Information and Computation 256 (2017) 372–380

Contents lists available at ScienceDirect

Information and Computation

www.elsevier.com/locate/yinco

Classification based on prototypes with spheres of influence

Martin Anthony [a,*], Joel Ratsaby [b]

[a] Department of Mathematics, The London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK
[b] Electrical and Electronics Engineering Department, Ariel University, Ariel 40700, Israel

ARTICLE INFO

Article history:
Received 10 December 2014
Received in revised form 16 January 2017
Available online 30 August 2017

Keywords:
Classification
Learning
Generalisation error

ABSTRACT

We present a family of binary classifiers and analyse their performance. Each classifier is determined by a set of 'prototypes', with given labels. The classification of a given point is determined through the sign of a discriminant function. For each prototype, its sphere of influence is the largest sphere centred on it that contains no prototypes of opposite label, and, given a point to be classified, there is a contribution to the discriminant function at that point from precisely those prototypes whose spheres of influence contain the point. This contribution is positive from positive prototypes and negative from negative prototypes. These contributions are larger in absolute value the closer the point is (relative to the sphere's radius) to the prototype. We quantify the generalization error of such classifiers in a standard probabilistic learning model which involves the values of the discriminant function on the points of a random training sample.

1. Introduction

Learning Vector Quantization (LVQ) and its various extensions introduced by Kohonen [14] are used successfully in many machine learning tools and applications. Learning pattern classification by LVQ is based on adapting a fixed set of labeled prototypes in Euclidean space and using the resulting set of prototypes in a nearest-prototype rule (winner-take-all) to classify any point in the input space. LVQ fails if the Euclidean representation is not well-suited for the data. To that end, several extensions of the LVQ algorithm exist which use a weighted Euclidean metric [11] that take advantage of samples for which a more confident (or a large margin) classification can be obtained. Generalization error bounds with dependence on this sample margin are stated in [11,18] and, as is usually the case for large-margin learning [1], the bounds are tighter than ones with no sample-margin dependence. The results of such work are important as they explain why LVQ works well in practice in Euclidean metric spaces.

In the world of big data, which deals with a rich variety of learning domains, there is a huge potential in doing prototype-based learning over non-Euclidean spaces. In this paper we present a family of binary classifiers for learning on any metric input space. We analyse their performance and present generalization learning error bounds that are sample-dependent and hence take advantage of samples that can be classified with a large margin. Each classifier is determined by a set of 'prototypes', whose classifications are given; and the classification of any other point depends on the classifications of the prototypes to which it is sufficiently close, and on how close it is to these prototypes. Thus, in contrast to the above-mentioned works, here a classifier's decision is not based only on the nearest prototype. In many domains of application, data can no longer simply be considered to be in Euclidean space. As has been pointed out in [12], data can take diverse

* Corresponding author.
E-mail addresses: m.anthony@lse.ac.uk (M. Anthony), ratsaby@ariel.ac.il (J. Ratsaby).

http://dx.doi.org/10.1016/j.ic.2017.08.004
0890-5401/

forms in areas such as linguistics and bioinformatics. For this reason, an approach that analyses data in a general metric space (such as that taken here) might be more useful.

More precisely, the classification of a given point is determined through the sign of a discriminant function. For each prototype, its sphere of influence is defined to be the largest sphere centred on it that contains no prototypes of opposite label. Given a point to be classified, there is a contribution to the discriminant function at that point from precisely those prototypes whose spheres of influence contain the point, this contribution being positive from positive prototypes and negative from negative prototypes. These contributions are larger in absolute value the closer the point is (relative to the sphere's radius) to the prototype. We quantify the generalization error of such classifiers in a standard probabilistic learning model, and we do so in a way that involves the values of the discriminant function on the points of the random training sample.

