Hans Journal of Data Mining 数据挖掘, 2018, 8(3), 79-88
Published Online July 2018 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2018.83009

Hans汉斯

Research on the Application of Data Mining in Soldier Selection

Chenyu Ji

Beijing Institute of Satellite Information Engineering, Beijing
Email: jichenyu2017@163.com

Received: Jun 3rd, 2018; accepted: Jun. 27th, 2018; published: Jul. 4th, 2018

Abstract

In recent years, with the increasing scale of information industry, the amount of data increased exponentially, and a large number of multi-source heterogeneous data brings a series of challenges at the same time, also brings enormous business opportunities. Therefore, data mining is raised, so that the knowledge and value of big data can be got. This paper is based on "Smart Defense" project. The most important job in the whole project is soldier's selection. Soldier selection produces a large amount of data every year, which contains multiple dimensions of candidate's characteristics. The job is totally finished by staff in the past, which has heavy workload. Therefore, we need to establish proper soldier selection model to provide the fixed intelligent analysis and auxiliary decision-making for the staff. The main content of this paper is the research of application of data mining in soldier selection, including the design of data mining model and the implementation of the model based on Python. Model design includes business understanding, data understanding, data preprocessing, model selection, evaluation and optimization. The paper focuses on the problem of imbalanced classification and logistic regression, decision tree, random forests, GBDT, and the parameters optimization and evaluation methods of the classification model. Finally we got the model which behaves best on the recall rate, F1 score and ROC_AUC. It can get better results on the training set, cross validation and test sets and has good generalization ability.

Keywords

Data Mining, Classification, Soldier Selection, Imbalanced Classification

数据挖掘在择优定兵中的应用研究

季晨雨

北京卫星信息工程研究所，北京
Email: jichenyu2017@163.com

收稿日期：2018年6月3日；录用日期：2018年6月27日；发布日期：2018年7月4日

文章引用：季晨雨. 数据挖掘在择优定兵中的应用研究[J]. 数据挖掘, 2018, 8(3): 79-88.
DOI: 10.12677/hjdm.2018.83009

摘　要

近年来，随着信息产业规模化程度的日益加深，数据量呈指数式爆炸增长，庞大数量的多源异构数据带来一系列挑战的同时，也带来了巨大的商机。于是，人们就提出了数据挖掘的概念，以便从大量的数据中发现有价值的规律和知识。本文以“智慧国防”综合应用系统为背景。在整个国防项目中，择优定兵环节是非常重要的环节，需要从所有适龄青年中挑选合适的应征兵员。每年的适龄青年都产生大量的数据，这些数据中蕴含着多个维度的属性特征，以往的择优定兵环节全程人工进行，具有很大的工作量。因此，对待定兵员数据进行数据挖掘，建立合适的择优定兵模型，为武装部工作人员提供择优定兵的智能分析和辅助决策，具有很重要的意义。本文的主要研究内容为数据挖掘在择优征兵工作中的应用，包括择优定兵问题上的数据挖掘模型设计及实现。模型设计具体包括业务理解，数据理解，数据预处理设计，模型的选择、评估与优化。重点研究了不平衡分类问题以及逻辑回归、决策树、随机森林、GBDT等分类模型，并对各个分类模型的参数优化方法、模型评估方法进行了研究。最终得到了在定兵类召回率、F1分数、ROC_AUC上均表现较好的择优定兵模型，该模型在训练集，交叉验证和测试集上均能得到较好的分类结果，具有较好的泛化能力。

关键词

数据挖掘，分类，定兵，不平衡数据分类

1. 引言

近年来，随着信息产业规模化程度的日益加深，数据量呈指数式爆炸增长，各行各业每时每刻都在产生着大量待分析的数据，数据规模从 MB、GB 级发展到 TB、PB 级甚至 EB、ZB 级。为了进一步提高数据信息的利用率，基于数据库的知识发现(Knowledge Discovery in Database，简称 KDD)和它的核心技术——数据挖掘(Data Mining)被提出。数据挖掘是从更大范围的概念，就是从大量数据中发现价值的一个处理过程。从数据挖掘中得到的价值和知识可以应用于各种领域，包括经济管理、工业控制、商业营销、项目管理以及科技创新等[1]。

