Neurocomputing 151 (2015) 259–267

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Robustness and generalization for metric learning

Aurélien Bellet [a,1,*], Amaury Habrard [b]

[a] Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA
[b] Laboratoire Hubert Curien UMR 5516, Université de Saint-Etienne, 18 rue Benoit Lauras, 42000 St-Etienne, France

ARTICLE INFO

Article history:
Received 10 February 2014
Received in revised form
18 July 2014
Accepted 24 September 2014
Communicated by M. Bianchini
Available online 7 October 2014

Keywords:
Metric learning
Algorithmic robustness
Generalization bounds

ABSTRACT

Metric learning has attracted a lot of interest over the last decade, but the generalization ability of such methods has not been thoroughly studied. In this paper, we introduce an adaptation of the notion of algorithmic robustness (previously introduced by Xu and Mannor) that can be used to derive generalization bounds for metric learning. We further show that a weak notion of robustness is in fact a necessary and sufficient condition for a metric learning algorithm to generalize. To illustrate the applicability of the proposed framework, we derive generalization results for a large family of existing metric learning algorithms, including some sparse formulations that are not covered by the previous results.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Metric learning consists in automatically adjusting a distance or similarity function using training examples. The resulting metric is tailored to the problem of interest and can lead to dramatic improvement in classification, clustering or ranking performance. For this reason, metric learning has attracted a lot of interest for the past decade (see [1,2] for recent surveys). Existing approaches rely on the principle that pairs of examples with the same (resp. different) labels should be close to each other (resp. far away) under a good metric. Learning thus generally consists in finding the best parameters of the metric function given a set of labeled pairs.[2] Many methods focus on learning a Mahalanobis distance, which is parameterized by a positive semi-definite (PSD) matrix and can be seen as finding a linear projection of the data to a space where the Euclidean distance performs well on the training pairs (see for instance [3–9]). More flexible metrics have also been considered, such as similarity functions without PSD constraint [10–12]. The resulting distance or similarity is used to improve the performance of a metric-based algorithm such as k-nearest neighbors [5,7], linear separators [12,13], K-Means clustering [3] or ranking [9].

Despite the practical success of metric learning, little work has gone into a formal analysis of the generalization ability of the resulting metrics on unseen data. The main reason for this lack of results is that metric learning violates the common assumption of independent and identically distributed (IID) data. Indeed, the training pairs are generally given by an expert and/or extracted from a sample of individual instances, by considering all possible pairs or only a subset based for instance on the nearest or farthest neighbors of each example, some criterion of diversity [14] or a random sample. Online learning algorithms [15,6,10] can still offer some guarantees in this setting, but only in the form of regret bounds assessing the deviation between the cumulative loss suffered by the online algorithm and the loss induced by the best hypothesis that can be chosen in hindsight. These may be converted into proper generalization bounds under restrictive assumptions [16]. Apart from these results on online metric learning, very few papers have looked at the generalization ability of batch methods. The approach of Bian and Tao [17,18] uses a statistical analysis to give generalization guarantees for loss minimization approaches, but their results rely on restrictive assumptions on the distribution of the examples and do not take into account any regularization on the metric. Jin et al. [19] adapted the framework of uniform stability [20] to regularized metric learning. However, their approach is based on a Frobenius norm regularizer and cannot be applied to other type of regularization, in particular sparsity-inducing norms [21] that are used in many recent metric learning approaches [22,8,23,9]. Independently and in parallel to our work,[3] Cao et al. [25] proposed a framework based on Rademacher analysis, which is general but rather complex and limited to pair constraints.

In this paper, we propose to study the generalization ability of metric learning algorithms according to a notion of *algorithmic*

* Corresponding author.
E-mail addresses: bellet@usc.edu (A. Bellet), amaury.habrard@univ-st-etienne.fr (A. Habrard).

[1] Most of the work in this paper was carried out while the author was affiliated with Laboratoire Hubert Curien UMR 5516, Université de Saint-Etienne, France.

[2] Some methods use triplets (x, y, z) such that x should be closer to y than to z, where x and y share the same label, but not z.

