ARTICLE IN PRESS
JID: PR [m5G;May 12, 2017;2:57]

Pattern Recognition 000 (2017) 1–8

Contents lists available at ScienceDirect

Pattern Recognition

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

Adaptive multiclass support vector machine for multimodal data analysis

Xin Zhang, Xiu Zhang*

[a] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin, China
[b] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin, China

ARTICLE INFO

Article history:
Received 4 March 2017
Revised 2 April 2017
Accepted 7 May 2017
Available online xxx

Keywords:
Artificial bee colony
Feature selection
Hyperparameter optimization
Multiclass classification
Support vector machine

ABSTRACT

Multimodal data commonly exists in human lives. Early analysis usually concentrates on mining information based on single modality. Recent studies show that learning tasks could be greatly enhanced by analyzing data from the aspect of multimodality. This paper deals with classifying multimodal data comprised of visual and acoustic contents. Different data features are fused under a hierarchical structure to achieve a good semantic understanding. Then, to accomplish accurate classification, an adaptive support vector machine method (ASVM) is proposed. The method is support vector machine with hyperparameters controlled by a novel and efficient artificial bee colony algorithm. First, a micro colony is set as the number of hyperparameters is usually less than 5. Second, one position inheritance based on roulette wheel selection is used. Third, discarded solutions are mutated by position shift operation instead of random reinitialization. The ASVM method is first verified on classical data sets demonstrating the goodness of the proposed method. Then the proposed method is applied on a multimodal data set. Each sample includes both image and audio data features. Experimental results show that the ASVM method is more effective and robust than the compared methods.

1. Introduction

Multimodal data analysis presents better performance in many applications than conventional unimodal data analysis. A multimodal content retrieval framework is able to combine unimodal heterogeneous similarities and capture semantic correlations of multiple modalities [1]. Deep learning is recently employed to deal with multimodal data learning. For classifying solar radio burst problem, multimodal learning framework is built to model solar radio spectrums. Autoencoder and structured regularization are taken to learn intra-modality and inter-modality information. Classification is performed with higher accuracy based on joint representation of multimodal features [2]. Visual features of image data can be represented in multimodality. To capture individual visual features, a supergraph is constructed by merging small prototype graphs, which results good scalability of the method [3]. Extracting features from both gray-scale ultrasound images and color doppler energy images could also assist the solving of placental maturity staging problem [4]. To recognize number and text detection in Marathon images, a multimodal method is useful which performs torso detection first and then text detection [5]. In image registration, multimodal optimizing of automatic correspondence also presents better performance than unimodal methods [6].

Multimodal data analysis is necessary for smart and autonomous systems such as autonomous cars or UAVs which are often equipped with radars, video/images, and audio sensors [7]. The multimodal data analysis will provide basis for situation understanding [8] to avoid potential danger or threat [9]. Multimodal data analysis and integration in smart and autonomous systems often happens in networked environments such as Internet of Things, sensor networks [10]. In this paper, we focus on online agent in social networks which needs multimodal data analysis. Mining information from multimodal data is a hot topic in machine learning and artificial intelligence [11]. Typical mining tasks include semantic extraction, object detection, tracking and so on. Early analysis of multimodal data usually concentrates on using features of single modality, while researchers find that analysis of such data using multimodal features is very useful for mining tasks [12,13]. Although a few researches have been reported, the study of multimodal data analysis from the aspect of multimodality still deserves further study to assure accurate and robust accomplishment of mining tasks.

As the growth spurt of multimodal data, this causes great challenge for data mining, browsing and organizing [14]. Semantic ex-

* Corresponding author.
E-mail addresses: ecemark@mail.tjnu.edu.cn (X. Zhang), ecezhang@mail.tjnu.edu.cn (X. Zhang).

http://dx.doi.org/10.1016/j.patcog.2017.05.006

traction is very helpful to assist users to obtain information from social networks [15]. By using multiple modality features, a good semantic model could be attained, and thus accurate classification is reached. This can greatly improve user based recommendation and quality of service for user experience [16].

Enhancing classification accuracy is very important in providing a high quality of service and attracting active users in social media [17], [18]. Moreover, multiple classes are commonly across as multimodal data contains various activities in daily lives. This causes trouble for conventional classifiers such as support vector machine (SVM) [19], which is initially designed for two class classification. On the other hand, many machine learning methods are developed to improve the prediction accuracy or reduce computational time of training data. As a huge amount of samples is a must in continuous monitoring, fast and effective classification method is of high demand [20,21].

SVM is one of the mostly studied machine learning techniques. It is based on Vapnik–Chervonenkis's statistical learning theory [22,23]. Its design philosophy is to trade off between the minimization of training set error and the maximization margin of hyperplanes. In solving the model of SVM, convex quadratic programming is used, which is very efficient in convergence rate and reaching global optimum. SVM method has been successfully employed in many applications such as handwritten digit recognition, speech recognition, text categorization, fingerprint recognition, and so on [24]. SVM model contains several parameters which have to be predefined by users. For example, penalty parameter C is typically used to balance training set error and generalization ability, which resulting C-SVM. Kernel functions of SVM model also involves parameters. For example, Gaussian kernel or radial basis function contains parameter γ, which is related with polar radius of kernel function. Such parameters are also called hyperparameters. The setting of hyperparameters is a notorious problem in SVM as their setting may heavily impact prediction accuracy [25].