本文以“智慧国防”综合应用系统为背景。在整个国防项目中，征兵业务是非常重要的部分，兵员的择优定兵环节则是重中之重。每年的待定兵员在定兵前产生大量的数据，这些数据中蕴含着兵员在多个维度的属性特征，以往的择优定兵环节全程由武装部工作人员人工进行，具有很大的工作量，费时费力，而且人工操作不可避免地会存在一些差错和局限性。因此，对待定兵员数据进行数据挖掘，建立合适的择优定兵模型，为武装部工作人员提供择优定兵的智能分析和辅助决策，具有很重要的意义。

本文的创新之处在于：

1) 设计出了一套适合兵员数据的数据预处理方案。

2) 设计并实现了专门针对定兵业务的定兵模型，该模型能够很好地对待定兵员的定兵结果进行预测，在准确率、查全率、F1 分数、ROC_AUC 评价指标上综合表现良好，具有较强的泛化能力。

3) 针对定兵类和不定兵类的数据不平衡问题，从数据、算法、评价指标三个层面提出了解决方案。

2. 数据挖掘介绍

数据挖掘是一个知识发现的过程，其流程可以由 CRISP-DM 模型来表示。CRISP-DM 模型即“跨行业数据挖掘标准流程”，由欧洲数据挖掘系统开发厂商和数据挖掘系统应用组织协会一起制订，总结了数据挖掘的流程，具有一定的代表性，得到了全行业的采用。

它将一个完整的数据挖掘周期分为六个阶段，见图 1，每个阶段分别对应不同的处理流程。这六个阶段包括业务理解(Business Understanding)、数据理解(Data Understanding)、数据准备(Data Preparation)、模型创建(Modeling)、模型评估(Evaluation)以及结果部署(Deployment) [2]。

其中，在业务理解阶段，集中于从业务的角度理解项目的目标和需求，将这种业务知识转换成对应的数据挖掘问题定义，并为达到目标建立初步的计划。

数据理解，需要完成采集原始数据，透视数据分布，确定挖掘任务，明确目标属性，评估数据质量。

数据准备阶段，需要完成数据清洗、转换、集成，归约，经过数据预处理，将数据变为可以直接用于挖掘模型的形式，数据准备阶段是整个数据挖掘中非常重要且耗时的一个阶段，该阶段的工作量能达到整个数据挖掘过程的 70%左右，数据预处理后的数据质量对模型的好坏和挖掘结果有着很大的影响。

模型创建阶段，需要采用各种模型建立的方法，按照挖掘的业务场景和数据集特点进行建模，并对建好的模型进行参数优化，将模型调整到最好的性能。

模型评估需要结合业务场景和挖掘目标，确定评估方法，按照选好的模型的评估方法对建好的模型进行评估，如分类模型的准确率，召回率等。

数据挖掘结果部署，需要产生数据挖掘报告，并对整个挖掘流程进行回顾，对挖掘结果进行保存和可视化展示及分析总结。

数据挖掘模型可以分为描述性模型和预测性模型。描述性模型是指对数据中的模式或关系进行识别，属于无监督学习，包括聚类、关联规则等；预测性模型是指利用已知结果对新的数据样本进行预测，属于有监督学习，包括分类、回归等。另外，在解决预测的数据挖掘任务时，可以先对样本数据使用描述型模型，以便发现数据中的关系，对数据有更好的认识。

3. 择优定兵模型设计

3.1. 业务理解

由于女兵的征兵工作与男兵的征兵工作有所不同，且女兵需要征集的人数较少，本文仅对男兵的择优定兵进行分析和研究。

人工定兵的流程是：每年征兵工作开始后，由市级、县级武装部层层布置征兵任务，基层武装部通过全国征兵网、公安部门收集所有适龄男青年的信息，并通过摸底对数据进行更新和完善，督促符合当年应征报名条件的适龄青年在网上进行应征报名。应征青年经过目测初检、体格检查、政治考核后成为双合格青年，也就是最后定兵的待定兵员。定兵的兵员择优过程指的就是将双合格青年根据评判标准进行综合考虑，按定兵的优先顺序进行排序，再根据当年的定兵指标，确定当年的定兵人选。定兵流程图见图 2。

