D*y*S: A Framework for Mixture Models in Quantification

André Maletzke, Denis dos Reis, Everton Cherman, Gustavo Batista
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
{andregustavo,denismr,evertoncherman}@usp.br, gbatista@icmc.usp.br

Abstract

Quantification is an expanding research topic in Machine Learning literature. While in classification we are interested in obtaining the class of individual observations, in quantification we want to estimate the total number of instances that belong to each class. This subtle difference allows the development of several algorithms that incur smaller and more consistent errors than counting the classes issued by a classifier. Among such new quantification methods, one particular family stands out due to its accuracy, simplicity, and ability to operate with imbalanced training samples: Mixture Models (MM). Despite these desirable traits, MM, as a class of algorithms, lacks a more in-depth understanding concerning the influence of internal parameters on its performance. In this paper, we generalize MM with a base framework called D*y*S: Distribution y-Similarity. With this framework, we perform a thorough evaluation of the most critical design decisions of MM models. For instance, we assess 15 dissimilarity functions to compare histograms with varying numbers of bins from 2 to 110 and, for the first time, make a connection between quantification accuracy and test sample size, with experiments covering 24 public benchmark datasets. We conclude that, when tuned, Topsøe is the histogram distance function that consistently leads to smaller quantification errors and, therefore, is recommended to general use, notwithstanding Hellinger Distance's popularity. To rid MM models of the dependency on a choice for the number of histogram bins, we introduce two dissimilarity functions that can operate directly on observations. We show that SORD, one of such measures, presents performance that is slightly inferior to the tuned Topsøe, while not requiring the sensible parameterization of the number of bins.

Introduction

Quantification is a task that is similar to classification in the sense that we are provided with a training set with labeled observations, and a test sample with unlabeled examples. However, in quantification, our objective is to predict the proportion of instances that belong to each class, rather than the label of each observation. The literature has proposed several methods that yield smaller quantification errors than simply classifying individual examples and counting the issued classes. González et al. (2017a) present a comprehensive survey of the area. A relevant finding is the fact that the errors committed by quantification methods are more consistent than those obtained with classifying and counting, since the absolute quantification error of the latter grows linearly around a predicted proportion for which the quantification error is zero (Forman 2008).

Better estimations for the proportion of the classes are important for applications where the interest is in analyzing tendencies and behaviors of groups of individuals rather than specific classifications. Examples are quality control for seminal material (González-Castro, Alaiz-Rodríguez, and Alegre 2013), estimation of insect population in a region (dos Reis et al. 2018a), and sentiment analysis in social media (Esuli, Sebastiani, and Abbasi 2010).

Among the quantification methods found in literature, a family of algorithms stands out due to its quantification accuracy, simplicity, and ability to work on imbalanced training samples: Mixture Models (MM) (Forman 2005).

MM methods constitute a family of quantification approaches where the probability distribution of each class is modeled individually and learned from a training set. As a test sample contains data from all classes at different proportions, its distribution is a parametric mixture of the classes' individual distributions, where the parameters are the proportions of the classes. Hence, the MM methods search for the parameters of such a mixture and, consequently, estimate the proportions of the classes in the test sample.

MM methods commonly represent the distribution of the classes using histograms as an approximation of the discretized Probability Density Function (PDF) (González-Castro, Alaiz-Rodríguez, and Alegre 2013). The values stored in such histograms are scores provided by classifiers and relate to the probability of each observation belonging to the positive class. This approach has three main advantages. First, memory compactness, since histograms summarize the original data as multiple examples are aggregated into a single histogram bin. Second, easiness and low computational cost for mixing distributions, as we use vectors to represent histograms and these vectors can be interpolated inexpensively. Finally, simplicity and low computational cost for comparing distributions, since we can use any dissimilarity function that operates on a vectorial space.

However, there are significant disadvantages that need to be addressed. While histograms with a small number of bins incur a more compact vision of the original data, they also incur in greater information loss. The lost information could otherwise be required to differentiate similar distributions. On the other hand, as MM methods typically resort to dissimilarity functions that compare aligned bins in isolation, as the Hellinger Distance (González-Castro, Alaiz-Rodríguez, and Alegre 2013), histograms with many bins incur spatial sparseness and demand more training and test observations to reasonably estimate the distributions. Finally, several distinct dissimilarity functions can be employed, although there is no consensus regarding which one is the most adequate.

In this paper, we formalize a framework that generalizes the Mixture Model approach for quantification. We name this framework DyS: Distribution y-Similarity. We make an extensive evaluation to rank the most suitable dissimilarity functions so that future work makes the best use of DyS. We empirically analyze the influence of the number of bins on quantification accuracy and provide recommendations for this parameter.

Finally, we introduce two dissimilarity functions that are compatible with the general framework proposed by DyS, even though they operate directly over observations rather than histograms. By using one of such distances, we are not required to summarize the original data and therefore lose information. Such distances can potentially better differentiate similar distributions while being immune to the curse of dimensionality. Our results show that Hellinger Distance used in HDy (González-Castro, Alaiz-Rodríguez, and Alegre 2013), a state-of-the-art MM approach, is outperformed by other measures. This finding is even more evident when we tune the number of histogram bins. Moreover, we propose a parameter-free distance for quantification that provides smaller quantification errors than all tested histogram distances when using the previous standard number of bins, as suggested by (González-Castro, Alaiz-Rodríguez, and Alegre 2013), and remains competitive even after this parameter is tuned.

Related Work

Quantification is a supervised Data Mining task that shares several similarities with classification. Both require the same feature-based representation for observations and a nominal output attribute describing the individual classes.

