In a number of applications involving classification, the final goal is not determining which class (or classes) individual unlabelled instances belong to, but estimating the *prevalence* (or "relative frequency", or "prior probability", or simply "prior") of each class in the unlabelled data. Estimating class prevalence for unlabelled data via supervised learning is known as *Learning to Quantify* (LQ) (or *quantification*, or *supervised prevalence estimation*).

LQ has several applications in fields (such as the social sciences, political science, market research, and epidemiology) which are inherently interested in characterizing *aggregations* of individuals, rather than the individuals themselves; disciplines like the ones above are usually *not* interested in finding the needle in the haystack, but in characterising the haystack itself. For instance, in most applications of tweet sentiment classification we are not concerned with estimating the true class (e.g., Positive, Negative, or Neutral) of individual tweets. Rather, we are concerned with estimating the relative frequency of these classes in the set of unlabelled tweets under study; or, put in another way, we are interested in estimating as accurately as possible the true distribution of tweets across the classes.

It is well known that performing quantification by classifying each unlabelled instance and then counting the instances that have been attributed the class (the "classify and count" method) usually leads to suboptimal quantification accuracy; this is a direct consequence of "Vapnik's principle" [11], which states

> If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.

As a result of the suboptimality of the "classify and count" method, learning to quantify has slowly evolved as a task in its own right, different (in goals, methods, techniques, and evaluation measures) from classification [3]. The research community has investigated methods to correct the biased prevalence estimates of general-purpose classifiers, supervised learning methods specially tailored to LQ, and evaluation measures for LQ. Specific applications of LQ have also been investigated, such as sentiment quantification, quantification in networked environments, or quantification for data streams. For the near future, it is easy to foresee that the interest in learning to quantify will increase, due (a) to the increased awareness that "classify and count" is a suboptimal solution when it comes to prevalence estimation, and (b) to the fact that, with larger and larger quantities of data becoming available and requiring interpretation, in more and more scenarios we will only be able to afford to analyse these data at the aggregate level rather than individually.

2. THE WORKSHOP

Uncharacteristically, LQ 2021 consisted of two half-day sessions which took place on two separate days, November 1 and 5, 2021, due to the fact that attending a single day-long event would have required night shifts on the part of most delegates, which were scattered across the globe. In fact, LQ 2021 had participants from Australia, Japan, Italy, Switzerland, Germany, Spain, US, and others.

The workshop consisted of two keynote speeches, six contributed talks, and a final brainstorming session. The six contributed talks were selected by a program committee consisting of 12 renowned LQ experts.

The first session featured a keynote talk and four contributed talks. The keynote talk ("Mixture proportion estimation in weakly supervised learning" [9]) was presented by **Masashi Sugiyama** (RIKEN and the University of Tokyo). In this talk, Sugiyama surveyed his work on mixture proportion estimation under different settings. The first part of the talk was devoted to semi-supervised classification under class-prior shift. His approach to solve this problem is based on distribution matching by density ratio estimation using different measures, including the Kullback-Leibler divergence;

interestingly, these distances can be minimised without estimating any density. The second part of Sugiyama's talk was on positive-unlabelled classification. The problem of this setting is that class prior is not identifiable in general. His first solution is based on the idea that class prior estimation can be computed by non-traditional classification, because it can be interpreted as partial distribution matching with Pearson divergence. In the most common case, when classes overlap, partial matching generally overestimates the true class prior. One solution to overcome this issue is to assume that there exists at least an anchor point, which allows deriving a nice and simple method; however, the anchor point assumption can be too strong in practical cases. Sugiyama described alternative approaches such as partial matching with penalized loss functions and regrouping.

Sugiyama's keynote was followed by four contributed presentations. In the first one [8], **Tetsuya Sakai** compared measures for evaluating ordinal LQ systems, i.e., systems that operate on a totally ordered set of $n > 2$ classes. In particular, he compared Normalised Match Distance (NMD – a normalised version of a particular case of the Earth Mover's Distance) with Root Normalised Order-aware Divergence (RNOD) in terms of their ability to provide stable system rankings regardless of the test set being used ("system ranking consistency"), and of their ability to differentiate, in a statistically significant way, among different systems under the same experimental conditions. Sakai's analysis concludes that both measures have their pros and cons, and that they should thus be used in parallel when evaluating ordinal LQ systems.

