Title: Quantification (machine learning)

URL: https://en.wikipedia.org/wiki/Quantification_(machine_learning)

PageID: 69310432

Categories: Category:Machine learning

In machine learning and data mining , quantification (variously called learning to quantify , or supervised prevalence estimation , or class prior estimation) is the task of using supervised learning in order to train models (quantifiers) that estimate the relative frequencies (also known as prevalence values) of the classes of interest in a sample of unlabelled data items . [1] [2] For instance, in a sample of 100,000 unlabelled tweets known to express opinions about a certain political candidate, a quantifier may be used to estimate the percentage of these tweets which belong to class `Positive' (i.e., which manifest a positive stance towards this candidate), and to do the same for classes `Neutral' and `Negative'.

Quantification may also be viewed as the task of training predictors that estimate a (discrete) probability distribution , i.e., that generate a predicted distribution that approximates the unknown true distribution of the items across the classes of interest. Quantification is different from classification , since the goal of classification is to predict the class labels of individual data items, while the goal of quantification it to predict the class prevalence values of sets of data items. Quantification is also different from regression , since in regression the training data items have real-valued labels, while in quantification the training data items have class labels.

It has been shown in multiple research works [3] [4] [5] [6] [7] that performing quantification by classifying all unlabelled instances and then counting the instances that have been attributed to each class (the 'classify and count' method) usually leads to suboptimal quantification accuracy. This suboptimality may be seen as a direct consequence of ' Vapnik 's principle', 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. [8]

In our case, the problem to be solved directly is quantification, while the more general intermediate problem is classification. As a result of the suboptimality of the 'classify and count' method, quantification has evolved as a task in its own right, different (in goals, methods, techniques, and evaluation measures) from classification.

Quantification tasks

The main variants of quantification, according to the characteristics of the set of classes used, are:

Binary quantification, corresponding to the case in which there are only $n = 2$ {\displaystyle n=2} classes and each data item belongs to exactly one of them;

Single-label multiclass quantification, corresponding to the case in which there are $n > 2$ {\displaystyle n>2} classes and each data item belongs to exactly one of them;

Multi-label multiclass quantification, corresponding to the case in which there are $n \geq 2$ {\displaystyle n\geq 2} classes and each data item can belong to zero, one, or several classes at the same time;

Ordinal quantification, corresponding to the single-label multiclass case in which a total order is defined on the set of classes.

Regression quantification, a task which stands to 'standard' quantification as regression stands to classification. Strictly speaking, this task is not a quantification task as defined above (since the individual items do not have class labels but are labelled by real values), but has enough

commonalities with other quantification tasks to be considered one of them.

Most known quantification methods address the binary case or the single-label multiclass case, and only few of them address the multi-label, ordinal, and regression cases.

Binary-only methods include the Mixture Model (MM) method, [3] the HDy method, [9] SVM(KLD), [6] and SVM(Q). [5]

Methods that can deal with both the binary case and the single-label multiclass case include probabilistic classify and count (PCC), [4] adjusted classify and count (ACC), [3] probabilistic adjusted classify and count (PACC), [4] and the Saerens-Latinne-Decaestecker EM -based method (SLD). [10]

Methods for multi-label quantification include regression-based quantification (RQ) and label powerset-based quantification (LPQ). [11]

Methods for the ordinal case include Ordinal Quantification Tree (OQT), [12] and ordinal versions of the above-mentioned ACC, PACC, and SLD methods. [13]

Methods for the regression case include Regress and splice and Adjusted regress and sum . [14]

Evaluation measures for quantification

Several evaluation measures can be used for evaluating the error of a quantification method. Since quantification consists of generating a predicted probability distribution that estimates a true probability distribution, these evaluation measures are ones that compare two probability distributions. Most evaluation measures for quantification belong to the class of divergences . Evaluation measures for binary quantification and single-label multiclass quantification are [15]

Absolute Error

Squared Error

Relative Absolute Error

Kullback–Leibler divergence

Pearson Divergence

Evaluation measures for ordinal quantification are

Normalized Match Distance (a particular case of the Earth Mover's Distance)

Root Normalized Order-Aware Distance

Applications

Quantification is of special interest in fields such as the social sciences , [16] epidemiology , [17] market research , and ecological modelling , [18] since these fields are inherently concerned with aggregate data. However, quantification is also useful as a building block for solving other downstream tasks, such as improving the accuracy of classifiers on out-of-distribution data, [10] [19] allocating resources , [3] measuring classifier bias , [20] and estimating the accuracy of classifiers on out-of-distribution data. [21]

Resources

LQ 2021: the 1st International Workshop on Learning to Quantify [22]

LQ 2022: the 2nd International Workshop on Learning to Quantify [23]

LQ 2023: the 3rd International Workshop on Learning to Quantify [24]

LQ 2024: the 4th International Workshop on Learning to Quantify [25]

LeQua 2022: the 1st Data Challenge on Learning to Quantify [26]

LeQua 2024: the 2nd Data Challenge on Learning to Quantify [27]

QuaPy: An open-source Python-based software library for quantification [28]

References