At first glance, classifying and counting seems to be a practical solution for quantification. However, research papers have shown that such a method generally produces poor quantification performance (Forman 2006; González-Castro, Alaiz-Rodríguez, and Alegre 2013; Gao and Sebastiani 2016; González et al. 2017b), and systematically under or overestimates the classes proportion as the test set class distribution changes (Forman 2008). These facts led the community to the proposal of several quantification methods. Due to lack of space, we refer readers to (González et al. 2017a) for a comprehensive survey on the subject and (Bekker and Davis 2018) for newer approaches that rely only on labeled positive observations.

Our purpose remains concentrated particularly in a class of methods known as Mixture Models – MM. MM methods represent the probability distribution of each class separately and model the test set distribution as a mixture of the classes' individual distributions, where the parameters are the proportions of the classes. Hence, the MM methods search for the parameters of such a mixture and, consequently, find the proportions of the classes in the test sample.

However, datasets are generally multidimensional. Accurately estimating a multidimensional probability distribution requires an increasing number of observations as the number of dimensions goes up, since increasing the number of dimensions also increases sparseness. Thus, using all original dimensions from the dataset raises not only the cost of data acquisition for training but also sets an elevated minimum size for test samples. Literature provides a simple way to avoid both undesirable characteristics: the indirect use of a scorer that maps the observations from the feature-space to $\mathbb{R}$ (González-Castro, Alaiz-Rodríguez, and Alegre 2013; dos Reis et al. 2018b; 2018a).

Generally speaking, a scorer outputs a number that is proportional to the probability of an observation belonging to the positive class, and is often an integral part of several classifiers, as Naïve Bayes and Support Vector Machines. Additionally, any classifier can be turned into a scorer when they are part of an ensemble. For example, Random Forests produces a score that is the proportion of the votes favoring the positive class. Unbiased scores can be individually obtained for the positive and negative class through k-fold cross-validation within the training set: for each k validation portion of the data, a scorer is induced with all $k-1$ other parts. Scores are obtained by applying such a scorer on the validation portion, and the scores given to positive and negative observations are kept apart (Forman 2005; González-Castro, Alaiz-Rodríguez, and Alegre 2013; Pérez-Gállego et al. 2019).

Scores are individual observations from a data distribution, and we need to estimate and represent such a distribution. A simple way of expressing a distribution that is also convenient for enabling a straightforward mixture of different distributions is the discretized Probability Density Function (PDF) (González-Castro, Alaiz-Rodríguez, and Alegre 2013). It consists of the aggregation of scores into normalized histograms with b bins, so that the sum of all bins equals one. These histograms can then be treated as vectors in the $\mathbb{R}^b$, and pairs of histograms can be mixed at varying degrees by performing a linear interpolation. This mixing approach means that individual observations are weighted instead of discarded, although they lose detail after being categorized into bins. Finally, any dissimilarity function that operates on a vector space can be applied to compare pairs of distributions. An equivalent approach for representing distributions that is out of the scope of this paper is to use the Cumulative Distribution Function (CDF) instead of PDF. Forman (2005) used this representation to propose the earliest MM method for quantification.

More recently, González-Castro, Alaiz-Rodríguez, and Alegre (2013) proposed the HDy algorithm. The method uses two normalized histograms (the normalization causes the sum of the bins to be one), P^+ and P^- that summarize the samples of scores S^+ and S^-. Such samples are obtained from two (possibly cross-validated) validation sets with exclusively positive and exclusively negative observations, respectively. When presented with an unlabeled test sample, the algorithm builds a histogram Q with the set of scores Z obtained by the same scorer on such a sample. These histograms, P^+, P^-, and Q, represent the distributions of the training set for each class and the distribution of the test sample, respectively. Finally, given the histograms P^+, P^-, and Q the $\mathrm{HDy}(P^+, P^-, Q)$ estimates the positive proportion rate as

$$\mathrm{HDy}(P^+, P^-, Q) = \\ \underset{0 \leq \alpha \leq 1}{\arg\min} \left\{ \mathrm{HD}\left(\alpha P^+ + (1-\alpha)P^-, Q\right) \right\} \quad (1)$$

where HD is the Hellinger Distance (Pollard 2002), and each histogram, with b bins, is represented as a vector in the $\mathbb{R}^b$. Hellinger Distance is a function that estimates the similarity between two probability distributions P and Q, where $P = \alpha P^+ + (1-\alpha)P^-$. The HDy authors use different numbers of bins from 10 to 110, with increments of 10, and the final proportion of positive labels in the test sample is the median of these 11 estimates. To estimate the value of α inside the algorithm, the authors of the original paper perform a linear search within the range $[0, 1]$.

Research papers have extended or adapted the HDy algorithm for building, for example, ensembles models for quantification (Pérez-Gállego et al. 2019), context identification methods (dos Reis et al. 2018a), and concept drift detection approaches (Maletzke et al. 2018). These achievements were obtained by promoting slight changes to the HDy setup. Additionally, dos Reis et al. (2018a) show that estimating α through Ternary Search makes HDy more efficient and generally more precise than through linear search, since local minima are very close to the global minimum. In this paper, we argue that HDy consists of an instance of a more general algorithm that remains informal with relevant parameters to be evaluated.

