Learning to Quantify: Methods and Applications (LQ 2021)

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

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

ABSTRACT

Learning to Quantify (LQ) is the task of training class prevalence estimators via supervised learning. The task of these estimators is to estimate, given an unlabelled set of data items D and a set of classes $C = \{c_1, \ldots, c_{|C|}\}$, the prevalence (i.e., relative frequency) of each class c_i in D. LQ is interesting in all applications of classification in which the final goal is *not* determining which class (or classes) individual unlabelled data items belong to, but estimating the distribution of the unlabelled data items across the classes of interest. Example disciplines whose interest in labelling data items is at the aggregate level (rather than at the individual level) are the social sciences, political science, market research, ecological modelling, and epidemiology. While LQ may in principle be solved by classifying each data item in D and counting how many such items have been labelled with c_i, it has been shown that this "classify and count" (CC) method yields suboptimal quantification accuracy. As a result, quantification is now no longer considered a mere byproduct of classification and has evolved as a task of its own. The goal of this workshop is bringing together all researchers interested in methods, algorithms, and evaluation measures and methodologies for LQ, as well as practitioners interested in their practical application to managing large quantities of data.

KEYWORDS

Quantification; Dataset Shift

ACM Reference Format:
Juan José del Coz, Pablo González, Alejandro Moreo, and Fabrizio Sebastiani. 2021. Learning to Quantify: Methods and Applications (LQ 2021). In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3459637.3482040

1 WORKSHOP THEME AND TOPICS

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; as [12] puts it, 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 – CC) usually leads to suboptimal quantification accuracy; this is a direct consequence of "Vapnik's principle" [22], 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.

A further reason why "classify and count" is suboptimal is the fact that many applicative scenarios suffer from *distribution shift*, the phenomenon according to which the class prevalence values in the training set may substantially differ from the class prevalence values in the unlabelled data that one needs to issue predictions for [20]. The presence of distribution shift means that the well-known IID assumption, on which most learning algorithms for training classifiers are based, does not hold; in turn, this means that CC will perform less than optimally on scenarios that exhibit distribution shift, and that the higher the amount of shift, the worse we can expect CC to perform.

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 [13]. The research community has investigated methods to correct the biased prevalence estimates of general-purpose classifiers [4, 11, 17], supervised learning methods specially tailored to quantification [1, 3, 6, 10, 14], and evaluation measures for quantification [9, 21]. Specific applications of LQ have also been

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11.
https://doi.org/10.1145/3459637.3482040

investigated, such as sentiment quantification [8, 12], quantification in networked environments [19], or quantification for data streams [18]. 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 analysing these data at the aggregate level rather than individually.

The main topics on which contributions have been solicited, and which form the main themes of the workshop, are

- Binary, multiclass, and ordinal LQ
- Supervised algorithms for LQ
- Semi-supervised / transductive LQ
- Deep learning for LQ
- Representation learning for LQ
- LQ and dataset shift
- Evaluation measures for LQ
- Experimental protocols for the evaluation of LQ
- Quantification of streaming data
- Quantifying text by topic and quantifying text by sentiment
- Novel applications of LQ

2 WORKSHOP OBJECTIVES, GOALS, AND EXPECTED OUTCOME

The goal of this workshop is bringing together all researchers interested in methods, algorithms, and evaluation measures for learning to quantify, as well as practitioners interested in their practical application to managing large quantities of data.

Work on learning to quantify has traditionally been published in a scattered way across different areas, e.g., information retrieval [6, 8, 17], data mining [10, 11], machine learning [2, 7], statistics [16], or directly in the areas to which these techniques get applied [5, 12, 15]. A further goal of this workshop is also to provide the first joint forum for quantification research, where researchers coming from these different walks of life can meet (albeit virtually) and share views.

This workshop is the first of its kind, and has never been held before, neither at CIKM nor at any other conference; we thus believe that it will generate durable benefit to the scientific community, and that the papers presented at the workshop will be useful resources for years to come.

The workshop will also be instrumental in generating interest and stimulating participation in the upcoming LeQua 2022 "Lab" (i.e., shared task) on learning to quantify, which is going to take place in 2022 as a satellite event of the CLEF 2022 conference (https://clef2022.clef-initiative.eu/).

ACKNOWLEDGMENTS

The work by AM and FS has been 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. The authors' opinions do not necessarily reflect those of the European Commission. The work by JdC and PG has been funded by MINECO (the Spanish Ministerio de Economía y Competitividad) under the research project PID2019-110742RB-I00.

