Tutorial: Supervised Learning for Prevalence Estimation

Alejandro Moreo and Fabrizio Sebastiani

Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, Italy
{firstname.lastname}@isti.cnr.it

Abstract. *Quantification* is the task of estimating, given a set σ of unlabelled items and a set of classes $\mathcal{C}$, the relative frequency (or "prevalence") $p(c_i)$ of each class $c_i \in \mathcal{C}$. 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. Quantification may in principle be solved via classification, i.e., by classifying each item in σ 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 that this "classify and count" (CC) method yields suboptimal quantification accuracy, one of the reasons being that most classifiers are optimized for classification accuracy, and not for quantification accuracy. 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 that deliver better prevalence estimates than CC. The goal of this tutorial is to introduce the main supervised learning techniques that have been proposed for solving quantification, the metrics used to evaluate them, and the most promising directions for further research.

1 Motivation

Quantification (also known as "supervised prevalence estimation" [2], or "class prior estimation" [5]) is the task of estimating, given a set σ 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 σ 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 σ 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 [9].

Quantification may in principle be solved via classification, i.e., by classifying each item in σ 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, 3, 7–9, 12]) 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. To see this consider that a binary classifier h_1 for which $FP = 20$ and $FN = 20$ (FP and FN standing for the "false positives" and "false negatives", respectively, that it has generated on a given dataset) is worse than a classifier h_2 for which, on the same test set, $FP = 18$ and $FN = 20$. However, h_1 is intuitively a better binary quantifier than h_2; indeed, h_1 is a perfect quantifier, since FP and FN are equal and thus, when it comes to class frequency estimation, compensate each other, so that the distribution of the test items across the class and its complement is estimated perfectly. In other words, a good quantifier needs to have small *bias* (i.e., needs to distribute its errors as evenly as possible across FP and FN). 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" [21], 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 is no longer considered a mere byproduct of classification, and has evolved as a task of its own, devoted to designing methods (see [10] for a survey) for delivering 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 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 [17]
   (a) Solving quantification via "Classify and Count"
   (b) Concept drift and distribution drift
   (c) Vapnik's principle
   (d) The "paradox of quantification"
2. Applications of quantification in machine learning, data mining, text mining, and NLP [9, 12]
   (a) Sentiment quantification
   (b) Quantification in the social sciences
   (c) Quantification in political science
   (d) Quantification in epidemiology
   (e) Quantification in market research
   (f) Quantification in ecological modelling
3. Evaluation of quantification algorithms [19]
   (a) Desirable properties for quantification evaluation measures
   (b) Evaluation measures for quantification
   (c) Experimental protocols for evaluating quantification
4. Supervised learning methods for binary and multiclass quantification [1, 3, 7, 8, 11, 12, 15, 18]
   (a) Aggregative methods based on general-purpose learners
   (b) Aggregative methods based on special-purpose learners
   (c) Non-aggregative methods
5. Advanced topics [4, 6, 13, 14, 16, 20]
   (a) Ordinal quantification
   (b) Quantification for networked data
   (c) Quantification for data streams
   (d) Cross-lingual quantification
6. Conclusions

References

1. Barranquero, J., Díez, J., del Coz, J.J.: Quantification-oriented learning based on reliable classifiers. Pattern Recognition **48**(2), 591–604 (2015). https://doi.org/10.1016/j.patcog.2014.07.032

2. Barranquero, J., González, P., Díez, J., del Coz, J.J.: On the study of nearest neighbor algorithms for prevalence estimation in binary problems. Pattern Recognition **46**(2), 472–482 (2013).
3. Bella, A., Ferri, C., Hernández-Orallo, J., Ramírez-Quintana, M.J.: Quantification via probability estimators. In: Proceedings of the 11th IEEE International Conference on Data Mining (ICDM 2010). pp. 737–742. Sydney, AU (2010).
4. Da San Martino, G., Gao, W., Sebastiani, F.: Ordinal text quantification. In: Proceedings of the 39th ACM Conference on Research and Development in Information Retrieval (SIGIR 2016). pp. 937–940. Pisa, IT (2016).
5. du Plessis, M.C., Niu, G., Sugiyama, M.: Class-prior estimation for learning from positive and unlabeled data. Machine Learning **106**(4), 463–492 (2017).
6. Esuli, A., Moreo, A., Sebastiani, F.: Cross-lingual sentiment quantification (2019), arXiv:1904.07965
7. Esuli, A., Sebastiani, F.: Optimizing text quantifiers for multivariate loss functions. ACM Transactions on Knowledge Discovery and Data **9**(4), Article 27 (2015).
8. Forman, G.: Quantifying counts and costs via classification. Data Mining and Knowledge Discovery **17**(2), 164–206 (2008).
9. Gao, W., Sebastiani, F.: From classification to quantification in tweet sentiment analysis. Social Network Analysis and Mining **6**(19), 1–22 (2016).
10. González, P., Castaño, A., Chawla, N.V., del Coz, J.J.: A review on quantification learning. ACM Computing Surveys **50**(5), 74:1–74:40 (2017).
11. González-Castro, V., Alaiz-Rodríguez, R., Alegre, E.: Class distribution estimation based on the Hellinger distance. Information Sciences **218**, 146–164 (2013).
12. Hopkins, D.J., King, G.: A method of automated nonparametric content analysis for social science. American Journal of Political Science **54**(1), 229–247 (2010).
13. Kar, P., Li, S., Narasimhan, H., Chawla, S., Sebastiani, F.: Online optimization methods for the quantification problem. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016). pp. 1625–1634. San Francisco, US (2016).
14. Maletzke, A.G., Moreira dos Reis, D., Batista, G.E.: Combining instance selection and self-training to improve data stream quantification. Journal of the Brazilian Computer Society **24**(12), 43–48 (2018).
15. Milli, L., Monreale, A., Rossetti, G., Giannotti, F., Pedreschi, D., Sebastiani, F.: Quantification trees. In: Proceedings of the 13th IEEE International Conference on Data Mining (ICDM 2013). pp. 528–536. Dallas, US (2013).
16. Milli, L., Monreale, A., Rossetti, G., Pedreschi, D., Giannotti, F., Sebastiani, F.: Quantification in social networks. In: Proceedings of the 2nd IEEE International Conference on Data Science and Advanced Analytics (DSAA 2015). Paris, FR (2015).
17. Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V., Herrera, F.: A unifying view on dataset shift in classification. Pattern Recognition **45**(1), 521–530 (2012).
18. Saerens, M., Latinne, P., Decaestecker, C.: Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. Neural Computation **14**(1), 21–41 (2002).
19. Sebastiani, F.: Evaluation measures for quantification: An axiomatic approach. Information Retrieval Journal (2019), to appear
20. Tang, L., Gao, H., Liu, H.: Network quantification despite biased labels. In: Proceedings of the 8th Workshop on Mining and Learning with Graphs (MLG 2010). pp. 147–154. Washington, US (2010).
21. Vapnik, V.: Statistical Learning Theory. Wiley, New York, US (1998)