D***y***S

We propose DyS, a generic framework for quantification based on the similarity of score distributions. Equation 2 formalizes our proposal. We note that DyS is a generalization of HDy.

$$\mathrm{D}y\mathrm{S}(S^+, S^-, Z) = \\ \underset{0 \leq \alpha \leq 1}{\arg\min} \left\{ \mathrm{DS}\left(\alpha H(S^+) + (1-\alpha)H(S^-), H(Z)\right) \right\} \quad (2)$$

where DS is a dissimilarity function, and S^+, S^-, Z are positive training sample, negative training sample, and test sample, respectively, and H is a function that converts a sample of scores into a representation that is compatible with DS and that supports mixing two distributions according to a factor α. For all tested histogram distances, H produces a histogram from the given sample, so that we obtain P^+ from S^+, P^- from S^- and Q from Z. As we discuss later in this paper, for the proposed dissimilarity functions SORD and MKS, $H(x) = x$, since the proposed methods operate directly over samples of scores rather than binned histograms. Moreover, while $\alpha S^+ + (1-\alpha)S^-$ is a simplified notation of the interpolation process that mixes two samples, it is accurate only for histograms. How the mixture is performed in SORD and KS is better described later.

Histogram Dissimilarity

We analyze the impact of switching the dissimilarity function across a plurality of measures from the literature that is suitable for comparing histograms.

In this section, we list all compared functions, and point the interested reader to (Cha 2008) for a thorough survey on these measures. In our work, we have selected the following dissimilarity functions: Squared Euclidean (SE), Manhattan (MH), Probabilistic Symmetric (PS), Topsøe (TS), Jensen Difference (JD), Taneja (TN), Hellinger (HD), Dice (DC), Jaccard (JC), Chebyshev (CB), Inner Product (IP), Kumar-Hassebrook (HB), Cosine (CS), and Harmonic Mean (HM). All employed histogram distances, except for ORD, are described in Table 1, where P and Q are two normalized histograms of same length b.

Table 1: Dissimilarity Functions.

SE	$\sum_{i=1}^{b}(P_i - Q_i)^2$
MH	$\sum_{i=1}^{b} \lvert P_i - Q_i \rvert$
PS	$2\sum_{i=1}^{b} \frac{(P_i - Q_i)^2}{P_i + Q_i}$
TS	$\sum_{i=1}^{b} \left(P_i ln\left(\frac{2P_i}{P_i+Q_i}\right) + Q_i ln\left(\frac{2Q_i}{P_i+Q_i}\right) \right)$
JD	$\sum_{i=1}^{b} \left[\frac{P_i lnP_i + Q_i lnQ_i}{2} - \left(\frac{P_i+Q_i}{2}\right) ln\left(\frac{P_i+Q_i}{2}\right) \right]$
TN	$\sum_{i=1}^{b} \left(\frac{P_i+Q_i}{2}\right) ln\left(\frac{P_i+Q_i}{2\sqrt{P_i Q_i}}\right)$
HD	$2\sqrt{1 - \sum_{i=1}^{b}\sqrt{P_i Q_i}}$
DC	$\frac{\sum_{i=1}^{b}(P_i - Q_i)^2}{\sum_{i=1}^{b} P_i^2 + \sum_{i=1}^{b} Q_i^2}$
JC	$\frac{\sum_{i=1}^{b}(P_i - Q_i)^2}{\sum_{i=1}^{b} P_i^2 + \sum_{i=1}^{b} Q_i^2 - \sum_{i=1}^{b} P_i Q_i}$
CB	$\max_i \lvert P_i - Q_i \rvert$
IP	$P \bullet Q \sum_{i=1}^{b} P_i Q_i$
HB	$\frac{\sum_{i=1}^{b} P_i Q_i}{\sum_{i=1}^{b} P_i^2 + \sum_{i=1}^{b} Q_i^2 - \sum_{i=1}^{b} P_i Q_i}$
CS	$\frac{\sum_{i=1}^{b} P_i Q_i}{\sqrt{\sum_{i=1}^{b} P_i^2}\sqrt{\sum_{i=1}^{b} Q_i^2}}$
HM	$2\sum_{i=1}^{b} \frac{P_i Q_i}{P_i + Q_i}$

Although Cha (2008) surveys a larger set of distances, we decided to exclude from our analysis all monotonic transformations of these distances since they are deemed to produce the same quantification estimates. This effect happens as, in DyS, we only search for the mixing parameter that produces the lowest dissimilarity value and disregard the value itself. Monotonic transformations do not change the order of the values, and therefore we should find the same parameters. We also eliminated common but asymmetric functions so that we need not impose an arbitrary order between the mixed training distribution and the test distribution.

Apart from the dissimilarity functions listed in Table 1, we also tested the use of ORD (Cha and Srihari 2002). To better understand such a distance, we note that in its original description, the histograms are normalized by multiplying every bin in P by $|Q|$ and every bin in Q by $|P|$. As such, each histogram is the allocation of $|P| \times |Q|$ units into b bins. The objective of ORD is to find the least number of movements that need be done to transform Q into P, where one movement is the transference of a unit from a bin to a *neighbor* bin. There is always a possible transformation since both normalized histograms have the same number of units. This setting happens to be a univariate case of the *minimum difference of pair assignment* (MDPA) of two distributions, which is a special case of the Earth Mover's Distance (EMD) (Rubner, Tomasi, and Guibas 1998). For this particular case, the proposers of ORD introduce a greedy algorithm that can compute the distance in $O(b)$, whereas the algorithms that are used to solve generic instances of EMD have higher time complexity. We note that the proposed algorithm also works when the histograms are normalized to have sum one, instead of $|P| \times |Q|$. The Algorithm 1 describes the rationale of such distance.

