Pattern Recognition Letters 104 (2018) 8–14

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Exploiting covariate embeddings for classification using Gaussian processes

Daniel Andrade[a,*], Akihiro Tamura[b], Masaaki Tsuchida[c]

[a] Security Research Laboratories, NEC Corporation, Japan
[b] Graduate School of Science and Engineering, Ehime University, Japan
[c] DeNA Co., Ltd., Japan

ARTICLE INFO

Article history:
Received 16 May 2017
Available online 16 January 2018

Keywords:
Logistic regression
Auxiliary information of covariates
Gaussian process
Text classification

ABSTRACT

In many logistic regression tasks, auxiliary information about the covariates is available. For example, a user might be able to specify a similarity measure between the covariates, or an embedding (feature vector) for each covariate, which is created from unlabeled data. In particular for text classification, the covariates (words) can be described by word embeddings or similarity measures from lexical resources like WordNet. We propose a new method to use such embeddings of covariates for logistic regression. Our method consists of two main components. The first component is a Gaussian process (GP) with a covariance function that models the correlations between covariates, and returns a noise-free estimate of the covariates. The second component is a logistic regression model that uses these noise-free estimates. One advantage of our model is that the covariance function can be adjusted to the training data using maximum likelihood. Another advantage is that new covariates that never occurred in the training data can be incorporated at test time, while run-time increases only linearly in the number of new covariates. Our experiments demonstrate the usefulness of our method in situations when only small training data is available.

1. Introduction

Classification is ubiquitous in many applications in machine learning and statistics. However, for small training data, classification performance is often insufficient, and, as a consequence, several types of additional knowledge is included:

- unlabeled data using semi-supervised learning techniques [1],
- assumptions about the generation process of the data [2],
- auxiliary information about samples [3],
- auxiliary information about covariates [4].

Here, in this work, we focus on incorporating auxiliary information about covariates that are given in the form of similarity information or embeddings. For text classification, where covariates are single words, covariate embeddings can be easily acquired from unlabeled documents using, for instance, word2vec [5] or GloVe [6]. Alternatively, similarities between covariates can be manually defined, and are available in resources like WordNet [7]. In the latter case, covariate embeddings can be easily learned using spectral decomposition of similarity matrices (see Supplementary Material, Section 2).

In order to incorporate the knowledge of covariates into logistic regression, we propose to model the interaction of the covariates by a Gaussian process (GP). The use of a Gaussian process allows us to directly model the joint covariate distribution by an appropriate covariance function that depends on the covariate embeddings. Our model assumes that the true (unknown) value of the covariates are generated from a GP, and the observed values are due to additive noise. By recovering the true covariate values, our model is able to adjust also values of related covariates that are not observed in the sample. In particular, for text classification, our method finds positive weights of semantically related words that do not explicitly occur in the document. Our proposed method performs effectively a kind of smoothing of the covariate vector that is controlled by the parameters of the covariance function and the noise variance.

Previous work using such covariate information mainly concentrates on designing ontology-specific kernels [4,8] or semantic smoothing kernels from unlabeled data that cannot be adjusted to the labeled training data at hand [9,10]. Wittek and Tan [11] proposes a wavelet kernel that can incorporate distance information between covariates. However, their method requires to create a

* Corresponding author.
E-mail address: s-andrade@cj.jp.nec.com (D. Andrade).

https://doi.org/10.1016/j.patrec.2018.01.011
0167-8655/

one-dimensional ordering of the covariates which can be inappropriate for some applications.

One simple method to use word embeddings for classification is to use the average [12] or weighted-average [13] of the word embeddings to represent a document. Our method compares favorable to such methods while at the same time enjoying the same good interpretability as a bag-of-words model (for an example, see Table 3).

A different solution, which is able to directly use a similarity measure between covariates, is known as bag-of-clusters (BOC), where similar covariates are grouped together [14,15]. However, fixing the clusters a-priori can be disadvantageous and can result in lower accuracy (as confirmed by our experiments in Section 3).

If large amount of training data is available, then a natural choice is to use neural network models for classification like suggested in [16,17]. However, due to their high number of model parameters, these models are not appropriate for small training data sets which is the focus of this work.

2. Proposed method

Let $\mathbf{x} = (x_1, \ldots, x_p)^T$ be the vector of p observed covariates, and y be the response variable. Furthermore, $\mathbf{e}_j \in \mathbb{R}^h$ denotes the embedding of covariate j, and $E = (\mathbf{e}_1, \ldots, \mathbf{e}_p)$ denotes the matrix of all covariate embeddings. Our proposed method assumes that the covariates $(x_1, \ldots, x_p)^T$ are disrupted by noise, and that E helps to recover a (true) noise-free covariate vector $\mathbf{f} = (f_1, \ldots, f_p)^T$. Formally, we model $p(y|\mathbf{x})$ as follows:

$$p(y|\mathbf{x}, \boldsymbol{\psi}, \boldsymbol{\theta}) = \int_{\mathbf{f}} p(y|\mathbf{f}, \boldsymbol{\psi}) p(\mathbf{f}|\mathbf{x}, E, \boldsymbol{\theta}) \,, \tag{1}$$

where $p(y|\mathbf{f}, \boldsymbol{\psi})$ is modeled by a logistic regression model (with parameter vector $\boldsymbol{\psi}$) using the noise-free covariate vector $\mathbf{f}$, and $p(\mathbf{f}|\mathbf{x}, E, \boldsymbol{\theta})$ is a Gaussian process model (with parameter vector $\boldsymbol{\theta}$) for recovering the noise-free covariate vector. The conditional independence assumptions of our model are shown in Fig. 2.