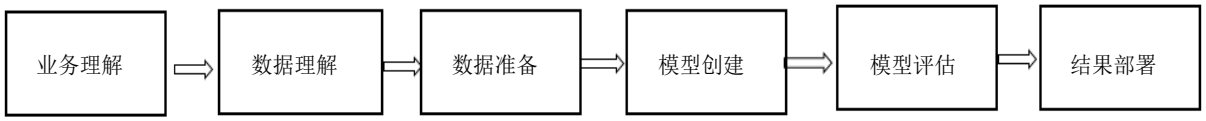

Figure 1. Data mining process
图 1. 数据挖掘流程

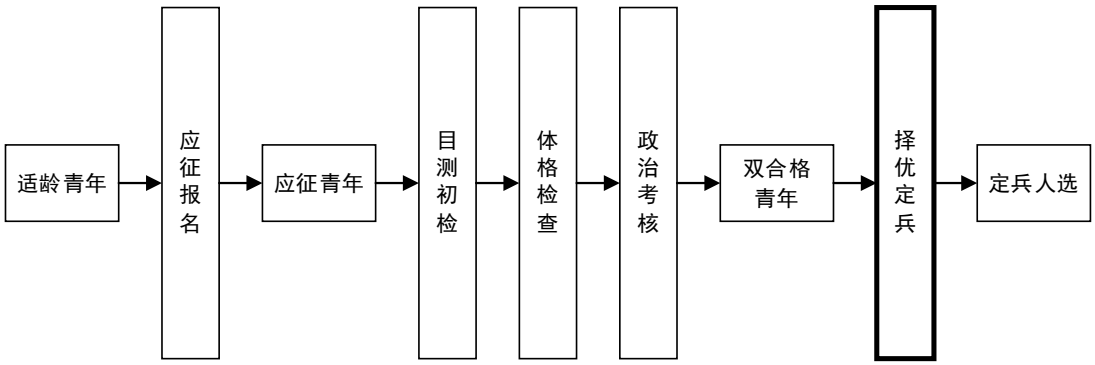

Figure 2. The flowchart of soldiers' selection
图 2. 定兵流程图

本文将数据挖掘应用在征兵工作中，通过分类模型，将应征青年分到定兵类和不定兵类，为定兵工作人员提供定兵的辅助决策。通过不断的优化模型，提高定兵类的召回率，如果模型的定兵类召回率可以达到 100%，即保证通过模型分到不定兵类里的应征青年在实际定兵时一定不会被选中，则可以直接淘汰掉被分到不定兵类的应征青年，从而减少了体检，政审和定兵的工作量，大大提高了定兵工作的效率。

一个好的择优定兵模型可以为武装部工作人员提供定兵的智能分析和辅助决策，减少工作量，提高效率，减少人为操作的差错。模型经过不断的训练和调优，可以达到较高的准确率，再将每年退兵人员的退兵原因进行收集整理，并添加到模型中，如适当增加退兵原因对应属性在整个模型中的重要程度，可以有效减少退兵率。

3.2. 数据理解

征兵工作中的数据来源较多，包括公安部门、医院、全国征兵网、学校、居民委员会、武装部走访调查等等，且数据的存储格式多样。例如，从公安部门导入的公民基本信息，是以光盘介质存储的 excel 表格，需要人工导入到数据库中；存储在智慧国防业务系统中的数据以关系型数据库 Mysql 存储，智慧国防大数据平台中的数据又存储在非关系型数据库 HBase 中。兵员数据的数据类型既包括了数值型的结构化数据，如年龄、身高、体重等，又包括了文本类型的非结构化数据，如走访调查过程中记录的一些描述性文字。

3.3. 数据准备

数据准备要完成数据清洗，数据集成，数据转换和数据归约等工作，最终将数据处理成能够满足数据挖掘模型需要的数据集形式[3]。

数据清洗时通过处理缺失值，光滑噪声数据，识别或删除离群点，解决数据的不一致性。在择优定兵的数据集中，存在一些属性值缺失的情况，比如有的待定兵员学历为高中及以下，在专业名称和专业代码属性上的值就为空，在数据清洗时可以通过将缺失值替换为 0 的方法来处理。

数据集成用来处理来自多个数据源的数据，同一概念的属性在不同的数据库中可能具有不同的名字，导致不一致性和冗余。

数据转换的目的是对数据进行规范化处理、连续变量的离散化以及变量属性的构造，将数据转化为适当的形式，以满足数据挖掘任务及算法的需要。包括 1、平滑 2、聚集 3、数据概化 4、规范化，包括最小－最大规范化、零－均值规范化、小数定标规范化等 5、属性构造。在待定兵员的属性中，有一些属性的值为非数值型，例如政治面貌这一属性，属性值包括中共党员，共青团员和群众，将该属性值进