Algorithm 1: Ordinal Distance

```
Data: Histograms to be compared $P$ and $Q$
Result: Dissimilarity between $P$ and $Q$
begin
    diffsum $\longleftarrow 0$ ;
    total_cost $\longleftarrow 0$ ;
    for $i \leftarrow 1$ to $length(P)$ do
        diffsum $\longleftarrow$ diffsum $+ (P[i] - Q[i])$ ;
        total_cost $\longleftarrow$ total_cost $+ |$diffsum$|$ ;
    end
    return total_cost ;
end
```

All histogram distances tested in our experiments, except for ORD, use each bin position in isolation. In other words, for each bin in one histogram, only one bin in the other histogram can affect the distance. For this reason, we believe that ORD is less susceptible to the curse of dimensionality and bad parameterization of the number of bins, when it is greater than the ideal. For all other distances, as we increase the number of bins and consequently their granularity, non-zeroed bins are more likely to contain fewer observations and be countered by zeroed bins in the opposing histogram. In fact, when the number of bins is infinity, only observations that are identical in both samples affect the distance and help decrease it. The distance is always one (the maximum value) when there are no identical observations. ORD, on the other hand, is less affected by an increasing number of bins, since the difference between mismatching bins is accounted for and any pair of bins (and non-identical observations) can affect the distance.

We introduce two dissimilarity functions that are suitable for DyS and do not rely on histograms, as they are able to operate directly on the observations from training and test samples. Such distances also carry the same benefits from the histograms approach, allowing for mixing pairs of distributions to be compared with a third distribution. However, they have the additional benefits of being immune to the curse of dimensionality while not losing information since they do not simplify the original data. We explain these distances in the following sections.

Mixable Kolmogorov Smirnov

The first distance is the Mixable Kolmogorov Smirnov (MKS) statistic. It is an adaptation of the Kolmogorov Smirnov (KS) (Kolmogorov 1933) statistic to compare two discrete empirical distributions and accounts for the first distribution being a weighted mixture of two distributions.

Equation 3 formalizes such a dissimilarity, where S^+ and S^- are the two samples that will be mixed together, according to the weights α and $1 - \alpha$, respectively, and Z is the third sample that represents the distribution to be compared with.

$$D_{MKS}(S^+, S^-, \alpha, Z) = \sup_x |\alpha F_{S^+}(x) + (1-\alpha) F_{S^-}(x) - F_Z(x)| \quad (3)$$

where $F_Y(x)$ is the proportion of the observations in Y that are lower or equal than x.

SORD

The second dissimilarity function is the Sample ORD (SORD). SORD can be viewed as a special case of ORD where the number of bins is infinity: while ORD is the minimum cost necessary to transform a histogram into another one, SORD is the minimum cost to transform a sample into another one.

If we are using SORD simply as a measurement of the difference between two samples S and Z, every observation x is weighted as $w(x)$, defined as follows.

$$w(x) := \begin{cases} |S|^{-1}, & x \in S \\ |Z|^{-1}, & x \in Z \end{cases} \quad (4)$$

This way, both samples share the same total weight and the transformation is feasible. The cost of transforming a fraction $f_{i,j}, 0 \leq f_{i,j} \leq 1$, of the i-th observation in S into the j-th observation in Z is $c(i,j) = |f_{i,j} w(S_i) S_i - w(Z_j) Z_j|$. The objective of SORD is therefore the following optimization problem.

$$\begin{aligned} & \underset{f_{i,j} \forall i,j}{\text{minimize}} \sum_i^{|S|} \sum_j^{|Z|} c(i,j) \\ & \text{subject to} \sum_j^{|Z|} f_{i,j} = 1 \quad \forall i \\ & \qquad w(Z_j) Z_j = w(S_i) S_i \sum_i^{|S|} f_{i,j} \quad \forall j \end{aligned} \quad (5)$$

For the purpose of quantification, one of the samples compared by SORD is a mixture of two other samples: one that contains positive training observations and another with negative training observations. This mixed sample is compared to a test sample. In this scenario, we have to adjust the weights of the observations in the mixed sample: positive observations share the same weight, proportional to α, and the negative ones share the same weight, inversely proportional to α.

SORD can be efficiently computed in $O(|S \cup Z| \log |S \cup Z|)$ with a greedy approach, where S and Z are the two samples being compared. Algorithm 2 fully describes the distance computation with the necessary change to the weights when S is a mixture (with parameter α) of two samples (S^+ and S^-).

Algorithm 2: SORD Dissimilarity Function

```
Data: Mixing samples $S^+, S^-$, mixing factor $\alpha$,
      comparing sample $Z$
Result: Dissimilarity between $\alpha S^+ + (1-\alpha)S^-$ and $Z$
begin
    $w'(x) := \begin{cases} \alpha|S^+|^{-1}, & x \in S^+ \\ (1-\alpha)|S^-|^{-1}, & x \in S^- \\ -|Z|^{-1}, & x \in Z \end{cases}$ ;
    $v \longleftarrow$ sorted array with $\forall\, x \in S^+ \cup S^- \cup Z$;
    acc $\longleftarrow w'(v[1])$ ;
    total_cost $\longleftarrow 0$ ;
    for $i \leftarrow 2$ to length($v$) do
        $\delta \longleftarrow v[i] - v[i-1]$ ;
        total_cost $\longleftarrow$ total_cost $+ |\delta \times$ acc$|$ ;
        acc $\longleftarrow$ acc $+ w'(v[i])$ ;
    end
    return total_cost ;
end
```

Experimental Setup

In this paper, we make a comprehensible experimental evaluation divided into two parts.

First, we hypothesize the existence of a relationship between the size of the test sample and the number of histogram bins that lead to the smallest error. The original HDy paper reports the estimated distribution based on the median across the use of varying number of bins from 10 to 110, with increments of 10. While this particular range may provide good quantification errors when the test sample has a large number of observations, we want to verify if the more general DyS, which includes HDy, is significantly influenced by the number of bins, and how.

