# Learning to Quantify: Estimating Class Prevalence via Supervised Learning

Alejandro Moreo and Fabrizio Sebastiani
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, Italy
firstname.lastname@isti.cnr.it

**ACM Reference Format:**
Alejandro Moreo and Fabrizio Sebastiani. 2019. Learning to Quantify: Estimating Class Prevalence via Supervised Learning . In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3331184.3331389

## 1 MOTIVATION

*Quantification* (also known as "supervised prevalence estimation" [2], or "class prior estimation" [7]) is the task of estimating, given a set $\sigma$ of unlabelled items and a set of classes $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$, the relative frequency (or "prevalence") $p(c_i)$ of each class $c_i \in \mathcal{C}$, i.e., the fraction of items in $\sigma$ that belong to $c_i$. When each item belongs to exactly one class, since $0 \leq p(c_i) \leq 1$ and $\sum_{c_i \in \mathcal{C}} p(c_i) = 1$, $p$ is a *distribution* of the items in $\sigma$ across the classes in $\mathcal{C}$ (the *true distribution*), and quantification thus amounts to estimating $p$ (i.e., to computing a *predicted distribution* $\hat{p}$).

Quantification is important in many disciplines (such as e.g., market research, political science, the social sciences, and epidemiology) which usually deal with aggregate (as opposed to individual) data. In these contexts, classifying individual unlabelled instances is usually not a primary goal, while estimating the prevalence of the classes of interest in the data is. For instance, when classifying the tweets about a certain entity (e.g., a political candidate) as displaying either a Positive or a Negative stance towards the entity, we are usually not much interested in the class of a specific tweet: instead, we usually want to know the fraction of these tweets that belong to the class [14].

Quantification may in principle be solved via classification, i.e., by classifying each item in $\sigma$ and counting, for all $c_i \in \mathcal{C}$, how many such items have been labelled with $c_i$. However, it has been shown in a multitude of works (see e.g., [1, 4, 12–14, 17]) that this "classify and count" (CC) method yields suboptimal quantification accuracy. Simply put, the reason of this suboptimality is that most classifiers are optimized for classification accuracy, and not for quantification accuracy. These two notions do not coincide, since the former is, by and large, inversely proportional to the sum $(FP_i + FN_i)$ of the false positives and the false negatives for $c_i$ in the contingency table, while the latter is, by and large, inversely proportional to the absolute difference $|FP_i - FN_i|$ of the two.

One reason why it seems sensible to pursue quantification directly, instead of tackling it via classification, is that classification is a more general task than quantification: after all, a perfect classifier is also a perfect quantifier, while the opposite is not true. A training set might thus contain information sufficient to generate a good quantifier but not a good classifier, which means that performing quantification via "classify and count" might be a suboptimal way of performing quantification. In other words, performing quantification via "classify and count" looks like a violation of "Vapnik's principle" [33], which asserts that

> 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, quantification has come to be no longer considered a mere byproduct of classification, and has evolved as a task of its own, devoted to designing methods and algorithms (see [15] for a survey) that deliver better prevalence estimates than CC.

There are further reasons why quantification is now considered as a task of its own. One such reason is that, since the goal of quantification is different from that of classification, quantification requires evaluation measures different from those used for classification. A second reason is the growing awareness that quantification is going to be more and more important; with the advent of big data, more and more application contexts are going to spring up in which we will simply be happy with analyzing data at the aggregate level and we will not be able to afford analyzing them at the individual level.

## 2 OBJECTIVES, AND RELEVANCE TO IR

The goal of this course is to introduce the audience to the problem of quantification and to its importance, to the main supervised learning techniques that have been proposed for solving it, to the metrics used to evaluate them, and to what appear to be the most promising directions for further research.

