Report on the 3rd International Workshop on Learning to Quantify (LQ 2023)

Mirko Bunse[1], Pablo González[2], Alejandro Moreo[3], and Fabrizio Sebastiani[3]
[1] Lamarr Institute for Machine Learning and Artificial Intelligence,
TU Dortmund University, 44227 Dortmund, DE
[2] Artificial Intelligence Center, University of Oviedo, 33204 Gijón, ES
[3] Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, 56124 Pisa, IT
mirko.bunse@cs.tu-dortmund.de, gonzalezgpablo@uniovi.es,
alejandro.moreo@isti.cnr.it, fabrizio.sebastiani@isti.cnr.it

ABSTRACT

The 3rd International Workshop on Learning to Quantify (LQ 2023)[1] took place on September 18, 2023 in Torino, IT, where it was organised as a satellite event of the 34th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2023). Like the main program of the conference, the workshop employed a hybrid format, with all presentations given in presence and with attendees participating in presence or online. This report presents a summary of the workshop, briefly summarising the individual works presented, and touching on the main issues that emerged during the final, open discussion.

1. INTRO: LEARNING TO QUANTIFY

Many fields such as the social sciences, political science, market research, or epidemiology (to name a few), are inherently interested in *aggregate* data, i.e., in how populations of individuals are distributed according to one or more indicators of interest. Researchers who operate in these specialty areas are instead little interested in the individual data items *per se*, since in these fields the individual data items are relevant only inasmuch as they are members of the population of interest; in other words, disciplines such as the above are interested not in finding the needle, but in characterising the haystack.

Sometimes, researchers active in these fields use supervised learning to obtain the data they need. For instance, epidemiologists interested in the distribution of the causes of death across different geographical regions may sometimes need to *infer* the cause of death of each person by classifying, via a machine-learned text classifier, a verbal description of the symptoms that affected a deceased person. However, epidemiologists are not specifically interested in the class (representing a given cause of death) to which an individual belongs; rather, their final goal is estimating the *prevalence* (i.e., *relative frequency*, or *prior probability*) of each class in the unlabelled data.

At a first glance, estimating these prevalence values via supervised learning looks like a direct application of classification, since one could simply (i) train a classifier on labelled data, (ii) use this classifier to issue label predictions for each unlabelled datapoint (i.e., for each individual) in the population of interest, (iii) count how many datapoints have been attributed to each of the classes of interest, and (iv) normalize the counts by the total number of labelled datapoints, thus obtaining the estimated relative frequencies of the classes. However, there is by now abundant evidence that such an approach, known in the literature as the "Classify and Count" (CC) method, yields poor class prevalence estimates when the distribution of the unlabelled datapoints across the classes differs substantially from the analogous distribution observed during training [12]. This latter condition is typically known as *prior probability shift* (or *label shift*); in the aforementioned disciplines this condition is ubiquitous since, quite obviously, there is interest in inferring a distribution *only* if we consider the possibility of this unknown distribution to be different from the known distribution of the training data. The main reason for which CC tends to fail in the presence of prior probability shift is a violation of the IID assumption on which many supervised learning methods are based.

Due to the suboptimality of CC, learning to quantify has slowly evolved as a task in its own right, different from classification in terms of goals, methods, techniques, and evaluation measures [9, 13]. The research community has investigated methods to correct the biased prevalence estimates of general-purpose classifiers, supervised learning methods specifically tailored to LQ, and evaluation measures for LQ. 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 an analysis of these data at the aggregate level rather than at the individual level. At ECML/PKDD 2023, the increase in awareness about LQ manifested in an LQ paper [8] being given the best student paper award.

[1] https://lq-2023.github.io/

2. THE WORKSHOP

LQ 2023 was a combined tutorial + workshop event, i.e., it consisted of a half-day (morning) tutorial on quantification (which was taught by the third and fourth authors of the present paper) plus a half-day (afternoon) workshop. The tutorial introduced the rationale for quantification (also discussing the relationship between quantification and various types of dataset shift), discussed a number of domains to which quantification has been applied to, presented the main evaluation measures and the main experimental protocols that have been used in the quantification literature, and gave an overview of the main classes of methods that have been used for performing quantification. The tutorial also included a "hands-on" session, which made use of the QuaPy open-source Python library for quantification developed by the instructors [18]. The workshop consisted instead of the oral presentations of seven papers submitted in response to the call for papers, plus a final brainstorming session, in which the perceived "burning issues" of the field were brought up and discussed. The combine event was attended by about 30 people, of which about 20 in-presence and about 20 online.