We note that until now, although the ideal number of bins has not been studied, a decision for this parameter may not be completely uninformed and can be based on important insights. Histograms with too many bins are negatively affected by two aspects. The first aspect is that if the sample size is not large enough, large histograms can become too sparse, each bin can have excessively low weight, and ultimately, the dissimilarity function can face the curse of dimensionality. Exception for this rule is the use of ORD. Notably, we note that ORD avoids being affected by such sparseness since the relation between neighbor dimensions is considered, rather than each dimension contributing in isolation to the magnitude of the distance. The second aspect is that a large number of bins has the implicit assumption of high precision for the scores. On the other hand, if there are too few bins, we may be unable to differentiate distributions.

To verify the impact of the number of the bins, in all experiments, we vary the number of bins from 2 to 20 with increments of 2, and from 20 to 110 with increments of 10. The test sample size, on the other hand, varies from 10 to 100 with increments of 10 examples, and from 100 to 500 with increments of 100 examples.

Once we figure a satisfactory range for the number of bins for each dissimilarity function, we proceed to the second part of our evaluation. We consider a satisfactory range to be one that minimizes the largest number of bins necessary to obtain the smallest quantification errors for at least 95% of the cases. We analyze the impact of using different histogram distances in the DyS framework for binary quantification. With a fixed range of bins for each distance, we rank them according to the median quantification error produced by their use so that the top-ranked distances are those which lead to the smallest errors in each one of our experiments. We vary the test sample size from 10 to 500 in the aforementioned way. We analyze the behavior of the ranks for each dissimilarity function with a box plot.

For all experiments, we vary the positive class proportion from 0% to 100% with increments of 1%, and for each proportion, we execute 10 runs with different test samples.

We performed preliminary experiments and concluded that Ternary Search (TSearch) suits all tested dissimilarity functions. For this reason, it is used for all of our experiments. We note that the α that minimizes Squared Euclidean distance can be algebraically deduced in $O(1)$. However, we also use TSearch for this distance to maintain experimental consistency across distances.

In the next section, we enumerate and describe the datasets used in our experiments.

Datasets

Each dataset was uniformly split into two halves: training and test. With the training half, we performed 10-fold cross-validation to obtain the training scores used by DyS. The full training half was also used to train a single scorer that was applied on the test half to get a test score set. Test samples were sampled from the test score set according to the settings described in the previous section regarding class proportion and size. One observation does not appear more than once in a single test sample, although it can appear in more than one test sample. This procedure was used to make the best use of our limited data.

As the time complexity of both SORD and MKS grows linearithmic for the total number of observations involved (including training and test), we undersampled training scores to up to $1,000$ per class, when using these distances. Despite such reduction, we believe this number of observation scores can still carry more information than a histogram.

We produced all scores using Random Forests with 200 trees. Also, we assess the performance of the quantifiers in our experiments using the Mean Absolute Error (MAE) (Sebastiani 2018). MAE is the average of absolute differences between true (p) and predicted ($\hat{p}$) quantifications for a set of classes C, as shown in Equation 6.

$$MAE(p, \hat{p}) = \frac{1}{|C|} \sum_{c \in C} |\hat{p}(c) - p(c)| \tag{6}$$

Table 2 presents a brief description of the datasets used in our experiments obtained from UCI (Dheeru and Karra Taniskidou 2017), OpenML (Vanschoren et al. 2013), PROMISE (Sayyad Shirabad and Menzies 2005), and Reis (dos Reis et al. 2018a) repositories. Specific citations are requested for Bank Marketing (Moro, Cortez, and Rita 2014), Credit Card (Yeh and Lien 2009), HTRU2 (Lyon et al. 2016), Mozilla4 (Koru, Zhang, and Liu 2007), Mushroom (Lincoff 1989), Nomao (Candillier and Lemaire 2012), and Occupancy Detection (Candanedo and Feldheim 2016). Additionally, we note that Jock A. Blackard and Colorado State University preserve copyright over Covertype.

Table 2: Datasets description.

Dataset	Size	Features	Repository
Anuran Calls	6,585	22	UCI
Bank Marketing	45,211	16	UCI
BNG (vote)	39,366	9	OpenML
Click Prediction	39,948	11	OpenML
CMC	1,473	9	UCI
Covertype-reduced	8,715	54	UCI
Credit Card	30,000	23	UCI
EEG Eye State	14,980	14	OpenML
HTRU2	17,898	8	UCI
JM1	10,880	21	PROMISE
Letter Recognition	20,000	16	UCI
MAGIC Gamma	19,020	10	UCI
Mozilla4	15,545	5	OpenML
Mushroom	8,124	22	OpenML
Nomao	34,465	118	OpenML
Occupancy Detection	20,560	5	UCI
Phoneme	5,404	5	OpenML
Spambase	4,601	57	UCI
Wine Type	6,497	12	UCI
AedesSex	24,000	27	Reis
AedesQuinx	24,000	27	Reis
ArabicDigit	8,800	27	UCI
Handwritten-QG	4,014	63	Reis
Wine Quality	6,497	12	UCI

Three observations about the datasets are due. First, Wine Type dataset is similar to Wine Quality. However, we want to differentiate between white and red wines, rather than the wine quality. Second, Covertype-reduced is a stratified sample from Covertype due to its considerable size. The abbreviated version is still large enough for our purposes. ArabicDigit is a preprocessed version of the original so that all examples have the same number of features (dos Reis et al. 2018a), and the objective is to predict the sex of the speaker rather than which digit is spoken. Finally, all the described datasets represent binary classification problems.

