Pattern Recognition Letters 94 (2017) 22–29

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Generalized mean based back-propagation of errors for ambiguity resolution

Shounak Datta, Sankha Subhra Mullick, Swagatam Das*

Electronics and Communication Sciences Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata-700 108, India

ARTICLE INFO

Article history:
Received 21 July 2016
Available online 27 April 2017

Keywords:
Ambiguity resolution
Generalized mean
Multiple labels
Back-propagation
Multi-Layer Perceptron

ABSTRACT

Ambiguity in a dataset, characterized by data points having multiple target labels, may occur in many supervised learning applications. Such ambiguity originates naturally or from misinterpretation, faulty encoding, and/or incompleteness of data. However, most applications demand that a data point be assigned a single label. In such cases, the supervised learner must resolve the ambiguity. To effectively perform ambiguity resolution, we propose a new variant of the popular Multi-Layer Perceptron model, called the Generalized Mean Multi-Layer Perceptron (GMMLP). In GMMLP, a novel differentiable error function guides the back-propagation algorithm towards the minimum distant target for each data point. We evaluate the performance of the proposed algorithm against three alternative ambiguity resolvers on 20 new artificial datasets containing ambiguous data points. To further test for scalability and comparison with multi-label classifiers, 18 real datasets are also used to evaluate the new approach.

1. Introduction

Given a training dataset $S = \{(\mathbf{x}_i, c_i)|\mathbf{x}_i \in P \subset \mathbb{R}^d; c_i \in \mathcal{C} = \{1, 2, \cdots, C\}\}$, consisting of data points $\mathbf{x}_i$ in the training set P and their corresponding labels c_i, the traditional supervised learning problem is to identify the mapping $f : \mathbb{R}^d \to \mathcal{C}$ so that $f(\mathbf{x}_i) = c_i \ \forall \mathbf{x}_i \in P$. Then, the label for a new data point $\mathbf{y} \in Q \subset \mathbb{R}^d$ (Q being the test set) can be predicted to be $f(\mathbf{y})$. However, many practical applications are characterised by *ambiguous* training data points, i.e., data points with multiple corresponding labels. Formally, the notion of ambiguity in context to supervised learning can be defined as follows.

Definition 1. For a supervised learning problem with dataset $S = \{(\mathbf{x}_i, C_i)|\mathbf{x}_i \in P \subset \mathbb{R}^d; C_i \subseteq \mathcal{C}\}$, a data point $\mathbf{x}_i \in P$ is said to be ambiguous if $|C_i| \geq 2$.

Ambiguity in a supervised learning problem may stem from a variety of reasons such as label noise, lack of sufficient information to be able to distinguish between classes or concepts, faulty label encoding scheme resulting in distinct labels being assigned to a single concept, information overlap between classes etc. Such datasets can be subjected to two distinct forms of supervised learning, viz. *ambiguity resolution* and *multi-label learning* [20], which are defined as follows.

Definition 2. For a dataset S containing ambiguous data points, the problem of ambiguity resolution is to identify a suitable mapping $f_1 : \mathbb{R}^d \to \mathcal{C}$ such that $f_1(\mathbf{x}_i) \in C_i \ \forall \mathbf{x}_i \in P$.

Definition 3. The problem of multi-label learning, on the other hand, is to identify a suitable mapping $f_2 : \mathbb{R}^d \to \mathcal{P}(\mathcal{C}) \setminus \{\Phi\}$ ($\mathcal{P}(\mathcal{C})$ being the power set of $\mathcal{C}$) so that $f_2(\mathbf{x}_i) \equiv C_i \ \forall \mathbf{x}_i \in P$.

Therefore, ambiguity resolution also differs from multi-label learning, in the treatment of a test point $\mathbf{y}$, in that $f_1(\mathbf{y}) \in \mathcal{C}$ is a single predicted label while $f_2(\mathbf{y}) \subseteq \mathcal{C}$ is a set of possible labels. Multi-label learning predicts potentially multiple labels for a given data point. Consequently, it is not suitable for applications where a single label (out of multiple ambiguous labels) should be selected for each data point. Let us look at a few scenarios where the need for such ambiguity resolution arises.

1. If the training dataset is multi-labeled but the user insists that a single label be assigned to a query point [3]. A befitting example can be, the task of identifying individual personalities using a facial recognition classifier which is trained on multi-labeled news-feed images, where no correspondence between the labels and the personalities in an image is specified [4].
2. If multiple experts are used to label a dataset, their personal opinions, feelings, knowledge, and biases can cause some of the data points to be assigned with multiple ambiguous labels, only one of which is the true label. For example, detecting emotions

* Corresponding author.
E-mail addresses: shounak.jaduniv@gmail.com (S. Datta), sankha_r@isical.ac.in (S.S. Mullick), swagatam.das@isical.ac.in, swagatamdas19@yahoo.co.in (S. Das).

http://dx.doi.org/10.1016/j.patrec.2017.04.019
0167-8655/

from speech (EMA dataset [17]), recognising faces in a picture (LOST dataset [5]),predicting medical condition from clinical reports [31], etc. Such type of problems have been previously dealt with in [32,39], etc.
3. A similar but more challenging problem arises when the labels are crowd-sourced, resulting in almost every data point being labeled with a potentially large set of ambiguous labels, only one of which is correct. Detailed discussions on this problem can be found in [13,23], etc.
4. There can also be cases where the labels of a dataset becomes noisy or corrupted due to faulty transmission, storage, etc. A description and simulation strategy of such problems can be found in the literature on partial label learning [35,36].

Hence, ambiguity resolution which is the general problem encompassing all the above-mentioned scenarios can be significantly important and useful in many real-life applications.

