# 基于 PSO 优化孪生支持向量机的自然语言处理

吴倍斯

四川托普信息技术职业学院 四川 成都 610000

**[摘 要]**在自然语言处理（NLP）中，文本分类、命名实体识别和机器翻译等任务是通过使用文本特征向量来进行训练和预测。然而，训练样本的数量通常相对较少，并且随着文本长度的增加，计算时间也会变得越来越昂贵。因此，在实际应用中，需要构建大量的训练集。但是，由于目前所使用的词向量和词袋模型（Bag of Words, BOW）无法捕获单词之间的关系和上下文信息，因此难以预测新的单词或词组。同时，对于自然语言处理中的许多问题（例如命名实体识别、机器翻译等），传统的支持向量机（Support Vector Machine, SVM）方法通常是最小化分类错误率。然而， SVM 存在一个明显的缺陷：如果参数设置不当，将导致过拟合。 粒子群优化（PSO）算法是一种基于群体智能原理的随机优化算法。其基本思想是通过调整种群中每一代个体的位置和速度来获得全局最优解。在本文中，我们提出了一种基于 PSO 优化孪生支持向量机（Twin Support Vector Machine, SWSVM）的方法。

**[关键词]**自然语言处理；孪生支持向量机；粒子群算法；惩罚因子

**[中图分类号]** G641 **[文献标识码]**A **[文章编号]**1647-9265(2024)-0097-11 **[收稿日期]**2024-01-21

## 一、算法简介

SVM 是一种无监督的机器学习方法，它利用最小化支持向量机模型中的超参数，并根据不同的情况来选择最优的超参数。SVM 已经被广泛用于各种基于文本的任务，例如文本分类、命名实体识别和机器翻译等。SVM 的基本思想是将一个向量空间映射到另一个向量空间。映射后的向量集被称为特征空间，映射前的向量集被称为目标空间。在特征空间中，可以选择若干个超参数来训练和预测目标函数。这些超参数中，有些是固定的，如核函数、惩罚因子和正则化系数；有些是可变的，如核函数和正则化系数。不同的超参数对训练结果有不同的影响。因此，选择最优超参数是非常重要的。

基于粒子群优化算法（PSO）优化 SVM 算法具有简单易实现、鲁棒性好等优点。但是， PSO 算法存在以下两个缺点：

当 PSO 算法遇到新问题时，它将采用一个新的适应度值来更新种群中每一代个体的位置和速度，然后再将这些新个体结合起来以获得全局最优解。这种方法缺乏严格性，因此对初始值非常敏感。

PSO 算法对初始值非常敏感。如果在训练过程中使用了过多的初始值，就会导致模型过于“稳定”和“单调”，从而使模型无法学习到真正有用的特征。

我们使用 SWSVM 作为孪生支持向量机（SWVM）模型（SV）以生成文本特征向量，并使用 PSO 算法优化这些特征向量以获得最优超参数（SVM 参数）。

我们首先在特征空间中使用一个新的超

参数来训练和预测目标函数，然后将所有新训练集（SV）输入到映射后的目标空间中来调整 SVM 参数，从而获得最佳超参数组合。最后将最佳超参数组合用于训练和预测文本特征向量。这两个过程是相互独立的。我们可以通过改变这两个过程来调整 SVM 参数以获得最优结果。

实验结果表明，与 SVM、GA-SVM、CBOW-SVM 等其他流行算法相比，SWSVM 在文本分类和命名实体识别任务中表现出了更好的性能。此外，在与其他经典算法相比时， SWSVM 表现出了更高的准确率、召回率和 F1 分数。

通过上述实验可以发现：对于文本分类、命名实体识别和机器翻译等任务，使用传统 SVM 方法可以获得较高的准确率；对于命名实体识别任务， SWSVM 与其他经典算法相比表现出了更好的性能；而对于机器翻译任务， SWSVM 可以获得更高的准确率、召回率和 F1 分数。因此，在本文中我们提出了基于 PSO 优化孪生支持向量机模型（SWSVM）以解决文本分类、命名实体识别和机器翻译等问题。

## 二、 SWSVM 模型的设计

SWSVM 算法的基本思想是基于孪生支持向量机（Twin Support Vector Machine, TSVM）模型的思想，将一个文本向量表示为一个支持向量和另一个文本向量的线性组合，其中两个文本向量之间的距离称为间隔。SVM 是一种线性回归模型，它能很好地拟合数据。但是， SVM 在处理非线性问题时表现并不理想。因此，我们提出了一种基于 PSO 优化 TSVM 模型的方法，它能有效地解决非线性问题。