Experimental Evaluation

In this section, we present and discuss our experimental results. One of our main questions regards a possible relationship between the number of bins and test sample size, to achieve the smallest quantification error. We are interested in knowing whether there is a test sample size for which it is better to use more or fewer bins than for another sample size, for the same distance. In Figure 1, we illustrate the Mean Absolute Error (MAE) across datasets while varying the sample size, using the distance function Topsøe, as it represents the general behavior of other distances.

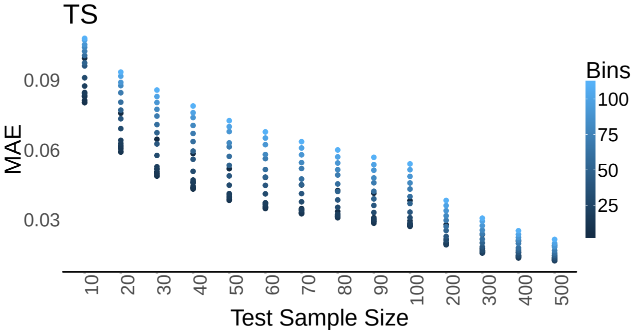

Figure 1: Mean absolute quantification error averaged for all datasets, obtained with DyS with Topsøe and varying test sample size and number of histogram bins.

We can form three observations. First, the error is lower for more significant test samples, which is expected, since we are provided with more information about the test distribution. Second, for the Topsøe distance function, a smaller number of bins generally leads to smaller errors across all assessed test sample sizes. Third, greater sample sizes are less negatively impacted by a higher number of bins. This is explained by the lower sparseness of the bins with more observations in a higher dimensionality.

Both observations hold for all tested distances, except Cosine, Harmonic Mean, Kumar-Hassebrook, Inner Product, and ORD. For the first four distances, errors are smaller for datasets with fewer observations, and a lower number of bins led to more significant errors. However, such distances performed very poorly: all of them led to errors greater than 70% on average, which is worse than a baseline that always predicts a positive class ratio of 50% and, consequently, obtains a maximum error of 50%. For this reason, Cosine, Harmonic Mean, Kumar-Hassebrook, and Inner Product will not be considered from now on.

On the other hand, ORD performed as well as the other distances while being almost invariant to the number of bins. This can be explained because each dimension is not used in isolation, which makes it less affected by sparseness. However, similarly to other distances, error decayed as test sample size increased. Figure 2 illustrates this finding. Our supplemental material website[1] contains figures for all other distances, which were omitted in this paper due to space constraints.

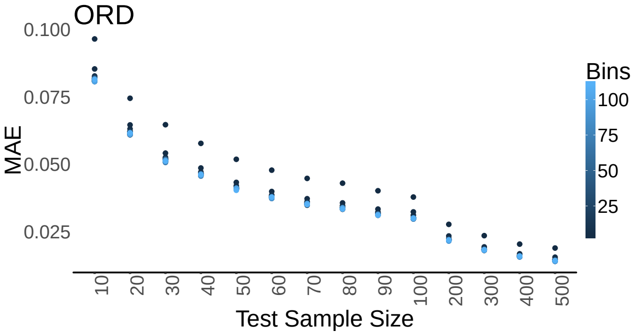

Figure 2: Mean absolute quantification error averaged for all datasets, obtained with DyS with ORD and varying test sample size and number of histogram bins.

In Figure 3, we observe that most of the best quantification results were obtained within up to 20 bins for all considered distances, except for ORD. In Table 3, we detail this finding: we present the smallest upper limit for the number of bins that was necessary to constrain 90%, 95% and 100% of the smallest quantification errors obtained for each distance. Hellinger Distance produced 95% of its best quantification results within the range from 2 to 14 bins. This finding contradicts the arbitrary range of 10 to 110 bins used by HDy's original authors to calculate the median positive class ratio.

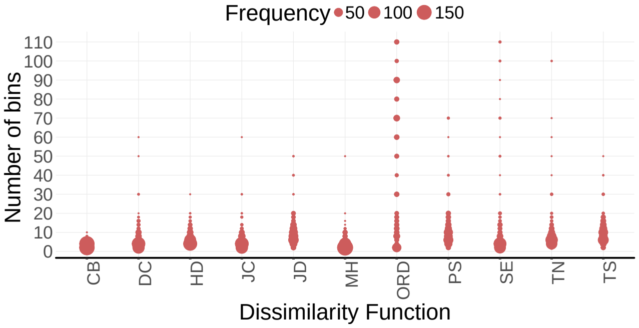

Figure 3: Frequencies of the number of bins that produced the smallest quantification error for each distance function, in DyS.

To rank the distances by quantification error, for each setting, we considered the median of the positive class ratio obtained with DyS while varying the number of bins. The number of bins ranges from 2 to the number of bins that are necessary to constrain 95% of the best results produced by the distance that was used, according to Table 3. The rankings are presented in Figure 4.

Table 3: Smallest upper bound for the number of histogram bins that encloses 90%, 95% and 100% of the smallest quantification errors produced by each distance function, in DyS.