REFERENCES

[1] Letizia , Anna Monreale, Giulio Rossetti, Fosca Giannotti, Dino Pedreschi, and Fabrizio Sebastiani. 2013. Quantification trees. In *Proceedings of the ICDM 2013*. Dallas, US, 528–536. https://doi.org/10.1109/icdm.2013.122

[2] Rocío Alaíz-Rodríguez, Alicia Guerrero-Curieses, and Jesús Cid-Sueiro. 2011. Class and subclass probability re-estimation to adapt a classifier in the presence of concept drift. *Neurocomputing* 74, 16 (2011), 2614–2623. https://doi.org/10.1016/j.neucom.2011.03.019

[3] José Barranquero, Jorge Díez, and Juan José del Coz. 2015. Quantification-oriented learning based on reliable classifiers. *Pattern Recognition* 48, 2 (2015), 591–604. https://doi.org/10.1016/j.patcog.2014.07.032

[4] Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. 2010. Quantification via probability estimators. In *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM 2010)*. Sydney, AU, 737–742. https://doi.org/10.1109/icdm.2010.75

[5] Dallas Card and Noah A. Smith. 2018. The importance of calibration for estimating proportions from annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2018)*. New Orleans, US, 1636–1646. https://doi.org/10.18653/v1/n18-1148

[6] Giovanni Da San Martino, Wei Gao, and Fabrizio Sebastiani. 2016. Ordinal text quantification. In *Proceedings of the 39th ACM Conference on Research and Development in Information Retrieval (SIGIR 2016)*. Pisa, IT, 937–940. https://doi.org/10.1145/2911451.2914749

[7] Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. 2017. Class-prior estimation for learning from positive and unlabeled data. *Machine Learning* 106, 4 (2017), 463–492. https://doi.org/10.1007/s10994-016-5604-6

[8] Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. 2018. A recurrent neural network for sentiment quantification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)*. Torino, IT, 1775–1778. https://doi.org/10.1145/3269206.3269287

[9] Andrea Esuli and Fabrizio Sebastiani. 2010. Sentiment quantification. *IEEE Intelligent Systems* 25, 4 (2010), 72–75.

[10] Andrea Esuli and Fabrizio Sebastiani. 2015. Optimizing text quantifiers for multivariate loss functions. *ACM Transactions on Knowledge Discovery and Data* 9, 4 (2015), Article 27. https://doi.org/10.1145/2700406

[11] George Forman. 2008. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery* 17, 2 (2008), 164–206. https://doi.org/10.1007/s10618-008-0097-y

[12] Wei Gao and Fabrizio Sebastiani. 2016. From classification to quantification in tweet sentiment analysis. *Social Network Analysis and Mining* 6, 19 (2016), 1–22. https://doi.org/10.1007/s13278-016-0327-z

[13] Pablo González, Alberto Castaño, Nitesh V. Chawla, and Juan José del Coz. 2017. A review on quantification learning. *Comput. Surveys* 50, 5 (2017), 74:1–74:40. https://doi.org/10.1145/3117807

[14] Víctor González-Castro, Rocío Alaiz-Rodríguez, and Enrique Alegre. 2013. Class distribution estimation based on the Hellinger distance. *Information Sciences* 218 (2013), 146–164. https://doi.org/10.1016/j.ins.2012.05.028

[15] Daniel J. Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54, 1 (2010), 229–247. https://doi.org/10.1111/j.1540-5907.2009.00428.x

[16] Gary King and Ying Lu. 2008. Verbal autopsy methods with multiple causes of death. *Statist. Sci.* 23, 1 (2008), 78–91. https://doi.org/10.1214/07-sts247

[17] Roy Levin and Haggai Roitman. 2017. Enhanced probabilistic classify and count methods for multi-label text quantification. In *Proceedings of the 7th ACM International Conference on the Theory of Information Retrieval (ICTIR 2017)*. Amsterdam, NL, 229–232. https://doi.org/10.1145/3121050.3121083

[18] André G. Maletzke, Denis Moreira dos Reis, and Gustavo E. Batista. 2018. Combining instance selection and self-training to improve data stream quantification. *Journal of the Brazilian Computer Society* 24, 12 (2018), 43–48. https://doi.org/10.1186/s13173-018-0076-0

[19] Letizia Milli, Anna Monreale, Giulio Rossetti, Dino Pedreschi, Fosca Giannotti, and Fabrizio Sebastiani. 2015. Quantification in social networks. In *Proceedings of the 2nd IEEE International Conference on Data Science and Advanced Analytics (DSAA 2015)*. Paris, FR. https://doi.org/10.1109/dsaa.2015.7344845

[20] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaíz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern Recognition* 45, 1 (2012), 521–530. https://doi.org/10.1016/j.patcog.2011.06.019

[21] Fabrizio Sebastiani. 2020. Evaluation measures for quantification: An axiomatic approach. *Information Retrieval Journal* 23, 3 (2020), 255–288. https://doi.org/10.1007/s10791-019-09363-y

[22] Vladimir Vapnik. 1998. *Statistical learning theory*. Wiley, New York, US.