附件 3

项目编号：_____

# 大学生科技创新项目

# 申　报　书

**项目名称：** 基于神经网络架构智能寻优的敏感信息识别模型研究与应用

**项目申报人：** 何伟斌

**学校名称：** 浙江工商大学

**申报日期：** 2020-12-8

**项目类别：** 个人项目☐　　团队项目☑

**浙江省大学生科技创新活动计划（新苗人才计划）实施办公室　制**

# 填写说明

一、填写申报书前，请先查阅《浙江省大学生科技创新活动计划(新苗人才计划)实施办法》及申报通知。

二、申报书要按照要求，逐项认真填写，填写内容必须实事求是，表达明确、严谨。

三、格式要求：申报书中各项内容以 Word 文档格式填写，表格中的字体为小四号仿宋体，1.5 倍行距；表格空间不足的，可扩展。

四、申报书由所在学校审查．签署意见并加盖公章后，报送浙江省大学生科技创新活动计划（新苗人才计划）实施办公室。

## 一、项目简介

<table>
<tr><td rowspan="4">项目概况</td><td>项目名称</td><td colspan="7">基于神经网络架构智能寻优的敏感信息识别模型研究与应用</td></tr>
<tr><td>项目性质</td><td colspan="7">（）基础研究　　（√）应用基础研究</td></tr>
<tr><td>项目来源</td><td colspan="7">（）自主立题　　（√）教师指导选题</td></tr>
<tr><td>起止时间</td><td colspan="7">自 2020 年 12 月　至 2021 年 12 月</td></tr>
<tr><td colspan="2">项目状况</td><td colspan="7">（√）研发阶段　（）中试阶段　（）批量（规模）生产　（选项打√）</td></tr>
<tr><td rowspan="2">项目申报人</td><td>姓名</td><td>何伟斌</td><td>性别</td><td>男</td><td>出生年月</td><td>2000.06</td><td>入学年月</td><td>2018 年 6 月</td></tr>
<tr><td>院系专业</td><td colspan="2">计算机与信息工程学院软件工程系</td><td>联系电话</td><td colspan="2">17395710636</td><td>电子信箱</td><td>1165810311@qq.com</td></tr>
<tr><td rowspan="5">项目组主要成员</td><td>姓名</td><td colspan="2">联系电话</td><td colspan="3">院系专业</td><td>年级</td><td>具体分工</td></tr>
<tr><td>瞿立涛</td><td colspan="2">17367108273</td><td colspan="3">软件工程</td><td>大三</td><td>方案构想，实验设计</td></tr>
<tr><td>李剑霏</td><td colspan="2">17367105378</td><td colspan="3">软件工程</td><td>大三</td><td>实验设计，代码编写</td></tr>
<tr><td>黄雅岚</td><td colspan="2">15985121304</td><td colspan="3">软件工程</td><td>大三</td><td>文档整理，数据获取</td></tr>
<tr><td>贾南辉</td><td colspan="2">17395717914</td><td colspan="3">软件工程</td><td>大二</td><td>数据获取，实验设计</td></tr>
<tr><td rowspan="5">项目指导老师</td><td>姓名</td><td colspan="2">联系电话</td><td colspan="3">所在单位</td><td>职称</td><td>主要研究方向</td></tr>
<tr><td>谢波</td><td colspan="2">13515719506</td><td colspan="3">浙江工商大学计算机与信息工程学院</td><td>副教授</td><td>数据挖掘、模式识别、自然语言处理、服务计算</td></tr>
<tr><td>张华</td><td colspan="2">15057148911</td><td colspan="3">浙江工商大学计算机与信息工程学院</td><td>副教授</td><td>数据挖掘、智能计算、生物信息学、自然语言处理</td></tr>
<tr><td></td><td colspan="2"></td><td colspan="3"></td><td></td><td></td></tr>
<tr><td colspan="9">近三年成果：国家级___等奖___项，省部级___等奖___项</td></tr>
<tr><td colspan="9">近三年科研经费 约 90 万元，年均 30 万元</td></tr>
</table>

| | |
|---|---|
| 项目主要内容简介 | 本项目依托数据挖掘技术，以粒子群优化算法（Particle Swarm Optimization 下述简称 PSO）作为进化算法，指导深度神经网络架构智能寻优（NAS），寻找最优的卷积神经网络架构，完成敏感数据分类、识别的任务。本项目采用的 PSO 算法为在传统 PSO 算法的基础上**改进的算法**，提出**新型的“双重 PSO”算法**，解决了传统粒子群优化算法与深度神经网络架构搜索相结合的缺陷。<br>项目主要是通过对文本数据集进行数据挖掘、特征提取，借助该算法**开发 Web 接口**，供企业使用，企业可以通过使用本算法对自己的平台上的信息、数据进行分类与甄别，提高含有敏感信息的数据的保护级别，避免或降低出现敏感信息泄露的事件发生。<br>与同类研究相比，本项目中预测方法准确率高，灵敏度高，运行效率高，有望发表学术期刊，同时与实际应用场景相结合，具有巨大的经济效益和商业价值。 |

## 二、项目背景、目的及意义

**项目的背景、实施必要性、目的、意义**

- **背景**

随着 2018 年 4 月的《大数据安全标准化白皮书》的发布，大数据安全现在已经上升为国家战略，而数据也被视为国家的基础性战略资源受到保护。现如今各行各业都在投入大量精力收集数据、分析数据、挖掘数据，进而推动整个行业的发展。企业信息化水平的不断提高，数据共享与开放对企业发展的作用日益凸显，数据已成为重要生产要素之一。企业在产业与服务、营销支持、业务运营、风险管控、信息披露和分析决策等经营管理活动中涉及到大量的业务数据，其中可能会包含企业的商业秘密、工作秘密，以及员工的隐私信息等。若因为管理不当，造成数据泄露，则有可能造成巨大的经济损失，或在社会、法律、信用、品牌上对企业造成严重的不良影响。