We note in passing that the idea of a sphere of influence is not new. In fact, RCE networks [15] have a hidden layer of activation units associated with a spherical decision region in the input space. There are some differences between our classifier and the RCE. RCE is essentially a classifier whose decision regions are union of spheres, which may not cover all of the input space and hence the classifier can in some cases reject making a decision. The radii of the spheres are parameters to be learnt. Learning RCE involves adapting the size of the radii in an incremental manner in response to whether sample instances are included or not in spheres that are associated with a mismatching class label. New spherical units, that is, prototypes, can also be added when sample points are not covered and not classified. In contrast to RCE, our classifier is non-parametric and the region of influence of each prototype, in resemblance to Voronoi cells in the nearest-neighbour classifier [7], is determined directly from the sample without any parameter such as a radius. The classifier's definition is intentionally left very general in that the set of prototypes can be *any* set of k points, in particular a subset of the sample, and can be determined via *any* algorithm. The error bounds that we state in the paper apply regardless of the algorithm that is used to learn these prototypes.

2. Classifiers based on spheres of influence

The classifiers we consider are binary classifiers defined on a metric space $\mathcal{X}$; so, they are functions $h : \mathcal{X} \to \{-1, 1\}$. We shall assume that $\mathcal{X}$ is of finite diameter with respect to the metric d and, for the sake of simplicity, that its diameter is 1. (The analysis can easily be modified for any other value of the diameter.) Each classifier we consider is defined by a set of *labeled prototypes*. More precisely, a typical classifier is defined by a finite set Π^+ of *positive prototypes* and a disjoint set Π^- of *negative prototypes*, with Π^+ and Π^- both being subsets of $\mathcal{X}$. The idea is that the correct classifications of the points in Π^+ (Π^-, respectively) are $+1$ (-1). We define the *sphere of influence* of each prototype as follows. Suppose $p \in \Pi^+$ and let

$$r(p) = \min\{d(p, p^-) : p^- \in \Pi^-\},$$

the distance to the closest oppositely-labeled prototype; and define $r(p)$ analogously in the case where $p \in \Pi^-$. Then the open ball $B_{r(p)}(p) = B_{r(p)}(p; d)$, of radius $r(p)$ and centred on p, is the sphere of influence of p. Suppose that $\Pi = \Pi^+ \cup \Pi^- = \{p_1, p_2, \ldots, p_k\}$, where $\Pi^+ = \{p_1, \ldots, p_t\}$ and $\Pi^- = \{p_{t+1}, \ldots, p_k\}$, and let r_i denote $r(p_i)$ where $0 < r(p_i) \le 1$. For $x \in \mathcal{X}$, let

$$\phi_i(x) = 1 - \frac{d(x, p_i)}{r_i}$$

and let

$$s_i(x) = [\phi_i(x)]_+ ,$$

where, for $z \in \mathbb{R}$, $[z]_+ = z$ if $z \ge 0$ and $[z]_+ = 0$ otherwise. Define the 'discriminant' function $f_\Pi : \mathcal{X} \to \mathbb{R}$ as follows:

$$f_\Pi(x) = \sum_{i=1}^{t} s_i(x) - \sum_{i=t+1}^{k} s_i(x). \tag{1}$$

The corresponding binary classifier defined by Π (and its labels) is $h_\Pi(x) = \text{sgn}(f_\Pi(x))$ where $\text{sgn}(z) = 1$ if $z \ge 0$ and $\text{sgn}(z) = -1$ if $z < 0$. (Note that $|f_\Pi(x)| \le k$ for all x.) We denote the class of all such f_Π by $\mathcal{F}$ and we denote by $\mathcal{H}$ the corresponding set of classifiers h_Π. In the context of learning, (1) defines the *margin* of h_Π at x.

To explain the idea behind this classifier, consider the contribution that a prototype p makes to the value $f_\Pi(x)$ of the discriminant function at x and suppose, without loss of generality, that p is a positive prototype. This prototype makes no contribution at all if x lies outside the sphere of influence of p. The rationale for this is simply that, in this case, there must be at least one negative prototype p^- whose distance from p is no more than the distance from x to p; and so there seems to be little justification for assuming x is close enough to p to derive some influence from the classification of p. If x does lie inside the sphere of influence of p, then there is a positive contribution to $f_\Pi(x)$ that is between 0 and 1 and is larger in absolute value the closer x is to p. The rationale here is that if x is deeply embedded in the sphere of influence of p (rather than being more on its periphery), and if we were considering how we should classify the point by taking into account