行数值化编码，中共党员替换为1，共青团员替换为2，群众替换为3，则一方面减少了数据存储所占空间，另一方面也便于模型对该属性值的分析利用。

数据归约目的是获得比原始数据小但不破坏数据完整性的挖掘数据集，该数据集可以得到与原始数据相同的挖掘结果。数据归约的方法：1) 数据立方体聚集：把聚集的方法用于数据立方体；2) 维归约：检测并删除不相关、弱相关或冗余属性；3) 数据压缩：选择正确的编码压缩数据集；4) 数值压缩：用较小的数据表示数据，或采用较短的数据单位、数据模型代表数据；5) 离散化和概念分层生成：使连续的数据离散化，用确定的有限个区段值代替原始值，概念分层是指用较高层次的概念替换低层次的概念，以此来减少取值个数。

3.4. 模型选择

数据挖掘模型分为描述性模型(关联规则分析、聚类算法等)和预测模型(回归和分类)两种。兵员择优定兵问题属于预测任务，即通过对样本数据(历史数据)的输入值和输出值关联性的学习，得到预测模型，再利用该模型对未来的输入值进行输出值预测。分类和回归作为预测模型，其区别在于，分类适合预测类别标号，即预测目标属性是离散的、无序的；而回归则适合用于建立连续型函数模型，即预测目标属性是连续的值。将待定兵员分为定兵和不定兵两个类别属于分类问题。

常用的分类算法包括：

决策树算法自上而下的树状结构，生成一系列规则进行分类。其中的C4.5算法使用信息增益率来选择属性，在树构造过程中进行剪枝，并且能够对不完整数据进行处理。

随机森林算法是用随机的方式建立一个森林，森林里面有很多决策树组成，且每一棵决策树之间没有关联。对于每个需要分类的样本，随机森林中的每一棵决策树分别对该样本所属的类进行判断，然后采取投票，哪个类被选择最多，就预测这个样本属于哪一类[4]。

逻辑回归算法适合解决二分类问题。它在线性回归基础上引入了sigmod函数，它的取值在[0,1]之间，代表属于某一类的概率。

支持向量机算法，简称SVM，是一种监督式学习的方法中。支持向量机将向量映射到一个更高维的空间里，尝试找到一个最大间隔超平面。

K最近邻分类算法(k-Nearest Neighbor, kNN)，是最简单的分类算法之一。该方法的思路是：在特征空间中，如果一个样本的k个最相似样本多数属于某一个类别，那么也将该样本分到这一类别。

3.5. 模型评估与模型优化

评价一个模型的好坏，需要看它能否在评价指标上达到可接受的水平；是否有较强的泛化能力，能够适应更广泛的数据集。

训练有监督的机器学习模型，比如分类模型，需要将数据集划分为训练集、验证集和测试集，以选出效果最好的，泛化能力最佳的模型。

训练集用来拟合模型，通过设置分类器的参数，训练分类模型。

验证集用来在训练集训练出的多个模型中找出效果最佳的模型。使用各个模型对验证集数据进行预测，并记录模型的准确率、召回率等，按照我们的分类评价指标选出效果最佳的模型所对应的参数，作为模型最终的参数。

测试集是用来衡量选出的最优模型的性能和分类泛化能力。通过训练集和验证集得出最优模型后，使用测试集进行模型预测，得到的分类结果与评价指标，作为评估模型性能的依据。测试数据绝对不能用来训练模型，否则，只能证明训练好的模型能够拟合原始数据，而不能保证模型在新样本上有很好的

表现，因为模型很可能出现了过拟合。

还有一种办法是将数据集分成训练集和测试集，然后对训练集采取 k-fold 交叉验证，k 可以取 5、10 等等。即将训练集分成 k 份，每次都用其中的 k − 1 份来训练模型，剩下的一份用来验证训练出的模型，并记录分类的结果和准确率、召回率等相关评价指标，然后再从这 k 份中取另一份作为验证集，剩下的 k − 1 份做训练集，直到所有的 k 份都做过验证集则交叉验证结束。计算这 k 个模型的准确率均值，并以准确率最高的模型使用的超参数作为最终模型的超参数[5]。