Distance	90%	95%	100%
CB	4	4	10
DC	10	14	60
HD	12	14	30
JC	10	12	60
JD	16	20	50
MH	8	10	50
ORD	90	100	110
PS	18	30	70
SE	16	40	110
TN	14	18	100
TS	16	18	50

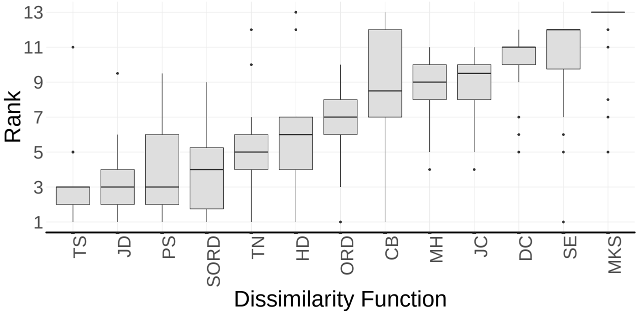

Figure 4: Aggregation of several rank positions for different distances in DyS. Each quantification was predicted as a median of estimates obtained for different numbers of histogram bins. The range of bins was individually tuned for each distance. Test sample size varied from 10 to 500.

We note that each distance had its range of bins tuned individually with the same datasets that were used for this comparison. Exceptions are SORD and MKS, which do not make use of this optimized parameter. This inserts bias into the comparison. On the other hand, if we had used the range from 10 to 110 bins, with increments of 10, as suggested by HDy's authors, we would obtain the ranking presented in Figure 5. In this scenario, the top five best distances are the same. We note that ORD jumps to first place as it is less affected by large histograms, and Topsøe's error increases.

For the remaining of our analysis, we consider the tuned range of bins again. We varied the test sample size from 10 to 100 with increments of 10 and from 100 to 500 with increments of 100. This difference of granularity provides small sample sizes (less than 100 observations) more weight than all bigger sizes combined since there are more test cases for the former case. We can see in Figure 6 that SORD performs only slightly worse than Topsøe for all sample sizes with a similar variance. However, we must keep in mind that SORD is a parameter-free algorithm, i.e., its results were obtained without previous tuning. On the other hand, the Topsøe results are a consequence of a time-consuming process aimed

[1] https://sites.google.com/site/andregustavom/research/dys

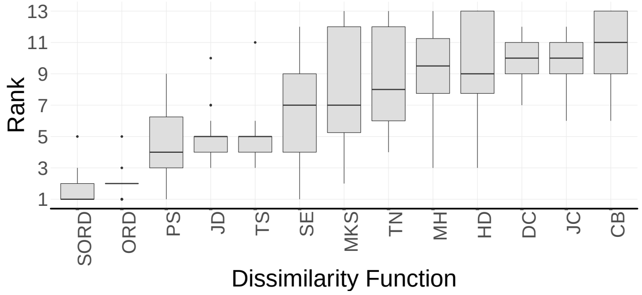

Figure 5: Aggregation of several rank positions for different distances in D*y*S. Each quantification was predicted as a median of estimates obtained for different numbers of histogram bins. The range of bins was [10,110] for all distances. Test sample size varied from 10 to 500.

at tuning the number of bins to be used on the same datasets. Finally, we also potentially limited the best performance that SORD could achieve since we limited the size of the samples, due to the algorithm's higher computational cost in comparison with the histogram distances.

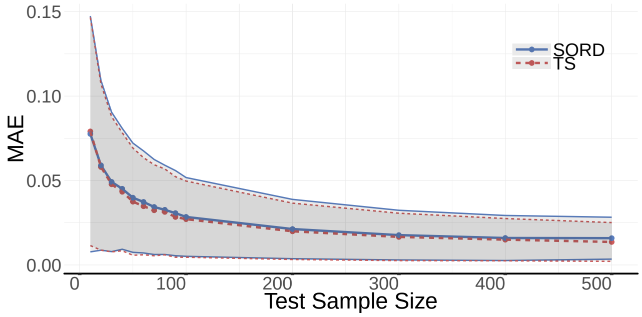

Figure 6: Comparison between SORD and Topsøe for varying test sample size. The shaded area corresponds to the standard deviations from the measured points, and thinner curves set the limit of the shaded areas.

Additionally, ORD also performs closely to Topsøe, after tuning, although not as close as SORD. This is true even though ORD fell from the second to the seventh rank position after the tuning process. However, this change of rank is due to an increase in the absolute performance of Topsøe, rather than a change in the performance of ORD. The latter is mostly unaffected by the rise in the number of bins (after a minimum at which the different distributions become discernible).

Conclusions and Future Work

In this paper, we introduced D*y*S, a framework of Mixture Models for quantification. We analyzed the use of several histogram distances and concluded that Topsøe offers the smallest quantification errors across several datasets and test sample sizes when we tune the number of histogram bins.

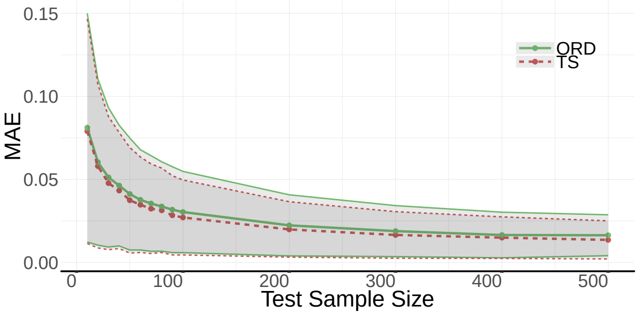

Figure 7: Comparison between ORD and Topsøe for varying test sample size. The shaded area corresponds to the standard deviations from the measured points, and thinner curves set the limit of the shaded areas.

We experimentally found that the best range for the number of bins in D*y*S varies for each distance function. While ORD is mostly unimpaired by an incorrect setting this parameter, for the majority of the distance functions, a suitable superior limit was below 20. Particularly, histograms with 14 or fewer bins provide at least 95% of the best quantification results when using Hellinger Distance. This finding opposes the arbitrary range from 10 to 110 bins used by HDy's original paper. Finally, we introduced a new dissimilarity function, SORD, that operates over observations rather than histograms, while still being compatible with the framework provided by D*y*S.

