AUTOGETS：基于知识的自动生成文本合成以提高文本分类

Chenhao Xue
Department of Engineering Science
University of Oxford
UK
chenhao.xue@eng.ox.ac.uk

Yuanzhe Jin
Department of Engineering Science
University of Oxford
UK
yuanzhe.jin@eng.ox.ac.uk

Adrian Carrasco-Revilla
FabLab, Inetum
Madrid
Spain
adrian.carrasco@inetum.com

Joyraj Chakraborty
Department of Engineering Science
University of Oxford
UK
joyraj.chakraborty@eng.ox.ac.uk

Min Chen
Department of Engineering Science
University of Oxford
UK
min.chen@eng.ox.ac.uk

ABSTRACT

在为现实世界应用开发文本分类模型时，一个主要挑战是难以为所有文本类别收集到足够的数据。在这项工作中，我们通过利用大型语言模型（LLMs）生成合成数据来应对这一挑战，并使用这些数据来提高模型的性能，而无需等待更多的真实数据被收集和标注。由于 LLM 会根据不同的输入示例生成不同的合成数据，我们制定了一个自动化的工作流程，搜索那些能够生成对提升模型更“有效”合成数据的输入示例。我们通过一系列详尽的实验研究了三种搜索策略，并利用实验结果设计了一种集成算法，根据类别的特征选择搜索策略。进一步的实验表明，这种集成方法比我们自动化工作流程中的每个单独策略在使用 LLM 改进分类模型方面更为有效。

Keywords Text Classification · Ensemble Algorithm · Synthetic Data · Data Augmentation · Large Language Model · Knowledge Map · Optimization

在文本分类的工业应用中，类别通常根据语义分组以及组织功能进行定义。开发用于此类应用的稳健文本分类模型的关键障碍包括（i）明显不平衡的类别大小、某些类别的数据稀缺，以及（ii）由于组织变更导致的分类方案变化。一组例子是不同公司和组织中的自动化工单系统，用于处理用户的消息并将其分发到不同的服务，例如，IT 问题、建筑问题、运营事故和服务请求。

如图 1 所示，模型最初是在一组标记过的票据上进行训练的。分类错误需要手动分类和重新分发，从而导致延迟和成本 [1] 。随着各个组织内部的逐渐变化，某些语义的信息可能会变得不再频繁，而其他语义的信息（包括新的语义）则变得更加频繁。随着时间的推移，模型的性能会下降 [2] 。一方面，这样的模型需要定期加以改进。另一方面，识别模型性能恶化通常没有充足的训练数据来改进模型。

合成数据已被用于克服现实世界数据的局限性，解决数据稀缺性、敏感性或收集成本问题，在许多领域中，如计算机视觉和自然语言处理、医学成像、自动驾驶系统、金融和网络安全等。在这项工作中，我们专注于提供一种新颖且具有成本效益的解决方案，以改进票务系统中短消息的分类模型。

特别是，我们进行了大量实验来研究寻找消息示例的方法，以便让大规模语言模型（LLMs）生成合成数据来改进模型。基于我们对实验结果的分析，我们提出了一种新算法，用于通过分配计算资源（例如，搜索示例、生成合成数据、重新训练模型和测试模型）来控制模型改进的工作流程，并依据系统探索算法空间获得的知识识别有效的消息示例。我们还使用其他数据集和数据合成方法验证了我们的发现。

1 相关工作

合成数据在许多领域的各种数据科学任务中被越来越多地使用 [3, 4] 。自助法 [5, 6] 是早期的数据合成方法之一。它对原始数据进行重新采样以模拟所需的分布并提高模型性能 [7] 。为了克服数据集不平衡的缺点，