2.1 The papers

The call for papers had asked for either completely original papers *or* papers that had recently (i.e., in 2023) been submitted / accepted / published in other workshops / conferences / journals; as a result, 3 papers of the first kind [4, 15, 21] and 4 papers of the latter kind [6, 7, 11, 19] were accepted for presentation at the workshop. The seven contributed talks were selected by a program committee consisting of 12 renowned LQ experts; each paper was reviewed by at least 3 members of the committee.

Dirk Tasche presented his work "Invariance assumptions for class distribution estimation" [21], which discusses various assumptions of invariance between the distributions of training and test data (covariate shift, factorizable joint shift, and sparse joint shift) and the implications that each of these assumptions has on quantification learning. While factorizable joint shift is shown to yield no consistent estimator of class prevalence values, sparse joint shift is shown to generalize prior probability shift, the type of shift which is most frequently assumed in quantification learning.

Gustavo Batista presented "MC-SQ and MC-MQ: Ensembles for multi-class quantification", a joint work with Zahra Donyavi and Adriane Serapião that was the subject of a recent conference publication [7]. The presentation started by noting that tackling multi-class quantification via the one-versus-all strategy (consisting of training an ensemble of independent binary quantifiers, one per class) generally yields poor performance. The authors thus propose *multiple-classifiers single-quantifier* (MC-SQ) and *multiple-classifiers multiple-quantifiers* (MC-MQ), two new ensemble-based models for quantification that couple different classifiers with different aggregative quantifiers. The experiments presented showed that MC-SQ and MC-MQ are new important contenders in the multiclass quantification arena.

Alessandro Fabris presented "Measuring Fairness under Unawareness of Sensitive Attributes: A Quantification-Based Approach", a joint work with Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani that was the subject of a recent journal publication [11]. In this paper, the authors use a method based on quantification in order to measure how fair (or how biased) a classifier is with respect to a given sensitive attribute (e.g., race, sex) which is not among the attributes (i.e., covariates) used by the classifier ("unawareness"). The authors conclude that the method has two important advantages, i.e., (a) it drastically outperforms methods based on standard classification, and (b) it manages, differently from methods based on classification, to obtain the desired aggregate-level predictions while not allowing undesired predictions at the individual level.

Mirko Bunse presented his work "Qunfold: Composable Quantification and Unfolding Methods in Python" [4], which documents a novel software package for quantification methods. Building on the findings of two works that the author presented at the 2$^{\text{nd}}$ edition of the LQ workshop, this easy-to-use package allows its users to compose novel quantification methods from existing loss functions and data representations [3], and it implements a powerful optimization technique [2] that the author shows to achieve lower prediction errors than other state-of-the-art implementations.

Pablo González presented "An Equivalence Analysis of Binary Quantification Methods", a joint work with Alberto Castaño, Jaime Alonso, and Juan J. del Coz. In this paper the authors analyse several algorithms falling under the distribution-matching framework [17], finding theoretical connections and equivalences between them. They also propose a new method called "QUANTy", based on average probabilities computed in different quantiles. The authors conclude by emphasizing the importance of using richer representations for characterising the distributions (in contrast to simplistic representations that may lose important information about them). The paper does not appear in the proceedings of LQ 2023 but is available as [6].

Kevin Kloos presented "Continuous Sweep: An Improved Binary Quantifier", a joint work with Julian D. Karch, Quinten A. Meertens, and Mark de Rooij. In this paper the authors propose a binary quantification method called "Continuous Sweep", based on the well-known "Median Sweep" quantification algorithm [12]. Continuous Sweep is a quantifier that uses parametric class distributions instead of empirical distributions. The paper presents theoretical derivations for the bias and variance of the method; simulation studies show that it outperforms its predecessor on datasets characterised by prior probability shift. The paper does not appear in the proceedings of LQ 2023 but is available as [15].

Alejandro Moreo presented the paper "Multi-Label Quantification", a joint work with Manuel Francisco and Fabrizio Sebastiani that was the subject of a recent journal publication [19]. In this paper the authors systematically investigate (by focusing on "aggregative" quantification) the case in which the data items can each have zero, one, or several labels at once (a task that had never been investigated before). The conclusion of their study is that methods that try to exploit the stochastic correlations among the classes *both* in the underlying classifier *and* in the aggregation policy, outperform all aggregative methods that either exploit these correlations in one of the two phases only or do not exploit them at all. The authors indeed propose two new methods that do exploit these correlations; an interesting aspect of these two methods is that they can also be used in conjunction with non-aggregative quantification methods.