假设一个文本数据集 T：其中文本的长度为 n，文本的维度为 n。文本中单词和单词之间的关系如下：

其中 E 和 K 分别是第 k 个单词与其在前 k 个单词中出现的概率，P 是一个词在文本中出现的次数。

由于词向量对单词之间关系和上下文信息无法有效捕获，因此本文使用 BOW 模型来提取文本特征。BOW 模型包括两个维度为 2-范数的向量（Vector）和一个参数为 r 和 w 的核函数（Kernel）。其中， var (v1,v2)是一个特征向量，表示一个词语在文本中出现的次数。Var (v2)是一个核函数，它将 var (v1)中出现过的单词进行连接。Var (v2)被认为是一个独立的向量。通过 SVM 来计算 Var (v1)和 Var (v2)之间的距离：

其中，s 是一个特征向量，l 是一个核函数，k 是一个特征向量和核函数之间的距离。在本文中，我们定义了一个关于 SVM 参数 r 和 w 的约束条件：

其中 C 为一组惩罚因子。惩罚因子 C 表示从所有训练样本中随机选取 N 个样本所需的最小惩罚金额；W 表示在每个训练样本中对 SVM 参数进行调整所需的最小权重；s 表示在每个训练样本中对 SVM 参数进行调整所需的最大权重。惩罚因子 C 和 w 均为常数。

通过以上定义，我们可以得到 SWSVM 模型如下：

其中 m 为训练样本数；w 为特征个数；m 为特征维度；c 为惩罚因子；V（C）表示对 SVM 参数进行调整所需最小惩罚金额。

在模型中，我们通过 PSO 算法来优化 TSVM 模型中每个参数 r 和 w 的值。PSO 算法是一种随机优化算法，其基本思想是利用群体智能原理来寻找最优解。为了方便起见，我们将 PSO 算法描述如下：

其中 s 表示粒子群中每个个体的适应度函数（Iterative function），本文使用 R（s）表示粒子群中每个个体在适应度函数中所占比例； vi 表示粒子群中每个个体在适应度函数中所占比例； xi 和 xj 分别代表第 i 个粒子和第 j 个粒子。

## 三、实验结果与分析

实验数据集取自 CNN-CQT，实验数据集包括 6384 个测试文本，每个文本包含 2000 个单词。每个单词有 60 维特征向量，包含三个维数为 300 和 500 的特征向量，作为训练集，另有一个 150 维的测试集，用于测试模型性能。在这里，我们使用 SWSVM 训练模型并对其进行评估。

本文使用的数据集与之前使用的数据集相比具有较大差异。我们在一组实验中使用 SWSVM，另外两组实验使用 BOW 和 SVM。在每种情况下，我们将准确率（Accuracy）和 F1 分数（F1-score）与现有的三种方法进行比较。实验结果如表 1 所示。

为了进一步探究 SWSVM 和 PSO-SWSVM 在实际应用中的有效性和可行性，我们将 SWSVM 和 PSO-SWSVM 分别与其他三种算法进行比较。在这些数据集中，我们使用混淆矩阵对损失函数进行度量，并对它们进行评估。

此外，我们还将模型与三种其他的基准模型进行了比较。可以看出，在本文中提出的方法能够有效地提高模型的性能。

参考文献：

[1]田桂丰,单志龙,廖祝华,等.基于 Faster R-CNN 深度学习的网络入侵检测模型[J].南京理工大学学报（自然科学版）.2021,(1).DOI:10.14177/j.cnki.32-1397n.2021.45.01.006.

[2]李舟军,范宇,吴贤杰.面向自然语言处理的预训练技术研究综述[J].计算机科学.2020,(3).DOI:10.11896/jsjkx.191000167.

[3]刘广灿,曹宇,许家铭,等.基于对抗正则化的自然语言推理[J].自动化学报.2019,(8).DOI:10.16383/j.aas.c190076.

**Natural language processing based on PSO**

Wu bei si

Sichuan Top Information Technology Vocational College, Sichuan Chengdu 610000

Abstract: In natural language processing (NLP), tasks such as text classification, named entity

recognition, and machine translation are trained and predicted by using text feature vectors. However, the number of training samples is usually relatively small, and as text length increases, computational time becomes increasingly expensive. Therefore, in practice, a large number of training sets need to be constructed. However, because the currently used word vector and bag of word models (Bag of Words, BOW) cannot capture the relationship and context information between words, it is difficult to predict new words or phrases. At the same time, for many problems in natural language processing (such as named entity recognition, machine translation, etc.), the traditional support vector machine (Support Vector Machine, SVM) methods usually minimize the classification error rate. However, the SVM has an obvious flaw: if the parameters are not set properly, it will cause overfitting. Particle swarm optimization (PSO) algorithm is a stochastic optimization algorithm based on the principle of swarm intelligence. The basic idea is to obtain the global optimal solution by adjusting the position and speed of each individual generation in the population. In this paper, we propose a method based on PSO optimization (Twin Support Vector Machine, SWSVM).

Key words: natural language processing; twin support vector machine; particle swarm algorithm; penalty factor