The topic of quantification is relevant to the SIGIR community, because when IR researchers apply classification techniques these researchers are often only interested in results at the aggregate level, which means that they should have used quantification techniques


*SIGIR '19, July 21–25, 2019, Paris, France*

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331389

instead. One typical example is sentiment classification in Twitter: almost nobody who engages in this task is interested in individual tweets *per se*. Researchers and practitioners who use classification when they should instead use quantification typically do so because they ignore that there is a difference between the two; one of the main goals of this tutorial is to raise awareness of this difference.

## 3 FORMAT AND DETAILED SCHEDULE

The structure of the lectures is as follows (each section also indicates the main bibliographic material discussed within the section):

(1) Introduction / Motivation
  (a) Solving quantification via "Classify and Count"
  (b) Concept drift and distribution drift [24, 31]
  (c) Vapnik's principle
  (d) The "paradox of quantification"
(2) Applications of quantification in machine learning, data mining, text mining, and NLP [14]
  (a) Sentiment quantification [11]
  (b) Quantification in the social sciences [5]
  (c) Quantification in political science [17]
  (d) Quantification in epidemiology [19]
  (e) Quantification in market research [11]
  (f) Quantification in ecological modelling [3]
(3) Evaluation of quantification algorithms
  (a) Desirable properties for quantification evaluation measures [30]
  (b) Evaluation measures for quantification [30]
  (c) Experimental protocols for evaluating quantification [10]
(4) Supervised learning methods for binary and multiclass quantification
  (a) Aggregative methods based on general-purpose learners [2, 4, 13, 20, 22, 27, 28]
  (b) Aggregative methods based on special-purpose learners [1, 12]
  (c) Non-aggregative methods [16, 17]
(5) Advanced topics
  (a) Ordinal quantification [6, 8]
  (b) Quantification for networked data [23, 32]
  (c) Quantification for data streams [18, 21, 29]
  (d) Cross-lingual quantification [9]
(6) Shared tasks [25, 26]
(7) Conclusions

## REFERENCES


[1] Barranquero, J., Díez, J. and del Coz, J. J. [2015], 'Quantification-oriented learning based on reliable classifiers', *Pattern Recognition* **48**(2), 591–604.

[2] Barranquero, J., González, P., Díez, J. and del Coz, J. J. [2013], 'On the study of nearest neighbor algorithms for prevalence estimation in binary problems', *Pattern Recognition* **46**(2), 472–482.

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

[4] Bella, A., Ferri, C., Hernández-Orallo, J. and Ramírez-Quintana, M. J. [2010], Quantification via probability estimators, *in* 'Proceedings of the 11th IEEE International Conference on Data Mining (ICDM 2010)', Sydney, AU, pp. 737–742.

[5] Ceron, A., Curini, L. and Iacus, S. M. [2016], 'iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content', *Information Sciences* **367/368**, 105—124.

[6] Da San Martino, G., Gao, W. and Sebastiani, F. [2016], Ordinal text quantification, *in* 'Proceedings of the 39th ACM Conference on Research and Development in Information Retrieval (SIGIR 2016)', Pisa, IT, pp. 937–940.

[7] du Plessis, M. C., Niu, G. and Sugiyama, M. [2017], 'Class-prior estimation for learning from positive and unlabeled data', *Machine Learning* **106**(4), 463–492.

[8] Esuli, A. [2016], ISTI-CNR at SemEval-2016 Task 4: Quantification on an ordinal scale, *in* 'Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)', San Diego, US.

[9] Esuli, A., Moreo, A. and Sebastiani, F. [2019], 'Cross-lingual sentiment quantification', *arXiv:1904.07965* .

[10] Esuli, A., Moreo, A., Sebastiani, F. and Trevisan, D. [2019], Evaluation protocols for quantification. Submitted for publication.

[11] Esuli, A. and Sebastiani, F. [2010], 'Sentiment quantification', *IEEE Intelligent Systems* **25**(4), 72–75.