[3] We posted a preliminary version of the present work on arXiv in 2012 [24].

http://dx.doi.org/10.1016/j.neucom.2014.09.044
0925-2312/© 2014 Elsevier B.V. All rights reserved.

robustness. This framework, introduced by Xu et al. [26,27], allows one to derive generalization bounds when the variation in the loss associated with two "close" training and testing examples is bounded. The notion of closeness relies on a partition of the input space into different regions such that two examples in the same region are considered close. Robustness has been successfully used to derive generalization bounds in the classic supervised learning setting, with results for SVM, LASSO, etc. We propose here to adapt algorithmic robustness to metric learning. We show that, in this context, the problem of non-IIDness of the training pairs/triplets can be worked around by simply assuming that they are built from an IID sample of labeled examples. Moreover, following [27], we provide a notion of weak robustness that is necessary and sufficient for metric learning algorithms to generalize well, confirming that robustness is a fundamental property. We illustrate the applicability of the proposed framework by deriving generalization bounds, using very few approach-specific arguments, for a family of problems that is larger than what is considered in previous work [17–19,25]. In particular, results apply to a vast choice of regularizers, without any assumption on the distribution of the examples and using a simple proof technique.

The rest of the paper is organized as follows. We introduce some preliminaries and notations in Section 2. Our notion of algorithmic robustness for metric learning is presented in Section 3. The necessity and sufficiency of weak robustness is shown in Section 4. Section 5 illustrates the wide applicability of our framework by deriving bounds for existing metric learning formulations. Section 6 discusses the merits and limitations of the proposed analysis compared to related work, and we conclude in Section 7.

2. Preliminaries

2.1. Notations

Let X be the instance space, Y be a finite label set and let $\mathcal{Z} = X \times Y$. In the following, $z = (x, y) \in \mathcal{Z}$ means $x \in X$ and $y \in Y$. Let μ be an unknown probability distribution over $\mathcal{Z}$. We assume that X is a compact convex metric space w.r.t. a norm $\|\cdot\|$ such that $X \subset \mathbb{R}^d$, thus there exists a constant R such that $\forall x \in X,\ \|x\| \leq R$. A similarity or distance function is a pairwise function $f : X \times X \to \mathbb{R}$. In the following, we use the generic term *metric* to refer to either a similarity or a distance function. We denote by $\mathbf{s}$ a labeled training sample consisting of n training instances $(s_1, \ldots, s_n)$ drawn IID from μ. The sample of all possible pairs built from $\mathbf{s}$ is denoted by $p_{\mathbf{s}}$ such that $p_{\mathbf{s}} = \{(s_1, s_1), \ldots, (s_1, s_n), \ldots, (s_n, s_n)\}$. A metric learning algorithm $\mathcal{A}$ takes as input a finite set of pairs from $(\mathcal{Z} \times \mathcal{Z})^n$ and outputs a metric. We denote by $\mathcal{A}_{p_{\mathbf{s}}}$ the metric learned by an algorithm $\mathcal{A}$ from a sample $p_{\mathbf{s}}$ of pairs. For any pair of labeled examples (z, z') and any metric f, we associate a loss function $l(f, z, z')$ which depends on the examples and their labels. This loss is assumed to be non-negative and uniformly bounded by a constant B. We define the generalization loss (or true loss) over μ as

$$\mathcal{L}(f) = \mathbb{E}_{z,z' \sim \mu} l(f, z, z'),$$

and the empirical loss over the sample $p_{\mathbf{s}}$ as

$$l_{emp}(f) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} l(f, s_i, s_j) = \frac{1}{n^2} \sum_{(s_i, s_j) \in p_{\mathbf{s}}} l(f, s_i, s_j).$$

We are interested in bounding the deviation between $l_{emp}(f)$ and $\mathcal{L}(f)$.