然而在享受大数据带来的红利时，数据被开发共享，交叉使用，其所带来的安全问题日益凸显。2018 年 11 月 14 日，由于 ElasticSearch 并未对公司内部的数据进行严格的保护，将含有隐私数据的文本同一般的新闻文本处理，导致有心人士趁机复制和窃取所有数据，使得最终公司泄露了 8200 万条敏感数据，其中包括 5700 万个人用户

数据，2600 万企业数据。可见大数据时代下敏感数据泄露问题发生范围广、危害大。**因此，对于公司内部的数据分类识别的任务迫在眉睫**。同时，政府也非常重视，围绕数据安全，国家近年密集颁布《网络安全法》、《民法典》、《数据安全法》（征求意见稿）、《个人信息保护法》（征求意见稿）等。

在企业中大部分数据是由用户产生的，其中不乏含有用户隐私数据的信息和一些企业的商业机密，因此为避免各敏感数据泄露的事件频发，对于企业而言首要的工作就是要对在自己平台上数据进行分类分级，在敏感数据进行分类后对自己平台上的高敏感数据进行更加高等级的保护或是及时提醒用户敏感数据的泄露情况。从而进一步围绕保护对象的全生命周期进行开放、动态的数据安全治理，解决数据开放共享与敏感数据保护的矛盾与统一。

伴随近年来机器学习、深度学习的飞速发展，使用机器学习算法对敏感数据进行分类、识别，已经成为了业界主流的方案。但是传统的机器学习算法用于数据的分类分级存在着许多问题，如不支持信息量较小的信息、极度依赖于外部数据库等问题，并不能良好的实现对敏感数据的识别与分类。随着卷积神经网络（CNN）在视觉图像处理方面表现出的越来越好的优越性，在 2014 年，Yoon Kim 针对卷积神经网络的输入层做了一些变形，提出了文本分类模型 textCNN，发现卷积神经网络在文本分类中也有着非常不错的效果，然而 TextCNN 的成功，不是网络结构的成功，Yoon Kim 使用的神经网络架构是非常简单的架构，后人在他的 TextCNN 的基础上，优化了卷积神经网络的架构并且得到了良好的效果，但一个有着良好效果的卷积神经网络架构，是由人工从零开始设计的，这会消耗大量的时间和经历，而且得到的结果并非最优的。

由于粒子群算法（Partical Swarm Optimization PSO）有着操作简单、收敛速度快的优点，可以利用群体最优和个体最优的思想智能地寻找到全局最优解。许多学者就提出了利用粒子群算法寻找最优的卷积神经网络架构的方法，然而这些方法普遍存在着不能对架构和超参数同时优化、过早收敛等问题，导致最终结果并非全局最优解，且寻优过程长，需要消耗大量的计算资源等问题。**因此迫切地需要一种算法能够解决粒子群算法与卷积神经网络相结合后存在的问题**，并且利用该算法寻找到对于敏感信息分类识别任务有着最优效果的卷神经网络架构，**实现对敏感数据高准确率的识别、分类，帮助企业动态提高隐私数据的保护级别的同时降级保护成本，减少或避免因敏感数据泄露所造成的不必要损失**。

- **研究现状**

在早些年敏感数据分类技术主要依赖于支持向量机（SVM）模型，但是对于信息量较短的文本而言，由于样本所包含的单词少，并且具有歧义性，导致支持向量机（SVM）不再直接适用敏感数据分类。后续，许多学者都提出了一些巧妙的策略来构建适用于敏感信息识别的分类模型。Yih W 和 Meek C 在《Improving similarity measures for short segments of text》中提出一种基于搜索引擎（Search Engine，SE）的分类方法，但是这种方式的分类结果准确性很大程度上**依赖于搜索引擎**，并且，分类过程需要搜索引擎的参与，**耗时长，不能实现高效、快速地分类分级**。此后，王荣波在他的《基于 Wikipedia 的短文本语义相关度计算方法》文章中提到另外一种观点，则通过引入外部数据库，通过知识库挖掘出信息之间的语义、语序和挖掘出词语同义词等信息，用于辅助分类。然而，由于拓展的效果由外部知识库的质量决定，对于知识库中没有关键词情形，无法直接进行拓展，而且对于敏感信息中的词汇可能有很大一部分不存在数据库中。并且它的**计算相对复杂，计算量较大，不具备对数据快速分类的能力**。

由于神经网络算法在自然语言领域处理的优越性，深度神经网络算法、卷积神经网络算法等算法对敏感信息识别的使用成了主流趋势，结果显示卷积神经网络算法的准确度优于其他算法。这是因为卷积神经网络(Convolutional Neural Network，CNN)因具有局部感受野(Receptive field)，共享权值和子采样等特点有着更强大的自主学习能力和表达能力。正是因为如此，CNN 在机器视觉、自然语言处理领域中有很泛的应用，而其他传统的分类、识别算法，依赖于现有的自然语言处理工具容易导致处理过程中的误差累积问题。**卷积神经网络方法为文本分类提供了一种直接端到端的解决方案，**相比传统机器学习方法，它不需要开发者准备先验信息，并且能够避免复杂的人工特征工程。2014 年 KIMY 的《Convolutional neural networks for sentence classification》首次开始将卷积神经网络应用到文本分类，其网络结构分别是多层卷积网络、基于单词向量的单层卷积网络和基于字符向量的多层卷积网络。这几个方法都取得了优于传统分类方法的性能。但是 KIMY 所提出的卷积神经网络架构相对简单，在性能和结果上并非最优的，并且这些模型并不能随着一个特定的任务或是数据集，改变其网络架构从而优化模型性能与预测结果。由于卷积神经网络的结构复杂多样，决定 CNN 结构的很大一部分因素是超参数，这些超参数无法通过网络训练得到，传统 CNN 架构设计中采用的深度神经网络多通过专家手动设计，**这是一个试错的过程，高度依赖专家经**