SORD outperforms all distances when they do not have their parameters tuned. On the other hand, when we tune the parameters, our parameter-free algorithm is outperformed by the Topsøe, Probabilistic Symmetric, and Jensen Difference dissimilarity functions, respectively. However, even then, SORD presents better results than the HDy, the function currently used by the state-of-art MMs, and is only slightly outperformed by Topsøe. ORD falls a little behind SORD, and even behind HDy. However, we argue that its performance is still competitive to Topsøe's in practical terms and, as it is mostly unaffected by a wrong parameterization of the number of bins. ORD is a viable and more time-efficient alternative to SORD when the tuning process cannot be done or is unreliable.

In future work, we plan on evaluating the impact of varying quality of scores on mixture models for quantification. The score quality is related to the difficulty of a dataset for classification. We plan on evaluating whether the ideal number of bins for histogram distances vary according to the quality of the scores. Additionally, we intend on assessing situations where there is a mismatch between the quality of scores produced for the training and the test. This situation can happen as a result of concept drift and incur circumstances where test samples are easier or harder to classify than the training set. Finally, we plan on extending all comparisons to other families of learning algorithms, to provide better positioning for Mixture Models inside Quantification literature as a whole.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code PROEX-6909543/D, the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant 306631/2016-4), the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP, grants 2016/04986-6, 2017/22896-7), and the United States Agency for International Development (USAID, grant AID-OAA-F-16-00072).

References

Bekker, J., and Davis, J. 2018. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence*.

Candanedo, L. M., and Feldheim, V. 2016. Accurate occupancy detection of an office room from light, temperature, humidity and co2 measurements using statistical learning models. *Energy and Buildings* 112:28–39.

Candillier, L., and Lemaire, V. 2012. Design and analysis of the nomao challenge active learning in the real-world. In *Proceedings of the ALRA: Active Learning in Real-world Applications, Workshop ECML-PKDD*.

Cha, S.-H., and Srihari, S. N. 2002. On measuring the distance between histograms. *Pattern Recognition* 35(6):1355–1370.

Cha, S.-H. 2008. Taxonomy of nominal type histogram distance measures. *City* 1(2):1.

Dheeru, D., and Karra Taniskidou, E. 2017. UCI machine learning repository.

dos Reis, D.; Maletzke, A.; Silva, D. F.; and Batista, G. E. A. P. A. 2018a. Classifying and counting with recurrent contexts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, 1983–1992. ACM.

dos Reis, D. M.; Maletzke, A. G.; Cherman, E.; and Batista, G. E. 2018b. One-class quantification. In *Proceedings of the European Conference on Machine Learning*, 564–575.

Esuli, A.; Sebastiani, F.; and Abbasi, A. 2010. Sentiment quantification. *IEEE Intelligent Systems* 25(4):72–79.

Forman, G. 2005. Counting positives accurately despite inaccurate classification. In *European Conference on Machine Learning*, 564–575. Springer.

Forman, G. 2006. 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 '06, 157–166. ACM.

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

Gao, W., and Sebastiani, F. 2016. From classification to quantification in tweet sentiment analysis. *Social Network Analysis and Mining* 6(1).

González, P.; Castaño, A.; Chawla, N. V.; and Coz, J. J. D. 2017a. A review on quantification learning. *ACM Computing Surveys (CSUR)* 50(5):74.

González, P.; Díez, J.; Chawla, N.; and del Coz, J. J. 2017b. Why is quantification an interesting learning problem? *Progress in Artificial Intelligence* 6(1):53–58.

González-Castro, V.; Alaiz-Rodríguez, R.; and Alegre, E. 2013. Class distribution estimation based on the hellinger distance. *Information Sciences* 218:146 – 164.

Kolmogorov, A. 1933. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.* 4:83–91.

Koru, A. G.; Zhang, D.; and Liu, H. 2007. Modeling the effect of size on defect proneness for open-source software. In *Predictor Models in Software Engineering, 2007. PROMISE'07: ICSE Workshops 2007. International Workshop on*, 10–10. IEEE.

Lincoff, G. H. 1989. The audubon society field guide to North American mushrooms. Technical Report No. 635.8 L5.

Lyon, R. J.; Stappers, B.; Cooper, S.; Brooke, J.; and Knowles, J. 2016. Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society* 459(1):1104–1123.

Maletzke, A.; dos Reis, D.; Cherman, E.; and Batista, G. 2018. On the need of class ratio insensitive drift tests for data streams. In *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 94 of *Proceedings of Machine Learning Research*, 110–124. ECML-PKDD, Dublin, Ireland: PMLR.

Moro, S.; Cortez, P.; and Rita, P. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62:22–31.

Pérez-Gállego, P.; Castaño, A.; Quevedo, J. R.; and del Coz, J. J. 2019. Dynamic ensemble selection for quantification tasks. *Information Fusion* 45:1–15.

Pollard, D. 2002. *A user's guide to measure theoretic probability*, volume 8. Cambridge University Press.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 1998. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*, 59–66. IEEE.

Sayyad Shirabad, J., and Menzies, T. 2005. The PROMISE repository of software engineering databases. School of Information Technology and Engineering, University of Ottawa, Canada.

Sebastiani, F. 2018. Evaluation measures for quantification: An axiomatic approach. *arXiv preprint arXiv:1809.01991*.

Vanschoren, J.; van Rijn, J. N.; Bischl, B.; and Torgo, L. 2013. Openml: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter* 15(2):49–60.

Yeh, I.-C., and Lien, C.-h. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications* 36(2):2473–2480.