Neural Networks 98 (2018) 87–101

Contents lists available at ScienceDirect

Neural Networks

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

Support vector machine with Dirichlet feature mapping

Ali Nedaie, Amir Abbas Najafi *

Faculty of Industrial Engineering, K.N. Toosi University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Received 20 February 2017
Received in revised form 5 November 2017
Accepted 7 November 2017
Available online 16 November 2017

Keywords:
Supervised learning
Support vector machine
Nonlinear mapping
Kernel function
Dirichlet distribution

ABSTRACT

The Support Vector Machine (SVM) is a supervised learning algorithm to analyze data and recognize patterns. The standard SVM suffers from some limitations in nonlinear classification problems. To tackle these limitations, the nonlinear form of the SVM poses a modified machine based on the kernel functions or other nonlinear feature mappings obviating the mentioned imperfection. However, choosing an efficient kernel or feature mapping function is strongly dependent on data structure. Thus, a flexible feature mapping can be confidently applied in different types of data structures without challenging a kernel selection and its tuning. This paper introduces a new flexible feature mapping approach based on the Dirichlet distribution in order to develop an efficient SVM for nonlinear data structures. To determine the parameters of the Dirichlet mapping, a tuning technique is employed based on the maximum likelihood estimation and Newton's optimization method. The numerical results illustrate the superiority of the proposed machine in terms of the accuracy and relative error rate measures in comparison to the traditional ones.

1. Introduction

Ever since support vector machine (SVM) was introduced by Vapnik (1998), it has received a great deal of attention from many researchers during the past decade. An SVM is a non-parametric max-margin classification technique aims at classifying data into two groups, making it useful for two-group classification problems. However, it has also been extended to solve multi-group classification problems (Crammer & Singer, 2001; López & Maldonado, 2016; Platt, Cristianini, & Shawe-Taylor, 1999; Weston & Watkins, 1998; Zangooei & Jalili, 2012), continuous output cases (Suykens & Vandewalle, 1999; Xu, An, Qiao, Zhu, & Li, 2013), semi-supervised problems (Reitmaier & Sick, 2015), rank learning (Kim, Ko, Han, & Yu, 2014) and so on. According to the previous researches, SVM has outperformed other classification techniques in terms of the accuracy measures (Auria & Moro, 2008; Joachims, 1998). Unfortunately, it suffers from some limitations which lead to inefficient separation in some datasets, so that the efficiency of this method associates to data structure. This fact is especially true for linear models of SVMs which are not flexible enough and consequently are strongly case-dependent. The traditional trick to avoid such inefficiency is to use kernel functions and other feature mappings combined with an SVM to establish a nonlinear classifier. Furthermore, training the SVM model and kernel parameters considerably affect its results. For a detailed description of training SVMs refer to Oneto, Ridella, & Anguita, 2016.

A rich literature can be found in the current area focused on improving the generalization of SVM based on combining methods or modifying separating hyperplanes, classification margin, boundary, etc. Zhang and Zhou (2016) presented a nonlinear combination of fisher discriminant analysis and SVM to obtain good statistical separability. Ma, Rana, Taghia, Flierl, and Leijon (2014) developed a minimum within-class and maximum between-class scatter machine in order to take the class distribution into consideration. Some researchers used ellipsoidal boundaries (Czarnecki & Tabor, 2014) and non-parallel hyperplanes (Shao, Chen, & Deng, 2014) to achieve better performance. Cuong and Van Thien (2016) proposed a method to reduce the number of support vectors to improve the convergence speed. Some related works in the literature also focused on improving the flexibility and accuracy of the SVM (Yue, Finley, Radlinski, & Joachims, 2007). In summary, one can categorize previous studies into two major groups: modifying and developing the base model of the SVM and using kernel or mapping functions which can be used to increase the SVM flexibility and improve accuracy in some cases (Chen, Wang, & Zhong, 2016; Tanveer, Shubham, Aldhaifallah, & Ho, 2016).

To provide a comprehensible report of our study, the remaining sections are organized as follows: In Section 2, related works will be surveyed to highlight the novelty of the proposed machine comparing to the similar works. Section 3 is designated to review traditional SVM. The Dirichlet distribution and its properties are described in Section 4. Some statements about the flexibility of this distribution are also given in this section. In Section 5, the main contribution of the current research is delineated and the

* Corresponding author.
E-mail address: aanajafi@kntu.ac.ir (A.A. Najafi).

https://doi.org/10.1016/j.neunet.2017.11.006
0893-6080/

model of the Dirichlet SVM is presented. Numerical experiments performed as a means of investigating are presented in Section 6, including both synthesized and real-case datasets. To provide more reliable results, Section 7 is dedicated to compare the accuracy and robustness of the models under different nuisance rates. Finally, Section 8 concludes the paper.

2. Related works and contribution

Many studies in the literature have focused on improving the SVM model from different viewpoints. In this regard, kernel and feature mapping function are interesting topics received much attention by the researchers. One of the earliest works in this area was conducted by Amari and Wu (1999) based on the Riemannian geometry; where, the spatial resolution around the separating boundary surface is enlarged such that the separability between classes is increased.

Yujian, Bo, Xinwu, Yaozong, and Houjun (2011) presented a solid geometric theory and developed piecewise linear classifiers for support vector machine. They provided two novel algorithms based on SVM called support conlitron algorithm and support multiconlitron algorithm. This scheme was extended to alternating version and simplified by Li and Leng (2015).