**验，耗时耗力**。且往往手工设计的深度神经网络存在较大的冗余，如参数较多，模型较大等，因此依然存在很大的优化空间。

并且由于 CNN 结构复杂,很难确定 CNN 使用何种网络结构,使用多少层,每一层有多少个神经元才能使网络性能好等诸如此类的问题都归结到 CNN 超参数的设置上。此外，对于传统的卷积神经网络在文本分类上的应用，即便是对某一问题确定了 CNN 结构，但是该 CNN 结构未必适用于其他问题。并且由于本项目的任务是对敏感数据的分类与识别，考虑到数据集的特殊性，因此又需要大费周章的寻求新问题的最佳 CNN 结构。为解决上述的问题，近年来有学者提出了采用优化方法与 CNN 相结合，从而自动寻找到全局最优的卷积神经网络架构与超参数，以解决传统卷积神经网络存在的问题。Arash Rikhtegar 等人在《Genetic algorithm-optimised structure of convolutional neural network for face recognition applications》中提出一种用 CNN 输出层的均方误差值作为 GA 的适应度函数,采用 GA 找出 CNN 的最优结构,并且通过用 SVM 集合替换 CNN 的最后一层来提高系统的性能。Wang Bin 等人在《Deep Convolutional Neural Networks by Variable-length Particle Swarm Optimization for Image Classification》中提出了一种由计算机网络启发的新型粒子群编码策略,并用其自动搜索 CNN 的体系结构。Giovanni L.F.S 等人在《Convolutional neural net work-based PSO for lung nodule false positive reduction on CT images》中使用 PSO 算法对 CNN 中的网络超参数进行优化。但是,现有的优化方法只针对 CNN 中**少部分的超参数进行优化**,还存在一些超参数需要依据经验设定,而这些超参数在 CNN 中扮演重要的角色,**对网络的性能有一定影响，**且现有的算法并且**没有做到对架构和超参数同时优化的效果**。

**综上**，传统机器学习方法并不适用在敏感数据分类上，其在该任务上存在着许多弊端。而传统的卷积神经网络架构设计中采用多是通过专家手动设计，这是一个试错的过程，高度依赖专家经验，耗时耗力。此后的将优化算法与卷积神经网络相结合的算法只针对 CNN 中少部分的超参数进行优化，或是智能优化架构、或是智能优化超参数，并不能做到二者同时优化的效果。虽然国内外研究者在该方面已经做了很多的创新和尝试，但仍然存在着极大的进步空间。考虑到本项目的任务是对敏感数据进行分类识别的，因此需要一种利用进化算法对卷积神经网络的架构和超参数进行**同时**优化的算法，且消耗尽量少的计算资源，使得最终的网络架构能良好的适应本任务，在本任务上具有较好的性能和预测结果，从而为使用本项目提供接口的企业提供优质的服务，降低企业对数据保护的成本，降低因敏感数据泄露给企业造成的不必要的损失。

- **实施必要性**

**今年来，国家高度重视数据安全、敏感数据保护等问题。**国家密集颁布《网络安全法》、《民法典》、《数据安全法》（征求意见稿）、《个人信息保护法》（征求意见稿）等，可见在国家、政府层面对数据保护的重视程度。

**对于企业而言企业发生信息泄露事件会导致企业在公众中的威望和信任度下降，会直接使他们改变原有选择倾向。**从这里不难推断，信息泄密事件可能会使企业失去一大批已有的或者潜在的客户。因此也可以说，在数据信息的作用与地位日益显要的今天，数据信息的安全问题是关乎企业声誉、公众信任感、经济利益、生死存亡的问题，企业数据信息的安全程度将会影响企业的外部竞争力，可见数据信息泄露对社会危害性极大与大数据时代下敏感数据泄露问题的危害程度。

**因此对于公司内部的数据进行数据治理的任务迫在眉睫。**而数据治理的首要问题是敏感数据分类，而如何高效安全的对数据分类并对敏感数据的识别成为重中之重，所以为了有效、规范保护各类敏感数据，其首要问题是对数据进行分级分类，**对于企业而言需要将敏感数据进行识别、分类，将高敏感的数据加以更严格的保护**，而不含敏感数据的信息可以采用成本更低的保护方式，**在降低数据泄露事件发生频率的同时，降低数据保护成本。**

因此迫切的需要一种算法来提供此类服务。现业界在敏感数据分类方面主要采用深度神经网络进行处理，其中卷积神经网络算法在该方面存在较为优秀的效果。然而，传统深层神经网络是从零开始人工设计的，但这种方法通常会消耗大量的时间和计算资源。这是一个试错的过程，高度依赖专家经验，耗时耗力。为解决此问题近年来有学者提出利用进化算法寻找卷积神经网络架构与超参数，但是仍存在以下这些问题：