2.2. Algorithmic robustness in classic supervised learning

The notion of algorithmic robustness, introduced by Xu and Mannor [26,27] in the context of classic supervised learning, is based on the deviation between the loss associated with two training and testing instances that are "close". Formally, an algorithm is said $(K, \epsilon(\mathbf{s}))$-robust if there exists a partition of the space $\mathcal{Z} = X \times Y$ into K disjoint subsets such that for every training and testing instances belonging to the same region of the partition, the variation in their associated loss is bounded by a term $\epsilon(\mathbf{s})$. From this definition, the authors have proved a bound for the difference between the empirical loss and the true loss that has the form

$$\epsilon(\mathbf{s}) + B\sqrt{\frac{2K \ln 2 + 2 \ln 1/\delta}{n}}, \tag{1}$$

with probability $1 - \delta$. This bound depends on K and $\epsilon(\mathbf{s})$. The latter should tend to zero as K increases to ensure that (1) also goes to zero when $n \to \infty$.[4] When considering metric spaces, the partition of $\mathcal{Z}$ can be obtained by the notion of covering number [28].

Definition 1. For a metric space (X, ρ), and $T \subset X$, we say that $\hat{T} \subset T$ is a γ-cover of T, if $\forall t \in T,\ \exists \hat{t} \in \hat{T}$ such that $\rho(t, t') \leq \gamma$. The γ-covering number of T is

$$\mathcal{N}(\gamma, T, \rho) = \min\{|\hat{T}| : \hat{T} \text{ is a } \gamma - \text{ cover of } T\}.$$

When X is a compact convex space, for any $\gamma > 0$, the quantity $\mathcal{N}(\gamma, X, \rho)$ is finite leading to a finite cover. If we consider the space $\mathcal{Z}$, note that the label set can be partitioned into $|Y|$ sets. Thus, $\mathcal{Z}$ can be partitioned into $|Y|\mathcal{N}(\gamma, X, \rho)$ subsets such that if two instances $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2)$ belong to the same subset, then $y_1 = y_1$ and $\rho(x_1, x_2) \leq \gamma$.

3. Robustness and generalization for metric learning

We present here our adaptation of robustness to metric learning. The idea is to use the partition of $\mathcal{Z}$ at the pair level: if a new test pair of examples is close to a training pair, then the loss value for each pair must be close. Two pairs are close when each instance of the first pair fall into the same subset of the partition of $\mathcal{Z}$ as the corresponding instance of the other pair, as shown in Fig. 1. A metric learning algorithm with this property is said robust. This notion is formalized as follows.

Definition 2. An algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$ robust for $K \in \mathbb{N}$ and $\epsilon(\cdot) : (\mathcal{Z} \times \mathcal{Z})^n \to \mathbb{R}$ if $\mathcal{Z}$ can be partitioned into K disjoints sets, denoted by $\{C_i\}_{i=1}^{K}$, such that for all sample $\mathbf{s} \in \mathcal{Z}^n$ and the pair set $p(\mathbf{s})$ associated to this sample, the following holds:

$$\forall (s_1, s_2) \in p(\mathbf{s}), \quad \forall z_1, z_2 \in \mathcal{Z}, \quad \forall i,\ j = 1, \ldots, K:$$

$$\text{if } s_1, z_1 \in C_i \text{ and } s_2, z_2 \in C_j \text{ then}$$

$$|l(\mathcal{A}_{p_{\mathbf{s}}}, s_1, s_2) - l(\mathcal{A}_{p_{\mathbf{s}}}, z_1, z_2)| \leq \epsilon(p_{\mathbf{s}}). \tag{2}$$

K and $\epsilon(\cdot)$ quantify the robustness of the algorithm and depend on the training sample. The property of robustness is required for every training pair of the sample; we will later see that this property can be relaxed.

Note that this definition of robustness can be easily extended to triplet based metric learning algorithms. Instead of considering all the pairs $p_{\mathbf{s}}$ from an IID sample $\mathbf{s}$, we take the admissible triplet set $trip_{\mathbf{s}}$ of $\mathbf{s}$ such that $(s_1, s_2, s_3) \in trip_{\mathbf{s}}$ means s_1 and s_2 share the same label while s_1 and s_3 have different ones, with the interpretation that s_1 must be more similar to s_2 than to s_3. The robustness property can then be expressed as follows: $\forall (s_1, s_2, s_3) \in trip_{\mathbf{s}}, \forall z_1,$

[4] This point will be made clear by the examples provided in Section 5.

Download English Version:

https://daneshyari.com/en/article/409712

Download Persian Version:

https://daneshyari.com/article/409712

Daneshyari.com