Dirk Tasche ("Minimising quantifier variance under prior probability shift" [10]) revisited the binary quantification problem by analysing theoretically the asymptotic variance of the maximum likelihood (ML) estimator. He found that this asymptotic variance is closely related to the Brier score for the regression of the class label against the input features under the test set distribution. This finding opens the door to methods based on learning base classifiers that minimize both the Brier score on the training data and the Brier score on the test data in order to reduce the variance of the ML estimator of the prevalence values on the test data. Due to the statistical consistency of ML estimators, by reducing the variance of the estimator its mean squared error is minimized too.

Gustavo Batista presented "The Risks of Using Classification Datasets in Quantification Assessment" [5], a joint work with Waqar Hassan and André Maletzke, in which a critical reassessment of the current protocols for evaluating the performance of quantification methods, i.e., the natural-prevalence protocol (NPP) and the artificial-prevalence protocol (APP), is carried out. In their paper, the shortcomings and potential risks that the adoption of the APP might bring for the performance assessment of quantification systems is highlighted, putting special emphasis on the fact that knowing in advance the expected value of the target prevalence values that the APP generates (e.g., $\mathbb{E}[p] = 0.5$ for the positive class in binary quantification) might be "maliciously" exploited by some methods in evaluation campaigns. The authors also propose that a trivial baseline that always returns $\mathbb{E}[p]$ should be adopted in comparative evaluations, and recommends the adoption of "radar charts" as a tool for visualizing and comparing results.

Finally, **Pablo González** presented his joint work with Juan José del Coz [4] which describes a deep neural network architecture for quantification, called HistNet, and based on differentiable histogram representations of the samples. HistNet has the ability to work without labels if only sample prevalences are available, but can also exploit label information through an extra (and optional) network connection. HistNet is applicable to many types of problems only by changing the feature extraction layer, and was tested on two public datasets, one from computer vision and the other from sentiment analysis. The authors present promising results for the two of them, in both the binary and the multiclass settings.

The second session consisted of two contributed talks, a keynote speech, and a final brainstorming session. While the four contributed talks of the first session had a "vertical" nature, the two contributed talks of the second session were of a "horizontal" type. In the first such talk, **Alejandro Moreo** presented (joint work with Andrea Esuli and Fabrizio Sebastiani) QuaPy (`https://github.com/HLT-ISTI/QuaPy`), an open-source Python-based software library for LQ. QuaPy (which was also the object of a presentation in the main CIKM 2021 conference – see [7]) provides implementations of both baseline and advanced LQ methods, of routines for LQ-oriented model selection, of several broadly accepted evaluation measures, and of robust evaluation protocols routinely used in the field. QuaPy also makes available datasets commonly used for testing quantifiers, and offers visualization tools for facilitating the analysis and interpretation of the results. The software is open-source and publicly available under a BSD-3 licence via GitHub, and can be installed via `pip`.

In the talk that followed, **Fabrizio Sebastiani** presented LeQua 2022 (`https://lequa2022.github.io/`), an upcoming shared task (jointly organized with Andrea Esuli and Alejandro Moreo) devoted to the evaluation of LQ systems [2]. LeQua 2022, an initiative held under the umbrella of the CLEF 2022 conference (`https://clef2022.clef-initiative.eu/`), is the first shared task ever whose main focus is quantification, and aims at pulling together the various subcommunities (from statistics, information retrieval, data mining, machine learning, etc.) that work on this task by providing a common, TREC-style evaluation framework. LeQua will provide subtasks for both binary quantification and multiclass quantification, and will cater both (by providing documents in vector form) for participants that are not interested/competent in generating vectorial representations of text, and (by providing documents in raw form) for participants wishing to engage in the optimization of end-to-end systems.