（1）**优化的超参数不全面。**现有的优化方法只针对 CNN 中少部分的超参数进行优化,还存在一些超参数需要依据经验设定,而这些超参数在 CNN 中扮演重要的角色,对网络的性能有一定影响。

（2）**不能同时优化架构与超参数。**一般来说，卷积神经网络的效果取决于两个方面:体系结构和超参数，只有当两者同时达到最优状态时，其性能、模型准确率等才能得到满足。但是现有的基于 PSO 的 CNN 网络结构搜索存在的主要问题在于没有做到二者同时优化的效果,只会单独优化参数或者是架构,本项目中使用的算法是基于优化后的 PSO 应用在 CNN 架构搜索上，可以实现二者同时优化，寻找到参数、架构都最优的模型。

（3）**需要消耗大量的计算资源。**一般来说，许多 PSO+CNN 方法选择直接方式来获得适应度，即，在经历了计算上昂贵的训练阶段之后，将在训练阶段期间不可见的验证数据集上计算真正的适应度，即使在高性能图形处理单元(GPU)上也需要数百个小时，但由于缺乏大量昂贵的计算资源，所以需要设计一些训练技巧，来缩短训练的时间。

（4）**编码策略不适用。**对于传统进化算法在神经网络架构搜索上的应用，对于架构和参数的编码主要存在两种主流的方法，一种是直接编码，一种是间接编码。然而，对于粒子群算法而言，可能不适合这类编码策略，需要重新设计。

因此本项目中提出一种改进后的粒子群优化算法与卷积神经网络相结合，智能寻找到在本项目的敏感数据分类识别任务下的最优卷积神经网络架构与超参数，**使得优化后的算法能智能寻找到准确率更高，灵敏度更高，运行效率更高的敏感信息识别、分类模型以克服传统数据治理的缺陷**，并开发对外的业务接口，旨在高效，精确，灵敏的分类数据、识别保护敏感数据，帮助企业动态提高隐私数据的保护级别的同时降级保护成本，减少或避免因敏感数据泄露所造成的不必要损失。

- **目的**

大数据技术的快速发展不断催生新的产业形态，正成为经济社会发展的新动能。与之相伴的是，数据安全风险日益成为影响产业发展。对于企业而言大数据是重要的商业资源和生产要素，数据安全治理能力已成为企业的重要竞争力，发展数字经济、加快培育发展数据要素市场，必须把保障数据安全放在突出位置。因此企业要做的第一步就是对公司内部的数据进行分类分级，这些数据可能是业务数据、用户产生的数据等等，识别出含有敏感数据的信息，将含有敏感数据的信息加以级别更高的保护，而不含敏感信息的数据可进行成本较低的保护，本项目旨在利用自主优化的算法提升数据安全治理能力的同时降低信息保护的成本。

针对上述问题及传统敏感数据分类、识别的缺陷，本项目依托自主优化改进的数据挖掘技术（基于粒子群算法的神经网络搜索），寻找最优的模型拟合，通过对文本数据集进行数据挖掘、特征提取，借助该算法开发对外接口，供企业提供更高效，精确，灵敏的技术服务，降低或避免由于隐私泄露给企业、用户带来的损失的事件发生。同时对信息共享时代下，对数据安全治理技术的提升做出贡献。

- **意义**

敏感数据分类、识别以及数据安全对信息化时代的推动具有极大意义。经对改进的算法及敏感数据识别分类模型研究后，利用该算法模型开发相关对外接口供企业使用，可为当今企业的数据安全治理提供保障，企业可以通过使用本算法对自己的平台上的信息、数据进行分类与甄别，提高含有敏感信息的数据的保护级别，避免或降低出现敏感信息泄露的事件发生。其意义可提现在以下几个方面：

**（1）社会方面**

大数据时代之下，信息共享成为形势所趋，而企业数据及用户隐私的保护成为重中之重的问题。而我们的项目可以解决该类问题，为企业的数据安全及隐私保护提供保障，消除社会上大众在信息时代对隐私泄露的恐惧，对数据安全治理的技术发展做出贡献，数据安全治理以及数据安全对信息化时代的推动具有极大意义。本项目具有很高的社会效益。

**（2）经济方面**

本项目具有较大商业价值，首先数据安全治理技术可为企业数据保护提供服务，其次敏感信息识别，及隐私保护也可为 App 开发在数据共享的需求下提供服务；其次，数据分类技术也可运用于文本分类，例如评论分类的应用、文本整理等等。可见该项目可提供较大的经济效益。具有较高的经济价值。

**（3）科技成果方面**

使用这套自主改进的基于 PSO 的 NAS 算法搜索出的模型准确率高，灵敏度高，运行效率高，有望发表学术期刊，同时与实际应用场景相结合，具有巨大的经济效益和商业价值。

## 三、项目研究方案

- **主要内容**

在本项目中，我们提出了一种基于改进后的粒子群优化算法的新算法实现卷积神经网络（CNN）架构的智能寻优，**与其他进化方法相比，该算法能够快速收敛，自动搜索最优的深度卷积神经网络结构，将该算法用于敏感数据识别、分类中**，并且设计了一种新的直接编码策略和一种速度算子，使得粒子群优化算法能够与神经网络一起优化使用，且实现深度卷积神经网络架构和超参数的共同优化的目的。本项目旨在通过改进后

的 PSO－CNN 的算法，智能寻优，找到在敏感数据分类识别任务下最好的卷积神经网络架构，从而来达到高准确率的敏感数据识别的效果，为企业提供敏感数据识别的服务，以避免或降低敏感数据泄露的事件发生。