Peng, Hu, Chen, and Dang (2015) proposed a mapping function to embed nominal attributes into a real space by minimizing an estimated generalization error for heterogeneous data and achieved more accuracy comparing to traditional machine.

Huang, Mehrkanoon, and Suykens (2013) developed a linear piecewise feature mapping based on hinging hyperplanes to provide a piecewise boundary for classifying data. They suggested a segmentation scheme on data and showed that using linear piecewise SVM can considerably improve the accuracy measure. Such a segmentation is challenging to perform and the relevant criteria is selected randomly. Thus, given the above, it is evident that different segmentation approaches result in unreliable accuracy. Another linear mapping function was developed by Huang, Suykens, Wang, Hornegger, and Maier (2017) called ℓ_1 Distance Kernel which is linear in subregions and nonlinear in global region.

A nonlinear mapping was introduced by Nedaie and Najafi (2016). They developed a new model called polar SVM based on the polar coordinate system. Angles and distances are used in polar SVM instead of the Cartesian coordinates. They made this choice because complex separators can be established in polar system more simply. Other similar works were performed by Burges (1998) and Li, Yang, Gu, and Zhang (2013).

In the current area, Parameter(s) tuning is an important issue in using kernel and feature mapping functions. In many researches, kernels and mapping functions are subjected to parameter tuning or other initializations, such as the number of segments (Carrizosa, Martín-Barragán, & Morales, 2014; Wu & Wang, 2009). In this regard, Yin and Yin (2016) showed that the performance of an SVM depends highly on the selection of the kernel function type and relevant parameters and proposed a novel index to serve as a better class separability criterion. In sum, a flexible feature mapping approach with an efficient tuning procedure is an essential aspect of SVM modeling. This notion has motivated our research, the aim of which is to introduce a new feature mapping based on a flexible function called Dirichlet distribution.

Closely similar to our contribution, Bdiri and Bouguila (2013) employed inverted Dirichlet distribution for kernel generation of support vector machine which can be applied in sequence data with different length; where, the performance of the model is greatly dependent on the posterior. To the best of our knowledge, this is the only research that uses inverse probability density function of Dirichlet to develop kernels (not feature mapping) for SVM. The proposed functions in this study is useful in classifying datapoints with different length i.e. text/image pattern recognition problems. Slightly apart from this research, our work aims at developing a new mapping function based on Dirichlet distribution that does not suffers from challenges of posterior information. In addition, the proposed function can be tuned based on Newton's algorithm; where, it will be shown that the convergence to optimal parameters is guaranteed.

A Dirichlet distribution is generalized form of Beta distribution. Due to the ability of the Dirichlet distribution to generate linear, convex and concave hyperplanes, it is very flexible for mapping data into the feature space. Generally, owing to its different shapes governed by its parameters, this distribution can be applied as an efficient feature mapping. Hence, our proposed approach combines a traditional machine with a Dirichlet mapping to provide an efficient model that incorporates the advantages of this function. The new model is introduced and investigated in the subsequent sections.

3. Support vector machine

Support vector machine is a popular and well-known supervised learning technique, aimed at finding a max-margin separator hyperplane to classify data (Sklar, 2014). When an SVM model is combined with a kernel function, it provides better flexibility relative to the simple form to apply in different data structures. For greater understanding of this concept, consider a dataset $\Xi = \{x_i, y_i\}_{i=1}^m$ to be classified, where $x_i \in R^n$ and $y_i \in \{-1, 1\}$ are datapoints and class labels, respectively. This case is a binary classification problem that seeks a function $f(x)$ in which $sign\ (f(x))$ is the most accurate separator for the given data and can be obtained by solving the quadratic programming problem as noted below:

$$\begin{aligned} &\min \frac{1}{2}\sum_{j=1}^{n}\omega_j + C\sum_{i=1}^{m}\xi_i \\ &\quad y_i\left(\omega^T\phi\ (x_i) + b\right) \geq 1 - \xi_i\,, \quad i = 1, \ldots, m \\ &\quad \xi_i \geq 0, \quad i = 1, \ldots, m \end{aligned} \tag{1}$$

where, ω_j and b are the normal vector and bias of the separator, respectively; ξ_i denotes the slack variables and the positive scalar C is penalty of misclassification. In addition, the term $\phi\ (x)$ is a feature mapping (usually a kernel function) which maps features to a higher or infinite dimensional space called feature space to provide a nonlinear separator and can be looked-up depending on the data structure. Generally, the nonlinear SVM performance is strongly dependent on selecting a suitable kernel function and tuning its parameter(s). While many methods for tuning SVM parameters have been proposed, their solutions may yield local optima only. Moreover, there is no rule-based procedure to select a mapping function or kernel and consequently making it difficult to attain the desired level of accuracy. As a more flexible kernel function would result in a more accurate model, the current work presents here introduces a flexible mapping function based on the Dirichlet distribution discussed below.

4. Dirichlet distribution

Dirichlet distribution is the generalized multivariate form of beta distribution. The flexibility provided by the Dirichlet distribution is used to model the shape of the different functions. This distribution is able to construct linear, convex and concave hulls by manipulating the parameters. Fig. 1 shows the flexibility of this distribution regarding the different values for parameters α. The following is the probability function of Dirichlet distribution:

$$f_X\left(x_1, x_2, \ldots, x_{n-1}\right) = \frac{\Gamma\left(\sum_{j=1}^{n}\alpha_j\right)}{\prod_{j=1}^{n}\left(\Gamma\left(\alpha_j\right)\right)}\prod_{j=1}^{n}x_j^{\alpha_j-1}. \tag{2}$$