对于二分类器，分类结果可以分为以下四种，由这四个指标组成了混淆矩阵。

1) 真正类(True Positive, TP)：被模型预测为正类的正样本。

2) 假正类(False Positive, FP)：被模型预测为正类的负样本。

3) 假负类(False Negative, FN)：被模型预测为负类的正样本。

4) 真负类(True Negative, TN)：被模型预测为负类的负样本。

分类模型的评价指标包括：

1) 正确率(Precision)，TP/(TP + FP)，给出的是预测为正类的样本中实际为正样本的比例。

2) 召回率(Recall)，TP/(TP + FN)，给出的是预测为正类的真实正样本占所有真实正样本的比例。

3) 准确率(accuracy)，(TP + TN)/(P + N)，即模型预测正确的样本占所有样本的比例。

F1 分数同时考虑了分类模型的准确率和召回率，可以看作是模型准确率和召回率的一种加权平均，F1 分数的分布在 0~1 之间。

$$\mathrm{F1} = \frac{2Precision * Recall}{Presicion + Recall}$$

ROC 曲线是一系列 threshold 下的(FPR, TPR)数值点的连线。

其中，

$$\mathrm{FPR} = \frac{FP}{N}, \mathrm{TPR} = \frac{TP}{P}$$

我们的目标是较小的 FPR 以及较大的 TPR，所以 ROC 曲线越接近最上角，该分类器的性能越好。

AUC (Area Under Curve) 被定义为 ROC 曲线下的面积，显然这个面积的数值不会大于 1。又由于 ROC 曲线一般都处于 y = x 这条直线的上方，所以 AUC 的取值范围一般在 0.5 和 1 之间。使用 AUC 值作为评价标准是因为很多时候 ROC 曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应 AUC 更大的分类器效果更好。

如果用完全没有参与模型训练的测试集来测试模型，能够得到较好的分类结果，就说明模型具有较好的泛化能力，也说明所选择的训练样本具备一定的代表性，能够体现整个数据集的整体特性。

可以采用以下几种方法提高分类模型的性能[6]。

1) 增加更多数据。

数据量越多，训练出的模型越具有普适性，越能代表数据的普遍特性，更多的数据允许数据进行更多的“自我表达”，当数据多到能对几乎整个样本空间进行充分覆盖，从而减弱对理论和模型的依赖时，这样的数据就足够“大”了。较大的数据量也可以防止过拟合现象发生。

2) 特征选择。

特征选择是指选择获得相应模型和算法最好性能的特征集，工程上常用的方法有以下：

① 计算每一个特征与因变量的相关性：工程上常用的手段有计算皮尔逊系数和互信息系数，皮尔逊系数只能衡量线性相关性而互信息系数能够很好地度量各种相关性，但互信息系数的计算相对复杂一些，

得到相关性之后就可以通过排序选择特征了；

② 构建单个特征的模型，通过模型的准确性为特征排序，借此来选择特征；

③ 通过 L1 正则项来选择特征：L1 正则方法具有稀疏解的特性，因此天然具备特征选择的特性，但是要注意，L1 没有选到的特征不代表不重要，原因是两个具有高相关性的特征可能只保留了一个，如果要确定哪个特征重要应再通过 L2 正则方法交叉检验；

④ 训练能够对特征打分的预选模型：Random Forest 和 Logistic Regression 等都能对模型的特征打分，通过打分获得相关性后再训练最终模型；

⑤ 通过特征组合后再来选择特征：如对用户 id 和用户特征组合来获得较大的特征集再来选择特征，这种做法在推荐系统和广告系统中比较常见，这也是所谓亿级甚至十亿级特征的主要来源，原因是用户数据比较稀疏，组合特征能够同时兼顾全局模型和个性化模型。

3) 使用多种算法。

使用正确的机器学习算法是获得更高准确率的理想方法。最终结论的得出来自于经验和不断尝试。有时候，有些算法比其他算法更适合特定类型的数据。因此，我们应该构建所有有关的模型，并检测其表现，再从中选取相对较优的模型。比如，可以尝试偏最小二乘回归算法，该算法可以很好地解决自变量之间的多重共线性问题。

4) 集成模型。

集成模型是把多个弱模型的结果组合在一起，以获得更好的结果。它可以通过 Bagging (Bootstrap Aggregating)和 Boosting 来实现。

5) 交叉验证。

交叉验证是将原始数据进行分组，一部分作为训练集，另一部分作为验证集，首先用训练集对模型进行训练，再利用验证集来测试训练集得到的模型。如十折交叉验证，将数据集分成十份，轮流将其中 9 份作为训练数据，1 份作为测试数据，进行试验，每次试验都会得出相应的正确率，10 次结果的正确率的平均值作为对算法精度的估计，一般还需要进行多次 10 折交叉验证，再求其均值，作为对算法准确性的估计。