（1） **项目开展的设计路线**如下图 1 所示：

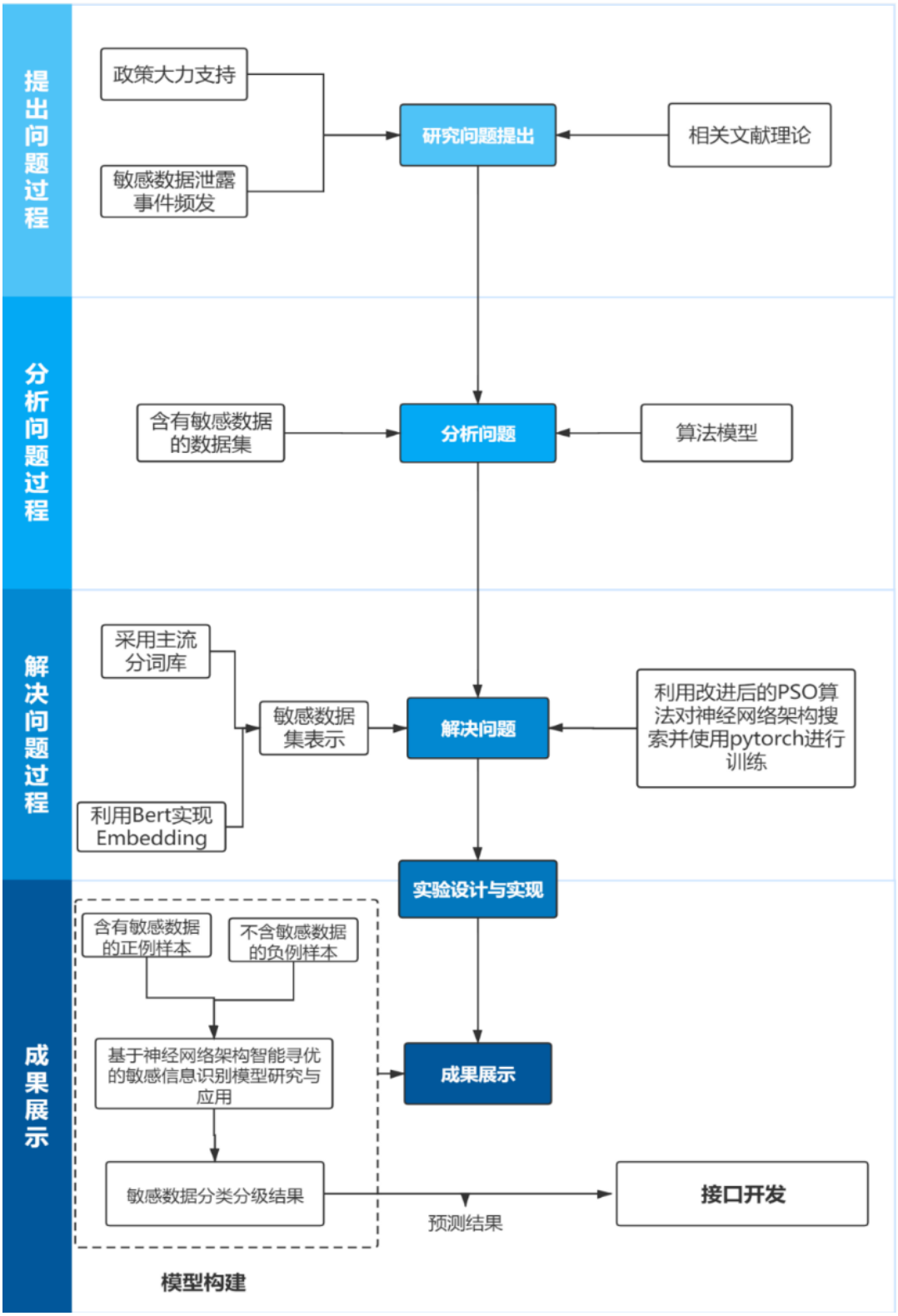


图 1 项目路线设计

**（2）项目研究的主要内容如下：**

**①基于 CNN 网络架构智能寻优的敏感数据识别模型研究**

首先团队通过对各大比赛如 kaggle、天池等比赛中相关数据集收集、利用网络爬虫技术获取相关数据集和通过 GitHub 上公开等各类数据集收集，获得了大量的数据集，以便后续模型的训练。其次，我们对于数据预处理首先利用大连理工大学提供的停用词和主流中文分词的 jieba 库对采集到的数据集进行分词，然后采用 Google 的 Bert 预训练模型对分词后的文本句子进行 Embedding，输入到卷积神经网络中。进一步，在搜寻最优神经网络架构的过程中，我们基于传统 PSO 算法（如图 2 所示）提出一种“架构优先”策略，从而解决架构与超参数同时优化的目的。