2.2 The discussion

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 Q&A sessions of the papers, and of allowing the LQ community to discuss, in an unconstrained, informal setting, what the participants perceived to be the "burning", still unresolved issues in this field.

2.2.1 Datasets for evaluating quantification

One of the major issues discussed was the lack of datasets for ideally supporting the evaluation of quantification. Most current datasets (including the one used in the recent LeQua 2022 data challenge, which was specifically devoted to quantification [10]) originated as classification datasets (i.e., are composed of a training set and a test set of individual labelled items), and were turned into quantification datasets by extracting from the test set, according to a certain extraction protocol, a number of test samples meant to cover the entire spectrum of prevalence distributions.

One consequence of this fact is that quantification research has been carried out on too many different datasets (as opposed to a few "standard" datasets) since, in the absence of true quantification datasets, it was easy for each author to pick her/his favourite classification dataset and turn it into a quantification dataset via a suitable extraction protocol; this has led to the fact that results reported in different research papers on quantification are often not comparable with each other.

Another consequence is that, in the above type of experimentation, the testing samples, which are generated by the extraction protocol, may arguably be unrealistic, at least in some cases. For instance, when applying quantification to *land cover mapping* (LCM),[2] the above extraction protocol would involve the generation of a sample by assembling different pixels "extracted" from different images, which is clearly unrealistic. In LCM, a realistic dataset would be composed of some training images and some test images, where each image is a sample of (labelled or unlabelled) items (the pixels); this scenario is unlike those seen in most experimental evaluations of quantification systems, in which (a) the training set is a set of individual labelled items (and not a set of *samples of* individual labelled items, as in LCM), and (b) the test set is a set of *artificially generated* samples (i.e., samples extracted via the protocol) of unlabelled individual items (and not a set of *naturally occurring* such samples, as in LCM).

LCM seems to evoke the so-called *natural prevalence protocol* (NPP – [9, §3.4.1]) for experimentation in quantification, where only naturally occurring samples are tested upon; other applications of quantification, such as monitoring insect populations [7], estimating the prevalence of different species of plankton in sea water samples [14], quantifying the number of damaged cells in biological samples [20], or seabed cover mapping [1], are also characterised by this property. Unfortunately, the use of the NPP tends to be feasible only when true quantification datasets (i.e., datasets in which both training set and test set naturally come as sets of samples) are available, and public datasets of this sort tend to be hard to obtain (note that there are several LCM datasets that are publicly available, but they contain *estimates* of the pixel labels, and not the ground truths that would be necessary for evaluation).

The discussion on this topic was concluded by a collective commitment to look out for candidate datasets that consist of naturally occurring samples. Future editions of the LQ workshop should plan to support this commitment by soliciting data-centric submissions in the call for papers.

[2]Land cover mapping is an application in which, given an aerial photograph of a portion of the Earth, one has to estimate the fractions of pixels that indicate Trees, or Shrubland, or Grassland, or other types of land cover [16].

2.2.2 A LeQua competition in 2024?

A related topic of discussion was a possible follow-up to LeQua 2022, the first data challenge specifically devoted to quantification [10]. Many attendees stated that LeQua 2022 turned out to be very useful because it provided a controlled environment in which different quantification methods could be experimentally compared and because it generated a reference collection on which systems could be (and have been) tested even after the end of the official challenge.

The attendees agreed that a 2nd edition of the challenge would be invaluable in deepening our collective understanding of quantification. The discussion ended with a collective commitment to think of possible formats for this 2nd edition, of possible subtasks of quantification to focus on, of possible conferences to co-locate it with, and of ways to increase participation in the challenge.

3. CONCLUSION

Overall, LQ 2023 was a success, both in terms of participation and, above all, in terms of the livelihood of interaction among the participants that emerged in the Q&A sessions after the individual papers and in the final brainstorming session. The LQ workshop series has proven itself to be an important reference point for researchers who are active in quantification, especially due to the fact that these researchers are scattered across different research communities (statistics, information retrieval, data mining, machine learning, and others) and are thus unlikely to meet in other contexts.

The proceedings of LQ 2023 appear in a self-published form as [5], and are freely available from the LQ 2023 website (`https://lq-2023.github.io/`).

Acknowledgements

The work by the first author has been funded by the Federal Ministry of Education and Research of Germany and the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence. The work by the second author has been funded by MINECO (Ministerio de Economía y Competitividad) and FEDER (Fondo Europeo de Desarrollo Regional), grant PID2019-110742RB-I00 (MINECO/FEDER). The work by the third and fourth authors has been supported by the AI4Media and SoBigData++ projects, funded by the European Commission (Grant 951911 and 871042, respectively) under the H2020 Programme ICT-48-2020, and by the SoBigData.it, FAIR, and QuaDaSh projects funded by the Italian Ministry of University and Research under the NextGenerationEU program. The authors' opinions do not necessarily reflect those of the funding agencies.