在本文中，由于数据量的采集存在一定限制，只能在尽可能的情况下增加数据量，避免训练出的模型存在过拟合现象。

在兵员择优定兵的场景下，将每个小的属性视为一个特征，建立逻辑回归模型，如现实表现方面中的几个属性 1、家庭条件 2、家教(严格)；3、吃苦耐劳程度；4、学校成绩；5、尊敬师长；6、有上进心。通过特征选择，选出对最终排序结果贡献不大的特征，可以简化模型，减少收集数据时需要收集的属性值，减少工作量。

采用多种分类模型如逻辑回归模型、kNN 模型、决策树模型、SVM 模型、随机森林模型、GBDT 模型及 VotingClassifier 集成分类模型建立定兵模型，并对比分类结果，再从中选择在定兵类召回率、F1 分数、ROC_AUC 等评价指标上表现最优的模型作为最后的定兵模型。

4. 不平衡数据分类问题

不平衡数据分类是分类问题中比较特殊的问题，主要特点是样本类分布不平衡。在不平衡的二分类问题中，表现为其中一类的学习样本远多于另一类的样本。

在实际应用中，有许多分类问题，属于不平衡分类问题，比如欺诈识别，入侵检测，医疗诊断、客户流失、广告点击预测等等。

在入选人员的定兵问题中，2016 年应征青年共 6706 人，其中包括 6551 名不定兵，只包括 155 个人

最后被定兵，属于典型的不平衡分类问题。

解决不平衡数据分类问题，可以从数据、算法、评价指标三个层面着手[7]。

1) 从数据的角度：通过改变原始数据集的分布，采用过采样或欠采样，即增加少数类样本或减少多数类样本，使不平衡数据集的正负类样本数达到平衡。

2) 在算法上：修改已有的分类器，使之适应不平衡数据的特征。主要包括代价敏感分类器，集成学习等方法。其中，代价敏感分类器对少数类样本和多数类样本分类错误的代价区别开来，将少数类错误地分到多数类将付出更大的代价，例如惩罚型的支持向量机 SVM。集成学习是在训练集上训练多个分类模型，预测时根据每个分类器的分类结果进行组合，得到最终的预测结果。常用的组合分类方法，包括 Bagging，Boosting 等。集成分类模型的结合策略包括平均法和投票法，其中，Voting Classifier 集成模型通过加权投票方法对基分类器进行集成。

3) 从评价指标上，对于一般的分类模型可以利用模型的准确率进行评估。由于不平衡数据集的特点，常用的分类评价指标准确率已经不能正确评价分类模型的好坏，因为对于 6706 个样本，其中只有 155 个定兵类，其余的 6551 个样本都是不定兵类，即使将所有样本都分到不定兵类，模型的准确率也有 97.69%，但这个分类模型是没有意义的。分类的目标是尽量使所有的实际被定兵样本被分到定兵类，尽量减少实际不定兵样本被分到不定兵类。因此可以采用召回率，给出预测为定兵类的定兵类样本占所有真实被定兵样本的比例，如果可以得到比较高的召回率，就代表很少的定兵类样本被错分到不定兵类，这也是我们想要的结果。还可以采用 F1 分数，F1 分数同时考虑了分类模型的准确率和召回率，适合作为不平衡数据分类问题的评价指标。

还可以采用 ROC 曲线下的面积 AUC 作为评价指标。因为 ROC 曲线有一个很好的性能，当测试集中的正负样本的分布变化时，ROC 曲线能够保持不变，因此适合作为不平衡数据集分类时的评价指标。

5. 实验结果与分析

原数据集样本个数 6706，其中包括 155 个属于定兵类的样本和 6551 个属于非定兵类的样本。将数据集按 0.75:0.25 的比例拆分成训练集和测试集，并固定随机数种子，便于比较各个分类模型的性能。训练集样本数共 5029，其中包括 128 个属于定兵类的样本和 4901 个属于非定兵类的样本。

为了解决应征兵员数据定兵类和不定兵类不平衡的问题，我们先对训练集进行 SMOTE 过采样。经过 SMOTE 过采样，训练集中正负样本数均为 4901，总的训练集样本个数为 9802。测试集样本总数 1677，其中包括 27 个属于定兵类的样本和 1650 个属于非定兵类的样本。