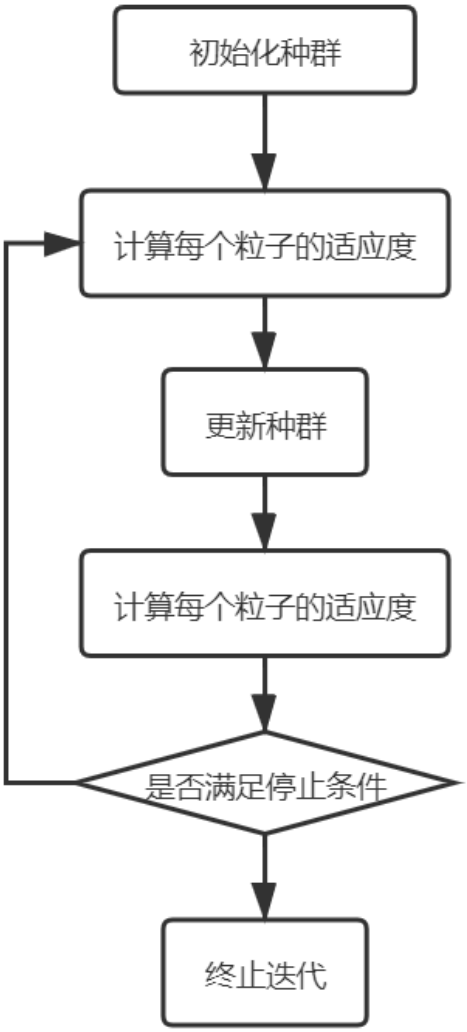


图 2 传统粒子群算法

②开发对外使用的 API 接口

本项目后期会利用训练好的模型搭建一个供企业付费使用的 API 接口，有需要的企业可以通过购买来使用本平台及调用相关 API。项目在接口开发方面，后端将采用 SpringCloud 进行分布式开发，应用 Mybatis 进行持久化层的交互，同时采用 Shiro 作为本项目的安全框架。采用前沿的技术，致力于打造一系列易用、实用的相关 API，给用户更好的体验。

- **计划目标**

（1）构建**基于神经网络架构智能寻优的数据安全敏感信息识别分类模型**。通过对文本数据集进行数据挖掘、特征提取，采用改进后的 PSO 算法智能寻找性能、准确率双高的卷积神经网络架构。最终的模型预测方法准确率高，灵敏度高，运行效率高，有望发表学术期刊，同时与实际应用场景相结合，具有较高的经济效益和商业价值。

（2）**借助该算法开发提供相关接口**，供企业使用。企业可以通过使用本算法对自己平台（尤其是对与有社交功能的平台）上的信息或是公司内部对外发送的信息进行过滤与甄别，避免或降低出现敏感信息泄露的事件发生。。

- **思路方法**

（1） **基于神经网络架构 PSO 寻优的敏感信息识别模型研究与应用**

本项目计划关于敏感信息识别模型研究提出一种新的 CNN 网络架构 PSO 寻优算法，基本流程如图 3 所示。架构寻优新算法的关键性要素组成如下：

①**实现神经网络架构和超参数的同步智能寻优**。为解决传统粒子群优化算法与卷积神经网络相结合存在的缺陷，在本项目中，提出一种类似于“种群最优”、“个体最优”的思路，即“架构优先”策略，使得最终能得到在敏感数据分类任务上的最优网络架构和超参数。

②**粒子没有大小限制**。新算法可以使用可变长度的粒子在深度卷积神经网络中搜索最优结构，没有大小限制，粒子被允许在没有上限的情况下增大尺寸。

③**设计一种新的用于计算两个粒子间差异的算子**。该算子允许具有不同网络层数

和参数的两个粒子进行比较，并允许粒子的速度在不使用实值编码方案的情况下进行更新。算子也允许我们使用几乎标准的粒子群算法进行搜索，避免使用多维粒子群算法。

至于**敏感信息的表示**，我们拟首先采用业界主流的分词工具对文本语句进行分词，然后利用 Bert 预训练模型对每个词进行 embedding，最后将获得的词向量表示输入到网络架构待优化的 CNN 网络中。