Since the calculation of the integral is infeasible, we approximate Eq. (1) as follows

$$\begin{aligned} p(y|\mathbf{x}, \boldsymbol{\psi}, \boldsymbol{\theta}) &= \int_{\mathbf{f}} p(y|\mathbf{f}, \boldsymbol{\psi}) p(\mathbf{f}|\mathbf{x}, E, \boldsymbol{\theta}) \\ &\approx \int_{\mathbf{f}} p(y|\hat{\mathbf{f}}, \boldsymbol{\psi}) p(\mathbf{f}|\mathbf{x}, E, \boldsymbol{\theta}) \\ &= p(y|\hat{\mathbf{f}}, \boldsymbol{\psi}) \,, \end{aligned}$$

where

$$\hat{\mathbf{f}} = \operatorname*{argmax}_{\mathbf{f}} p(\mathbf{f}|\mathbf{x}, E, \boldsymbol{\theta}) \,.$$

We note that the relationship between Eq. (1) and its approximation, is analogously to the relationship between a fully Bayesian posterior predictive distribution and a maximum-a-posterior (MAP) predictive distribution.

In the following, we explain in detail our model for $p(\mathbf{f}|\mathbf{x}, E, \boldsymbol{\theta})$ (Section 2.1), the combined logistic regression model $p(y|\hat{\mathbf{f}}, \boldsymbol{\psi})$ (Section 2.2) and how to estimate the parameters $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ from the training data (Section 2.3).

2.1. Recovering the noise-free covariate vector

We assume that $\mathbf{f}$ is a function of the covariate embedding. Let f denote the function that maps the covariate embedding of covariate j to its value f_j, i.e. $f_j = f(\mathbf{e}_j)$, where $\mathbf{e}_j \in \mathbb{R}^h$ is the embedding of covariate j.

We assume that $f_j \in \mathbb{R}$ is distributed according to a Gaussian process with a fixed mean $m_j \in \mathbb{R}$ and covariance function $k(\mathbf{e}_{j_1}, \mathbf{e}_{j_2})$. As covariance function, we use the squared exponential covariance function

$$k(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}) = exp\Big(-\frac{1}{2l^2} \cdot r^2_{j_1,j_2}\Big),$$

where l defines the characteristic length scale [18]. Here, we set r_{j_1,j_2} to the euclidean distance between the covariates j_1 and j_2, i.e.

$$r_{j_1,j_2} := ||\mathbf{e}_{j_1} - \mathbf{e}_{j_2}||_2 \,.$$

The above choice of the covariance function achieves that two covariates have similar values, if their embeddings are similar.

Furthermore, we make the assumption that the true, but unobserved, covariate value f_j is disturbed by isotropic Gaussian noise, leading to the observed covariate value x_j. In summary, our generative model can be written as:

1. Sample f from a GP

   $$f \sim GP(m_j, k(\mathbf{e}_{j_1}, \mathbf{e}_{j_2})) \,.$$

2. Add isotropic noise

   $$x_j = f(\mathbf{e}_j) + \epsilon \,,$$

   where the noise ϵ is sampled from $N(0, \sigma^2)$.

Given a sample with covariate vector $\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$, our goal is to get an estimate of the true covariate value f_j. We choose as a point-estimate of f_j the maximum a posteriori (MAP) estimate. It is easy to show[1] that $p(f_j|x_1, \ldots, x_p)$ is a Gaussian distribution with mode $\operatorname{argmax}_{f_j} p(f_j|x_1, \ldots, x_p)$ equaling

$$m_j + \mathbf{k}_j^T (K_l + \sigma^2 I)^{-1} (\mathbf{x} - \mathbf{m}) \,, \tag{2}$$

where $\mathbf{m} = (m_1, \ldots, m_p) \in \mathbb{R}^p$ is the mean of the GP, and $I \in \mathbb{R}^{p \times p}$ is the identity matrix. The co-variance matrix $K_l \in \mathbb{R}^{p \times p}$ is defined by

$$K_{l_{j_1,j_2}} := k(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}) \,.$$

The vector $\mathbf{k}_j \in \mathbb{R}^p$ denotes the jth column vector of covariance matrix K_l. Assuming that the mean of each sample is close to 0, or alternatively, the data is centered, we can use a zero mean GP. This way, we can get an estimate of the noise-free covariate vector by left-multiplication of $\mathbf{x}$ with the smoothing matrix $S_{\boldsymbol{\theta}}$ defined as

$$S_{\boldsymbol{\theta}} := K_l (K_l + \sigma^2 I)^{-1} \,,$$

where $\boldsymbol{\theta}$ denotes the covariance parameters $\{l, \sigma\}$. We note that $S_{\boldsymbol{\theta}}$ can be calculated in advance at training time. The calculation of $S_{\boldsymbol{\theta}}$ is feasible for more than 10k covariates by using Cholesky factorization. In summary, using Eq. (2), our noise-free estimate of the covariate vector is

$$\hat{\mathbf{f}} := \operatorname*{argmax}_{\mathbf{f}} p(\mathbf{f}|\mathbf{x}, E, \boldsymbol{\theta}) = S_{\boldsymbol{\theta}} \mathbf{x} \,. \tag{3}$$

The proposed method has the desired effect that covariates that are related to many other covariates that have high observed values, will also get high values. For example, in text classification, words that are related to the document but have zero value (e.g. words that did not occur in the document) will get positive values.

To illustrate the effect, let us consider an example document which contains two words: “funny” and “tears” both with value 5 (e.g. tf-idf weight). The text does not contain the words “melancholic”, “sad”, and “humor”, i.e. these values are 0. Furthermore, only for illustration, assume that the covariate embeddings are one-dimensional, then the observed value of each covariate can be illustrated as shown on the left-hand side of Fig. 1. For example,

[1] See e.g. [18] page 27.