对训练集采取 5-fold 交叉验证，并以准确率最高的模型使用的超参数作为最终模型的超参数。将训练并验证好的模型应用在测试集上，以得到的分类结果作为分类模型性能的比较依据。

考虑到定兵数据集的不平衡性，常用的分类评价指标准确率已经不能正确评价分类模型的好坏，选择定兵类的召回率、F1 分数和 ROC_AUC 作为定兵模型的评价指标，三个评价指标均是值越大，分类模型的性能越好。

分别利用逻辑回归模型、kNN 模型、决策树模型、SVM 模型、随机森林模型、GBDT 模型及 VotingClassifier 集成分类模型得到的分类结果如下表 1 所示。

对设计好的模型进行基于 Python 的模型实现，包括逻辑回归、k 最近邻模型、决策树模型、支持向量机等基本分类器和随机森林、GBDT、VotingClassifier 等集成分类器，并进行模型参数的调优。最后经过比较，选择在定兵类召回率，F1 分数和 ROC_AUC 上均表现较好的集成分类模型 VotingClassifier 作为最后的择优定兵模型。

VotingClassifier 定兵模型的分类结果混淆矩阵见图 3。在测试集中，共有 22 个定兵类样本，1650 个

不定兵类样本。利用 VotingClassifier 定兵模型对测试集样本的定兵结果进行预测，全部定兵类样本被正确预测，1650 个不定兵类样本中有 1510 个样本被正确预测，模型在定兵类召回率和准确率上均表现良好。

VotingClassifier 定兵模型的分类结果 ROC 曲线见图 4，可以看出 ROC 曲线下的面积即 ROC_AUC 接近于 1，证明模型的分类性能良好。

6. 总结

本文完成了数据挖掘在择优定兵中的应用研究，包括择优定兵模型设计及实现。模型设计具体包括业务理解，数据理解，数据预处理设计，模型的选择、评估与优化。重点研究了不平衡分类问题以及逻辑回归、决策树、随机森林、GBDT 等分类模型，并对各个分类模型的参数优化方法进行了研究，对模型的评估方法进行了设计。经过对各分类模型的综合比较，得出最终的择优定兵分类模型。

Table 1. Explaining variable
表 1. 预测变量

模型	准确率	定兵类召回率	F1 分数	ROC_AUC
逻辑回归	0.898032200358	0.68	0.24	0.858949297495
kNN	0.843768634466	0.75	0.19	0.842638973732
决策树	0.966607036374	0.55	0.44	0.748533903482
SVM	0.855694692904	0.80	0.21	0.873457544288
随机森林	0.955277280859	0.57	0.38	0.922755039707
GBDT	0.976744186047	0.45	0.48	0.919525045816
VotingClassifier	0.916267942584	1.0	0.24	0.983856749311

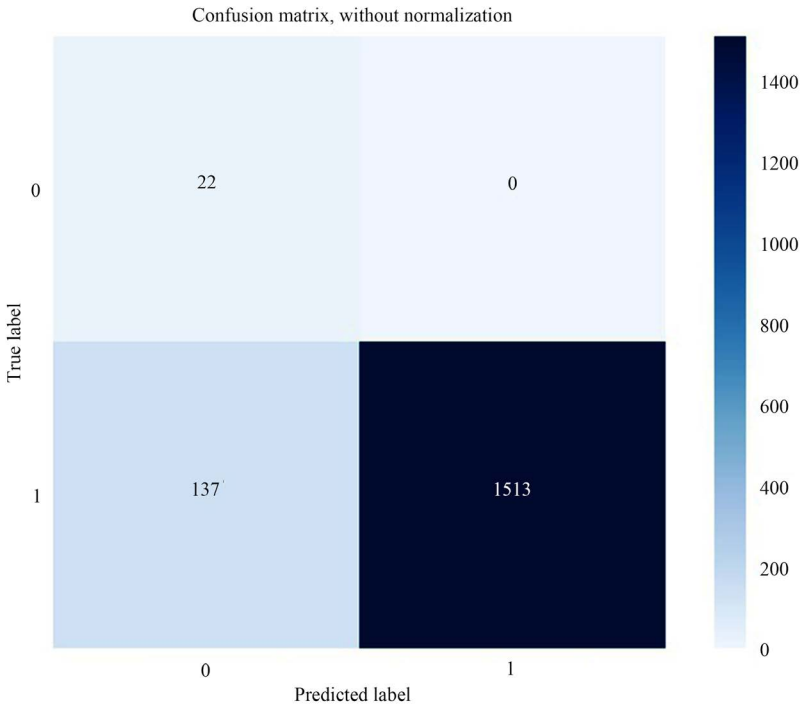

Figure 3. The confusion matrix of Voting Classifier
图 3. Voting Classifier 模型混淆矩阵