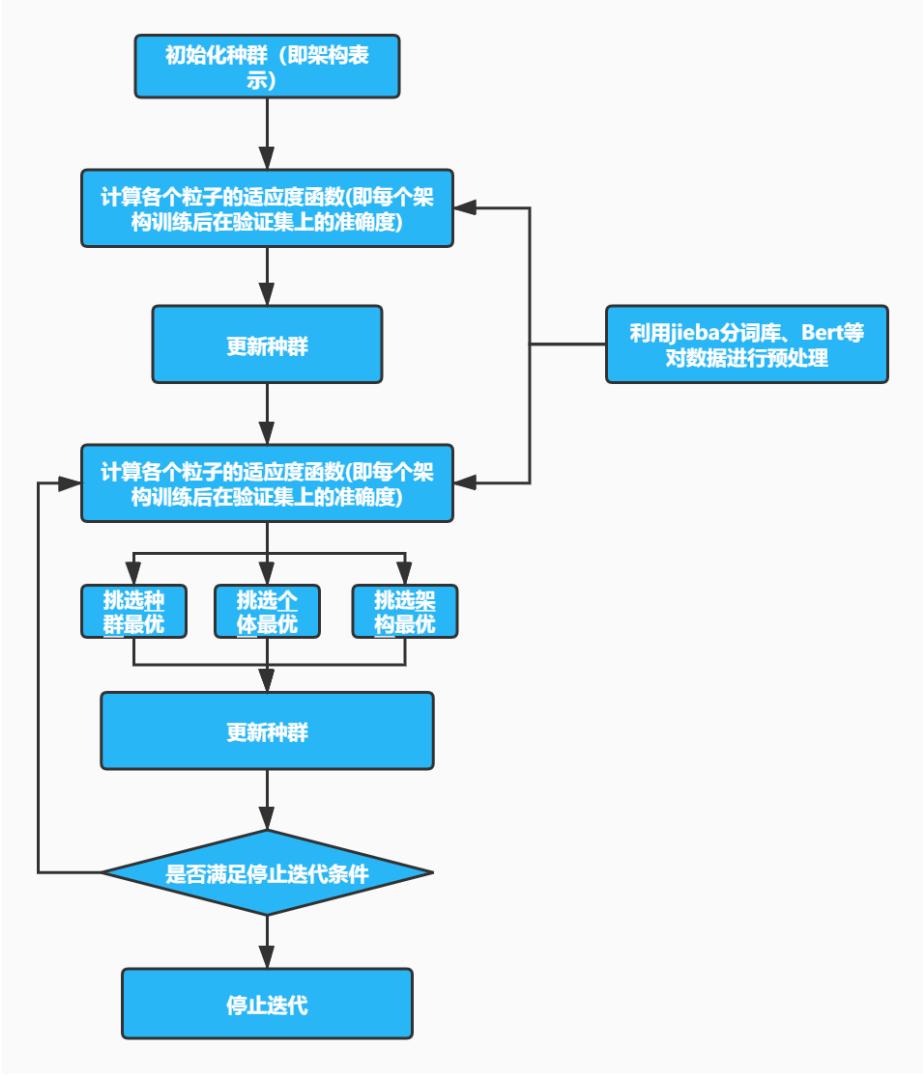


图 3　算法基本流程

**（2） 预测服务 API 开发**

基于本项目提出的算法模型开发一系列用于敏感数据分类、识别的 API 接口。API 接口的开发采用 SpringCloud 进行分布式开发，Mybatis 进行持久化层的交互，同时利用 Shiro 作为本项目的安全框架。使用前沿的技术，致力于打造一系列易用、实用的

API 接口，给用户更好的体验和保障。

➢ **进度安排**

| 起止年月 | 任务 |
| --- | --- |
| **2020 年 9 月——2020 年 10 月** | 论文与研究资料的收集、整理，现存研究成果的对比与总结。前沿特征选择算法的学习与实现。 |
| **2020 年 10 月——2021 年 11 月** | 提出解决方案，编写代码并调试 bug。 |
| **2021 年 11 月——2021 年 2 月** | 在敏感数据集上进行大量实验，不断改进方案直至提出的算法效率、准确度均达到较高的水平。 |
| **2021 年 2 月——2021 年 5 月** | 整理实验结果，开始撰写论文并尽快发表；<br>绘制 web 公共服务平台线框图，整理基础功能表单。 |
| **2021 年 5 月——2021 年 9 月** | 完成 web 公共服务平台的搭建；<br>完成论文的公开发表。 |
| **2021 年 9 月——2021 年 12 月** | 总结、补充和完善相关资料，为验收作项目总结与结题好准备。 |

## 四、项目研究条件及创新之处

➢ **所具备的基础**

**（1）前期成果**

项目组目前已基本完成全部代码，实现了基于神经网络架构智能寻优的数据安全敏感数据分类识别模型并且应用在敏感数据分类识别的数据集上，目前团队的实验结果明显优于现在主流的模型，如 DE-CNN、Text-CNN、Char-CNN 等等。

API 服务正在筹划搭建中，后续工作正在稳步推进。目前正在致力挖掘更好的数据特征量、构建更好的数据模型，来提高敏感数据分类识别的准确率。

**（2）知识条件**

团队成员有着较强的开发能力、学习能力、建模能力，且有四位的专业方向都是大数据方向。每位成员都已经熟练掌握数据科学基础、数据挖掘基础、深度学习等知识，并且团队成员的成绩均在班级前列，有着扎实的编程基础和较强的学习能力。其中三人参加了 2019 年中国大学生服务外包创新创业大赛，分别取得一等奖、二等奖、三等奖

的优秀名次；一人获得全国大学生数学建模竞赛省二等奖。约半年前，在两位专业知识能力强的指导老师的耐心细心地领导下，团队做了大量的前期准备、技能知识学习与研究分析，对“PSO 在深度卷积神经网络上的应用”已经有了比较全面的知识储备，并且认为该研究领域还存在着很大的改进空间，同时有足够的信心做好该课题。目前，我们已经学习了 Python 语言、数据挖掘（分类和回归算法、支持向量机、决策树）、机器学习、算法导论等，并且能熟练使用 Pytorch、Tensorflow 等深度学习框架和 Scipy、Sklearn 等第三方库，具备了较强的建模能力。

**（3）政策优势**

面对日益严峻的数据安全威胁，世界主要国家全面加强数据保护的立法和监管。全球共有 115 个国家和地区制定了专门的个人信息保护法，确立了个人信息收集、使用以及安全保护等数据保护规则。2018 年 5 月 25 日，被称为史上最严格数据保护法的欧盟《一般数据保护条例》（GDPR）正式实施，成为全球数据安全保护的重要标杆。我国《网络安全法》也将个人信息保护纳入网络安全保护的范畴，围绕数据保护的相关配套法规和管理标准相继出台。纵观全球，各国数据保护的相关法律法规持续升级，对企业数据安全合规提出了更高的要求。因此本项目是顺应时代发展和社会需要，我们着眼于数据安全治理的敏感数据分类识别，之后也将深挖数据安全治理领域。

**（4）队伍优势**

项目所需硬件环境可依托指导老师所在的服务计算与软件外包两大实验室，有足够的台式电脑供我们团队运行计算实验，并且所需软件均已配备安装。实验室中有投影仪、无线网络、扫描仪和打印机供我们使用。同时也配备有项目相关方面的书籍，如《数据挖掘基础教程》、《大数据与智能计算》、《深度学习》等书籍，硬件与软件条件都满足我们进行该项目的要求，能帮助我们顺利地完成任务。

- **项目风险**

**(1) 政策风险**

本项目基于正规软件平台的开发，符合国家对软件产业的相关规定，提供敏感信息识别、分类模型也符合国家的相关规定，因此本项目在政策方面不存在风险。

**(2) 技术风险**

首先，项目成员采用 Python 作为主要的数据处理与数值计算语言，利用前沿、成熟深度学习框架——Pytorch、基于 Python 的科学计算模块——Scipy 等工具进行编码开