4. REFERENCES

[1] O. Beijbom, J. Hoffman, E. Yao, T. Darrell, A. Rodriguez-Ramirez, M. Gonzalez-Rivero, and O. Hoegh-Guldberg. Quantification in-the-wild: Datasets and baselines. CoRR abs/1510.04811 (2015). Presented at the NIPS 2015 Workshop on Transfer and Multi-Task Learning, Montreal, CA, 2015.

[2] M. Bunse. On multi-class extensions of adjusted classify and count. In *Proceedings of the 2nd International Workshop on Learning to Quantify (LQ 2022)*, pages 43–50, Grenoble, IT, 2022.

[3] M. Bunse. Unification of algorithms for quantification and unfolding. In *Proceedings of the Workshop on Machine Learning for Astroparticle Physics and Astronomy*, pages 459–468, Hamburg, DE, 2022.

[4] M. Bunse. Qunfold: Composable quantification and unfolding methods in Python. In *Proceedings of the 3rd International Workshop on Learning to Quantify (LQ 2023)*, pages 1–7, Torino, IT, 2023.

[5] M. Bunse, P. González, A. Moreo, and F. Sebastiani, editors. *Proceedings of the 3rd International Workshop on Learning to Quantify (LQ 2023)*. Torino, IT, 2023.

[6] A. Castaño, J. Alonso, P. González, and J. J. del Coz. An equivalence analysis of binary quantification methods. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI 2023)*, pages 6944–6952, Washington, US, 2023.

[7] Z. Donyavi, A. Serapio, and G. Batista. MC-SQ: A highly accurate ensemble for multi-class quantification. In *Proceedings of the 23rd SIAM International Conference on Data Mining (SDM 2023)*, pages 622–630, Minneapolis, US, 2023.

[8] B. Dussap, G. Blanchard, and B.-E. Chérief-Abdellatif. Label shift quantification with robustness guarantees via distribution feature matching. arXiv:2306.04376 [stat.ML], 2023.

[9] A. Esuli, A. Fabris, A. Moreo, and F. Sebastiani. *Learning to quantify*. Springer Nature, Cham, CH, 2023.

[10] A. Esuli, A. Moreo, F. Sebastiani, and G. Sperduti. A detailed overview of LeQua 2022: Learning to quantify. In *Working Notes of the 13th Conference and Labs of the Evaluation Forum (CLEF 2022)*, Bologna, IT, 2022.

[11] A. Fabris, A. Esuli, A. Moreo, and F. Sebastiani. Measuring fairness under unawareness of sensitive attributes: A quantification-based approach. *Journal of Artificial Intelligence Research*, 76:1117–1180, 2023.

[12] G. Forman. Quantifying trends accurately despite classifier error and class imbalance. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 157–166, Philadelphia, US, 2006.

[13] 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.

[14] P. González, A. Castaño, E. E. Peacock, J. Díez, J. J. Del Coz, and H. M. Sosik. Automatic plankton quantification using deep features. *Journal of Plankton Research*, 41(4):449–463, 2019.

[15] K. Kloos, J. D. Karch, Q. A. Meertens, and M. de Rooij. Continuous Sweep: An improved, binary quantifier. arXiv:2308.08387 [stat.ML], 2023.

[16] P. Latinne, M. Saerens, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: Evidence from a multi-class problem in remote sensing. In *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*, pages 298–305, Williamstown, US, 2001.

[17] A. Maletzke, D. Moreira dos Reis, E. Cherman, and G. Batista. DyS: A framework for mixture models in quantification. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI 2019)*, pages 4552–4560, Honolulu, US, 2019.

[18] 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.

[19] A. Moreo, M. Francisco, and F. Sebastiani. Multi-label quantification. *ACM Transactions on Knowledge Discovery and Data*, 18(1):Article 4, 2023.

[20] L. Sánchez, V. González, E. Alegre, and R. Alaíz. Classification and quantification based on image analysis for sperm samples with uncertain damaged/intact cell proportions. In *Proceedings of the 5th International Conference on Image Analysis and Recognition (ICIAR 2008)*, pages 827–836, Póvoa de Varzim, PT, 2008.

[21] D. Tasche. Invariance assumptions for class distribution estimation. In *Proceedings of the 3rd International Workshop on Learning to Quantify (LQ 2023)*, pages 56–71, Torino, IT, 2023.