These two talks were followed by a keynote speech titled "An Improved Method of Automated Nonparametric Content Analysis for Social Science" [6] in which **Connor Jerzak** discussed (joint work with Gary King and Anton Strezhnev) an improved version of the well-known ReadMe system developed by Gary King and his then-colleagues more than 10 years ago. The original ReadMe system allows for direct estimation of class prevalence values without resorting to classification (i.e., it squarely belongs to the *non-aggregative* group of quantification algorithms), and was originally developed in order to work with textual data (verbal autopsies) converted into a bag-of-words format. The improvements carried out in this revised version of ReadMe focus on devising more robust means for representing text; in particular,

moving to dense vectorial representations (i.e., word embeddings), and endowing the system with strategies aiming to accommodate possible changes in the use of language, to improve the discriminative power of word representations, and to capitalize on non-redundant features. The extensive evaluation served to showcase the relative merits of the improved variant over the original one and other baseline quantification systems.

The workshop ended with a brainstorming session, which had the double purpose of acting as an overflow space for all those questions for which there had been no time during the various Q&A sessions, and of allowing the LQ community to discuss, in an unconstrained, informal setting, what participants felt are the burning, yet unresolved issues in this discipline.

3. CONCLUSION

LQ 2021 was the first event entirely devoted to learning to quantify. The workshop turned out to be a success, despite the difficulties inherent in participating remotely to events taking place at inconvenient hours of the day, and was especially successful in terms of the (very high) level of interactivity that characterized the Q&A sessions and the brainstorming session. It is our impression that, for LQ 2021, the online format was not an undesired side effect of the COVID-19 pandemic but a key to success, because researchers active on LQ are, as mentioned above, too scattered across different communities (statistics, information retrieval, data mining, machine learning, etc.), and, as a result, it would be hard to identify an in-presence conference alongside which to organize the workshop and to which all these different sub-communities would gladly travel.

The proceedings of LQ 2021 appear in the collective volume that hosts the proceedings of all workshops co-located with CIKM 2021, as published in the CEUR Workshop Proceedings series [1].

4. ACKNOWLEDGEMENTS

The work by the third and fourth authors was supported by the SoBigData++ project, funded by the European Commission (Grant 871042) under the H2020 Programme INFRAIA-2019-1, and by the AI4Media project, funded by the European Commission (Grant 951911) under the H2020 Programme ICT-48-2020. These authors' opinions do not necessarily reflect those of the European Commission.

5. REFERENCES

[1] G. Cong and M. Ramanath, editors. *Proceedings of the CIKM 2021 Workshops, co-located with the 30th ACM International Conference on Information and Knowledge Management (CIKM 2021), Virtual Event, November 1–5, 2021*, CEUR Workshop Proceedings. CEUR-WS.org, 2021. Forthcoming.

[2] A. Esuli, A. Moreo, and F. Sebastiani. LeQua@CLEF2022: Learning to Quantify. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR 2022)*, pages 374–381, Stavanger, NO, 2022.

[3] P. González, A. Castaño, N. V. Chawla, and J. J. del Coz. A review on quantification learning. *ACM Computing Surveys*, 50(5):74:1–74:40, 2017.

[4] P. González and J. J. del Coz. Histogram-based deep neural network for quantification (abstract). In *[1]*, 2021.

[5] W. Hassan, A. Maletzke, and G. Batista. The risks of using classification datasets in quantification assessment. In *[1]*, 2021.

[6] C. T. Jerzak, G. King, and A. Strezhnev. An improved method of automated nonparametric content analysis for social science. *Political Analysis*, 2022.

[7] A. Moreo, A. Esuli, and F. Sebastiani. QuaPy: A Python-based framework for quantification. In *Proceedings of the 30th ACM International Conference on Knowledge Management (CIKM 2021)*, pages 4534–4543, Gold Coast, AU, 2021.

[8] T. Sakai. A closer look at evaluation measures for ordinal quantification. In *Proceedings of the CIKM 2021 Workshop on Learning to Quantify*, Virtual Event, 2021.

[9] M. Sugiyama. Mixture proportion estimation in weakly supervised learning (abstract). In *[1]*, 2021.

[10] D. Tasche. Minimising quantifier variance under prior probability shift. In *[1]*, 2021.

[11] V. Vapnik. *Statistical learning theory*. Wiley, New York, US, 1998.