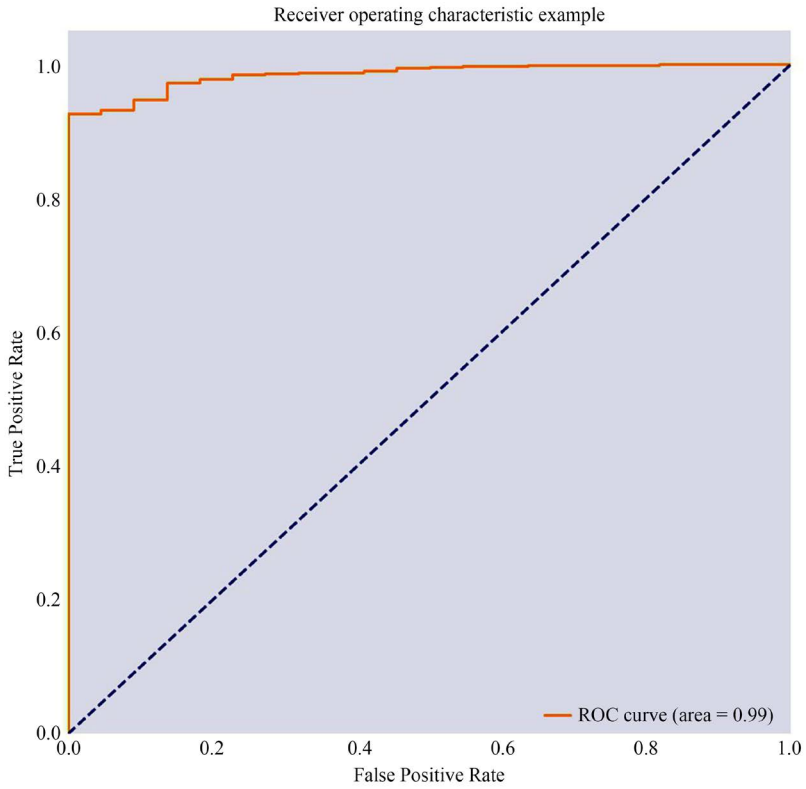

Figure 4. The ROC of Voting Classifier
图 4. Voting Classifier 模型 ROC 曲线

由于得到的分类模型在训练集，交叉验证和测试集上均能得到较好的分类结果，说明模型具有较好的泛化能力，也说明所选择的训练样本具备一定的代表性，能够体现整个数据集的整体特性。

那么，在以后的定兵工作中，就可以根据当年兵员数据的相关属性值，对该兵员的定兵结果进行较为准确的预测，提高定兵工作的效率。如果分类结果中，定兵类的召回率可以达到 1.0，完全可以将被分到不定兵类的兵员淘汰，而无需对他们进行接下来的政审、体检及走访调查，可以大大减少征兵工作的工作量，节省人力资源。因此，本文提出的择优定兵模型具有一定的研究意义和实用价值。

参考文献

[1] 陈良臣. 大数据挖掘与分析的关键技术研究[J]. 数字技术与应用, 2015(11): 93.

[2] Han, J. and Kamber, M. (2011) Data Mining: Concepts and Techniques. 2nd Edition, Morgan Kaufmann, Burlington, 1-18.

[3] PANG-NINGTAN, MICHAELSTEINBACH, VIPINKUMAR. 数据挖掘导论: 完整版[M]. 北京: 人民邮电出版社, 2011.

[4] 马骊. 随机森林算法的优化改进研究[D]: [硕士学位论文]. 广州: 暨南大学, 2016.

[5] Lobo, J.M., Jiménez-Valverde, A. and Real, R. (2008) AUC: A Misleading Measure of the Performance of Predictive Distribution Models. Global Ecology & Biogeography, 17, 145-151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

[6] 孙科. 基于 Spark 的机器学习应用框架研究与实现[D]: [硕士学位论文]. 上海: 上海交通大学, 2015.

[7] 翟云, 杨炳儒, 曲武. 不平衡类数据挖掘研究综述[J]. 计算机科学, 2010, 37(10): 27-32.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2163-145X，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：hjdm@hanspub.org