[12] Esuli, A. and Sebastiani, F. [2015], 'Optimizing text quantifiers for multivariate loss functions', *ACM Transactions on Knowledge Discovery and Data* **9**(4), Article 27.

[13] Forman, G. [2008], 'Quantifying counts and costs via classification', *Data Mining and Knowledge Discovery* **17**(2), 164–206.

[14] Gao, W. and Sebastiani, F. [2016], 'From classification to quantification in tweet sentiment analysis', *Social Network Analysis and Mining* **6**(19), 1–22.

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

[16] González-Castro, V., Alaiz-Rodríguez, R. and Alegre, E. [2013], 'Class distribution estimation based on the Hellinger distance', *Information Sciences* **218**, 146–164.

[17] Hopkins, D. J. and King, G. [2010], 'A method of automated nonparametric content analysis for social science', *American Journal of Political Science* **54**(1), 229–247.

[18] Kar, P., Li, S., Narasimhan, H., Chawla, S. and Sebastiani, F. [2016], Online optimization methods for the quantification problem, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)', San Francisco, US, pp. 1625–1634.

[19] King, G. and Lu, Y. [2008], 'Verbal autopsy methods with multiple causes of death', *Statistical Science* **23**(1), 78–91.

[20] Levin, R. and Roitman, H. [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, pp. 229–232.

[21] Maletzke, A. G., Moreira dos Reis, D. and Batista, G. E. [2018], 'Combining instance selection and self-training to improve data stream quantification', *Journal of the Brazilian Computer Society* **24**(12), 43–48.

[22] Milli, L., Monreale, A., Rossetti, G., Giannotti, F., Pedreschi, D. and Sebastiani, F. [2013], Quantification trees, *in* 'Proceedings of the 13th IEEE International Conference on Data Mining (ICDM 2013)', Dallas, US, pp. 528–536.

[23] Milli, L., Monreale, A., Rossetti, G., Pedreschi, D., Giannotti, F. and Sebastiani, F. [2015], Quantification in social networks, *in* 'Proceedings of the 2nd IEEE International Conference on Data Science and Advanced Analytics (DSAA 2015)', Paris, FR.

[24] Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V. and Herrera, F. [2012], 'A unifying view on dataset shift in classification', *Pattern Recognition* **45**(1), 521–530.

[25] Nakov, P., Farra, N. and Rosenthal, S. [2017], SemEval-2017 Task 4: Sentiment analysis in Twitter, *in* 'Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)', Vancouver, CA.

[26] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F. and Stoyanov, V. [2016], SemEval-2016 Task 4: Sentiment analysis in Twitter, *in* 'Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)', San Diego, US, pp. 1–18.

[27] Pérez-Gállego, P., Quevedo, J. R. and del Coz, J. J. [2017], 'Using ensembles for problems with characterizable changes in data distribution: A case study on quantification', *Information Fusion* **34**, 87–100.

[28] Saerens, M., Latinne, P. and Decaestecker, C. [2002], 'Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure', *Neural Computation* **14**(1), 21–41.

[29] Sanya, A., Kumar, P., Kar, P., Chawla, S. and Sebastiani, F. [2018], 'Optimizing non-decomposable measures with deep networks', *Machine Learning* **107**(8-10), 1597–1620.

[30] Sebastiani, F. [2018], 'Evaluation measures for quantification: An axiomatic approach', *arXiv:1809.01991* .

[31] Storkey, A. [2009], When training and test sets are different: Characterizing learning transfer, *in* J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer and N. D. Lawrence, eds, 'Dataset shift in machine learning', The MIT Press, Cambridge, US, pp. 3–28.

[32] Tang, L., Gao, H. and Liu, H. [2010], Network quantification despite biased labels, *in* 'Proceedings of the 8th Workshop on Mining and Learning with Graphs (MLG 2010)', Washington, US, pp. 147–154.

[33] Vapnik, V. [1998], *Statistical Learning Theory*, Wiley, New York, US.