Many works have been reported in model selection and hyperparameter setting problems of SVM. Among these works, evolutionary computing approaches are popular. Such approaches can be divided into evolutionary algorithms and swarm intelligence. The former includes genetic algorithm (GA), evolutionary strategies, genetic programming, etc.; while the latter includes particle swarm optimization (PSO), firefly algorithm, artificial bee colony (ABC), neighborhood field optimization, etc. The main advantage of evolutionary computing approaches is these approaches assume hyperparameter setting as a black-box, also do not need gradient information. Because the relation of hyperparameter with SVM model is hard to define, let alone the computing of gradient or Hessian matrix, evolutionary computing approaches become appropriate choices for hyperparameter selection problem.

In contrast to GA and PSO, ABC is as effectiveness as them in solving mathematical functions and engineering design with high dimensionality, though ABC involves less algorithmic parameters than GA and PSO [26]. Hence, the paradigm of ABC is selected as basis to handle the hyperparameter selection problem. The resulting method is named as adaptive support vector machine (ASVM). As the number of hyperparameters in SVM model is less than 5, this fact is considered in the proposed method. ABC is assigned a micro colony to reduce the number of iterations so that training time could be greatly reduced. Moreover, a modified one-position inheritance scheme is designed to promote convergence rate of ABC. It is helpful in information exchange of current hyperparameters in colony. Furthermore, shift operation substitutes reinitialization of discarded hyperparameters in scout stage. Each dimension is probably being shifted depending on the trials it fails to produce fitness improvement. The ASVM method is first examined on classical classification data sets as well as comparing with other classification methods. The proposed method is then employed to classify multimodal data collected from social media websites.

The remainder of the paper is presented as follows. Section 2 reports SVM model, ABC and related works. Section 3 gives the proposed ASVM algorithm and its design philosophy. Section 4 shows experimental results results as well as result analysis and discussion. Section 5 concludes the paper.

2. Methodologies and related works

This section introduces SVM model and standard ABC algorithm as well as related works of these two aspects.

2.1. Support vector machine

In classification task, suppose a training data set is given by pairs $(\mathbf{x}_i), y_i,\ i = 1, 2, \ldots, l$. Each sample is comprised of instance $\mathbf{x}_i$ and its associated class label y_i. The C-SVM refers to the following model:

$$\begin{array}{ll} \min\limits_{\mathbf{w},b,\boldsymbol{\xi}} & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum\limits_{i=1}^{l}\xi_i \\ s.t. & y_i\left(\mathbf{w}^T\phi(\mathbf{x}_i) + b\right) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{array}, \tag{1}$$

where $\mathbf{w}$ and b defines the hyperplane to classify training samples. As samples are usually non-separable in real world applications, relaxation factor $\boldsymbol{\xi}$ is introduced to create a flexible classification. C is the so called regularization parameter, which is a tradeoff of maximizing class margin and minimizing training error of samples [27]. For complex problems, samples could be difficult to separate. Hence, instances are often mapped to higher dimensional space so that instances become separable in higher dimensional space. $\phi(\mathbf{x}_i)$ is such a function and $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is called kernel function [28]. Typical kernel functions are as follows:

(1) Linear kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$;
(2) Polynomial kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma\mathbf{x}_i^T\mathbf{x}_j + r)^d, \gamma > 0$;
(3) Radial basis function: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2), \gamma > 0$;
(4) Sigmoid function: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma\mathbf{x}_i^T\mathbf{x}_j + r)$.

In the above kernel functions, γ, r and d are kernel parameters. Clearly, different kernel may have different parameters. Linear kernel does not require parameters, though its performance is not effective enough to satisfy the need of users. Polynomial function, radial basis function (RBF) and sigmoid function show good performance but additional parameters have to be set for them.

Besides C-SVM, the ν-support vector machine (ν-SVM) is another commonly used SVM model.

$$\begin{array}{ll} \min\limits_{\mathbf{w},b,\boldsymbol{\xi},\rho} & \frac{1}{2}\mathbf{w}^T\mathbf{w} - \nu\rho + \frac{1}{l}\sum\limits_{i=1}^{l}\xi_i \\ s.t. & y_i\left(\mathbf{w}^T\phi(\mathbf{x}_i) + b\right) \geq \rho - \xi_i \\ & \xi_i \geq 0 \\ & \rho \geq 0 \end{array}, \tag{2}$$

where parameter ν has been proved to be an upper bound on the fraction of error samples over training samples, also it is a lower bound on the fraction of support vectors over training samples.

The setting of parameters of SVM model (C-SVM, ν-SVM) or kernel functions is considered as hyperparameter setting problem, which has been studied and recent researches are summarized in the following. Ch et al. studied the coupling of firefly algorithm with SVM to predict the malarial incidences in Jodhpur and Bikaner area [29]. Their algorithm accomplished more accurate prediction than artificial neural network (ANN) and SVM without tuning hyperparameters. Ding et al. proposed to improve