Surprisingly, the present literature on learning with ambiguous data points abounds with paradigms of multi-label learning [20], while the equally (if not more) important problem of ambiguity resolution has received little attention. Some learners designed for handling multi-label problems are [22,28,34,37,38] etc. Another approach to deal with ambiguity is *preference learning* (more specifically *label ranking*), where each data point has a preference ranking corresponding to each label [14,33]. While it may help in resolving ambiguity, by assigning an ambiguous point with the label having maximum preferability (if there exists such an unique label), it usually requires prior preference information [9] which is often unavailable or costly. The major work in ambiguity resolution is that of Bullinaria [3], in which a common Multi-Layer Perceptron (MLP) is proposed to handle ambiguities in the $(g+1)$th epoch by drawing each ambiguous data point $\mathbf{x}_i$ towards the label $c_i^{(g)} \in C_i$ which generates minimum error for $\mathbf{x}_i$ in the gth epoch. The assumption behind this approach is that the ambiguity gets resolved automatically as the network gets trained on the non-ambiguous data points. However, the use of discrete minimum function prevented the application of the back-propagation algorithm directly to the non-differentiable error function. Moreover, such an approach also completely ignores the affinity that a data point may have towards other labels. The reliance on a large number of hidden nodes, as demonstrated by experiments in [3], is possibly a side-effect of the unusual learning method adopted.

In this article, we propose an elegant improvement over Bullinaria's early milestone by using the concept of the *generalized mean* [11].

Definition 4. The generalized mean μ_ρ of a set of real numbers $A \subset \mathbb{R}$ is defined as

$$\mu_\rho(A) = \left(\frac{1}{|A|} \sum_{a \in A} a^\rho \right)^{\frac{1}{\rho}}. \tag{1}$$

It is well known that the generalized mean of a set of values tends towards the minimum value, for a sufficiently small choice of the exponent, i.e. $\mu_\rho(A) \to \min(A)$ as $\rho \to -\infty$. The generalized mean function being both continuous as well as differentiable, unlike the minimum function, can be directly subjected to back-propagation based learning. Furthermore, the affinity to the minimum value can be controlled by varying the exponent ρ. Because of these desirable characteristics, we are motivated to utilize generalized mean for ambiguity resolution using MLPs.

The major contributions of the current study are summarized below:

1. We put forth a novel error function for back-propagation based learning of MLPs, which is able to handle non-ambiguous and ambiguous data points alike. We minimize the generalized mean of errors of each data point w.r.t. each of its target labels. Notice that the generalized mean of errors boils down to the traditional error function for an umambiguous data point.
2. We prepare a set of 20 artificial datasets having ambiguously labeled data points. The datasets, which are diverse in terms of structure, dimensions, and extent of ambiguity, can be found at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FO4RIRM.
3. The proposed method is tested on the 20 artificial datasets created by us and on 10 other real-life ambiguous datasets (without ground truth information), from various fields like bioinformatics, video annotation, etc. We compare its performance with those of three alternative ambiguity resolution strategies and a neural network based multi-label classifier called BP-MLL [37].
4. We establish the better performance of our proposed ambiguity resolver compared to the multi-label classifier BP-MLL on datasets where ground truth is available. To simulate noisy labeled datasets we use 6 real-life datasets from the UCI repository [18] following [35]. To illustrate our algorithm's improved immunity against inexperienced, misguided and/or biased experts we also conduct experiments on the LOST and EMA datasets. We conduct Wilcoxon signed rank [6] and Mann–Whitney U tests [10] to establish the superiority of the proposed learner in a statistically significant way.

Organization of this paper is in order. We derive the expressions for the proposed back-propagation method in Section 2. Subsequently, in Section 3, we describe the used datasets and the experimental procedure. Next in the same section, we present the experimental results and analyse them. We finally conclude the article with a brief summary and remarks in Section 4.

2. Generalized mean multi-layer perceptron

MLP [12] is a popular non-parametric supervised learner having a network architecture. Hence, it does not require any prior assumptions about the class distributions of the datasets to be learned and can effectively generate non-linear separation boundaries to distinguish between structurally complex classes. The structure of an MLP is simple and highly parallel in nature, making it suitable for high-dimensional data processing. Back-propagation of errors [25], scaled conjugate gradient method [21], and many other learning algorithms have been designed to train MLPs. All these factors have influenced us to use MLP as the underlying supervised learner for ambiguity resolution. We refer to the proposed MLP based ambiguity resolver as *Generalized Mean Multi-Layer Perceptron* (GMMLP), the details of which are presented in the rest of this section.

2.1. Generalized mean based error function

Let us consider an MLP consisting of an input layer of $(d+1)$ nodes, a single hidden layer having α nodes, and an output having C nodes (as many nodes as the number of possible classes). Let u_{kb} denote the weight of the connection from the kth input node to the bth hidden node, and let u_{0b} denote the bias term of the bth hidden node. Similarly, let v_{br} denote the weight of the connection from the bth hidden node to the rth output node, and let v_{0r} denote the bias term of the rth output node. Moreover, let $U = [u_{kb}]_{d\times\alpha}$, $\mathbf{u}_0 = [u_{0b}]_{\alpha\times 1}$, $V = [v_{br}]_{\alpha\times C}$, and $\mathbf{v}_0 = [v_{0r}]_{C\times 1}$ denote the matrices and vectors of the weights and the bias terms. Let the activation function be the sigmoid function

$$\psi(x) = \frac{1}{1+e^{-\gamma x}}, \tag{2}$$

where we set the skewness parameter $\gamma = 1$, in keeping with general conventions.