发，由于这些技术和工具都是现如今较为成熟的工具，在近几年中不会过时，且团队成员能熟练的应用和掌握这些工具，且团队的老师有着非常丰厚的自然语言处理的经验，所以在技术上不存在较大的风险。

其次，深度神经网络架是通过多个隐层进行复杂计算去探求最佳结果的过程，需要根据计算结果的好坏进行不断地调整参数，因而存在一定不确定性，也就存在一定的技术风险。但本项目是基于他人相对成熟的实验结果所优化而来，加之，我们已经获得的初步结果看，预测效果已接近国际同等水平，若继续努力优化模型，调整参数，敏感数据的过滤效果很希望再进一步提升。因此本项目在技术方面不存在风险。

**(3) 人员风险**

人员的变更是破坏项目稳定性的重要因素。本项目的成员多数为在校软件工程专业高年级学生，拥有专业知识，自主学习能力和充沛的精力。因此本项目在人员方面不存在风险。

**(4) 资金风险**

本项目的开发成本主要来自于人力成本和在线 GPU 的使用。因开发人员均为在校学生，故人力成低。加之，在线 GPU 的使用成本也较低，所以总体成本较低。因此本项目在资金方面不存在风险。

- **项目创新点**

**（1） 提出一种改进后的 PSO 算法用于深度卷积神经网络架构的寻优**

一般来说，卷积神经网络的性能取决于两个方面:体系结构和权值，只有当两者同时达到最优状态时，卷积神经网络的性能、模型准确率等才能得到满足。但是现有的基于 PSO 的 CNN 网络结构搜索存在的主要问题在于没有做到二者同时优化的效果，只会单独优化参数或者是架构，并且考虑到本项目的任务是对敏感数据进行分类识别，因此使用的算法是基于优化后的 PSO 应用在 CNN 架构搜索上，可以实现对架构和超参数的同时优化，寻找到参数、架构在本项目所要实现的任务中表现最优的模型。

**（2） 提出一种较为新颖的速度更新算子与粒子表示策略。**

该算法使用的粒子表示策略可以使用可变长度的粒子在深度卷积神经网络中搜索最优结构，而实际上没有大小限制。粒子被允许在没有上限的情况下增大尺寸。可这个速度算子允许我们使用几乎标准的粒子群算法进行搜索，避免使用多维粒子群算法。该

粒子群算法比遗传算法收敛更快。通过将其快速收敛与搜索 CNN 架构的能力相结合，所提出的算法可以比竞争算法花费更少的时间来超越最先进的结果。

**（3）与实际应用场景相结合。**

将该算法结合实际场景应用，迎合当前政策趋势，搭建 web 端、API 接口等供给企业使用。

## 五、项目预期成果

- **知识产权成果**

  **(1)** 公开发表 1-2 篇 SCI/EI 收录论文；

  **(2)** 构建一个敏感数据识别分类的平台，并申请专利、软著等。

- **知识产权归属**

  项目研发过程中发表的文章和获得的专利均属浙江工商大学和本研究团队所有。

- **社会效益**

深度神经网络架构在数据安全治理方面的应用一直受到业界的关注。依托深度神经网络架构，通过对敏感数据进行分类分级识别，将快速促进数据安全保障体系完善。改善后的智能寻优算法能对数据进行基于内容的分类分级，而数据的分类分级是数据安全治理的核心环节。本项目的敏感数据识别模型将进一步提高对垃圾和无用数据的识别和过滤，从而推动上述领域的发展。

## 六、项目财务预算

项目将本着专款专用的原则，保证将其全部用在本项目上，并取得预计成果。

| 支出科目 | 金额 | 计算根据理由 |
|---|---|---|
| 实验材料费 | 2500 | 硬件设备与耗材购置、软件工具配置 |
| 图书资料费 | 2200 | 购买书籍 1000+其他资料 200 |
| 打印费 | 800 | 资料打印费用 |
| 调研差旅费 | 3500 | 调研及参加学术活动段差旅费和通信费 |
| 其他费用 | 1000 | 人员培训、平台域名注册费+空间费+维护费+意外情况和其他支出 |
| 总计 | 10000 | |

## 七、审核流程

| | |
|---|---|
| 承诺书 | 1．本报告中所填写的各栏目内容真实，准确。<br>2．提供验收的技术文件和资料真实、可靠，技术（或理论）成果事实存在。<br>3．提供验收的实物（样品）与所提供鉴定的技术文件和资料一致，并事实存在。<br>4．本项目的知识产权或商业秘密明晰完整，未剽窃他人成果．未侵犯他人的知识产权或商业秘密。<br>5．项目实施经费合理有效，由承担项目的学生使用，无弄虚作假行为。<br>若发生与上述承诺相违背的事实，由项目组承担全部法律责任。<br>签名（全体成员）：<br>2020 年 12 月 8 日 |
| 指导教师意见 | 提出的算法具有创新性，并已积累有关研究基础。同意推荐申报！<br>签名：谢波、张华<br>2020 年　12 月　15 日 |
| 学院审核意见 | 盖章：<br>年　月　日 |
| 学校审核意见 | （无需填写、盖章）<br>盖章：<br>年　月　日 |
| 专家组审核意见 | （无需填写、盖章）<br>签名：<br>年　月　日 |
| 省实施办公室审核意见 | （无需填写、盖章）<br>盖章：<br>年　月　日 |