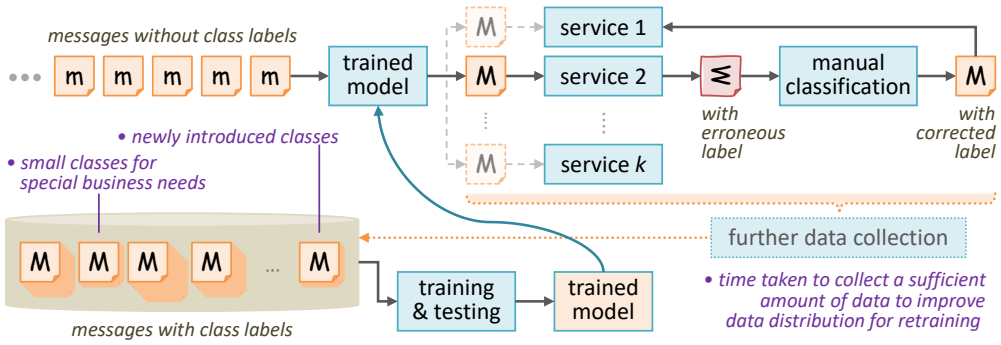


Figure 1: 在工业票务系统中开发和部署分类模型的工作流程，以及影响模型性能的主要障碍。

合成数据与常用方法结合使用，以过采样少数类和欠采样多数类 [8] 。如今，数据增强包括一系列技术，这些技术转换数据集中现有的数据以生成具有所需属性的合成数据，例如增加多样性、改变分布、补全缺失数据等 [9] 。

机器学习模型，如生成对抗网络（GANs）[10] ，被用于生成具有高度真实感和复杂性的合成数据。研究表明，基于 GAN 生成数据训练的模型通常可以与基于真实数据训练的模型媲美 [11, 12] 。例如，Frid-Adar 等人 [13] 使用 GAN 生成的图像增强了肝病灶的诊断。Yale 等人 [14] 使用 GAN 生成的电子健康记录在 ICU 病人预测上展示出可比的性能。Croce 等人 [15] 证明了其在生成用于 NLP 任务的逼真文本方面的有效性。He 等人 [16] 探讨了面向特定任务的文本生成。然而，用于文本分类的 GAN 生成数据通常缺乏语义上的连贯性和特定任务的相关性 [17] 。

最近在大型语言模型（LLMs）方面的进展，例如 GPT-2 [18] ，提供了克服这些限制的新方法。LLMs 在少样本和零样本学习中表现出色 [19, 20] ，能够适应未见过的任务并生成上下文相关的数据，从而提高模型的鲁棒性。Yoo 等人的 GPT-3Mix [21] 展示了 LLMs 通过精心编制的提示工程生成多样而有效的文本分类合成数据的能力。提示优化策略显示，精心设计输入提示可以显著影响生成数据的质量 [22] 。用于识别最佳提示的自动搜索技术，例如 AutoPrompt 中使用的那些，提供了一种改善合成数据生成的潜在解决方案。

在提示工程中，一个方面是选择现有的数据对象作为输入提示（例子）。这些示例数据对象的效果在合成数据生成中成为关键因素。已经提出了几种方法来实现有效输入例子的选择，从均匀分布到借助可视化的人为选择 [23, 24] 。受工业应用的启发，本文旨在提供一种自动化技术，该技术能比人为选择方法 [24] 更快地进行例子选择，同时模拟人为选择过程中的一些行为。

除了数据合成和提示工程之外，集成方法也成为提高合成数据生成的另一种有前途的方法。Xu 等人 [25] 提出了 AdaDEM，它在不同粒度级别集成多个卷积网络，以优化用于文本分类的合成数据选择。类似地，Zhou 等人 [26] 介绍了 MetaAugment，这是一个基于强化学习的集成框架，通过动态选择每个类别的增强策略来确保自适应增强。Agbesi 等人 [27] 开发了 MuTCELM，集成多个子分类器以在集成框架中捕获不同的语言特征。

在这项工作中，我们在模型改进工作流程 AutoGeTS 中提取测试数据中的有用知识，并利用这些知识选择一组有效的方法来识别示例数据。

2 方法

鉴于之前的研究 [24] 已经确认，使用 LLMs 生成的合成数据可以提高自动工单系统中分类模型的性能，本研究着眼于图 2 上部所示的三个研究问题。首先，我们进行了大量的结构化实验，以理解不同算法变体在寻找合适消息示例上对模型改进性能的影响。理想情况下，我们可能会发现一种更优越的算法。我们将在本节中描述我们的实验方法，并在第 **??** 节中报告结果和分析。我们的分析显示出没有这样一个优越的算法。这导致了第二和第三个研究问题，这将在第 **??** 节中讨论。

模型优化目标。部署在特定组织环境中的工单系统通常面临不同的，有时是相互矛盾的要求。典型的业务需求和相关的性能指标包括：

R1. 每个类别的准确性应该尽可能高，并且超过某一阈值。可以使用基于类别的平衡准确率或 F1-分数作为目标函数来优化模型，并将每个阈值作为约束条件。

R2. 模型的整体分类准确性应该尽可能高，并且超过某一阈值，因为错误分类的信息会导致不良后果。可以使用全局性能指标来优化模型，比如整体平衡准确率和整体 F1-分数。

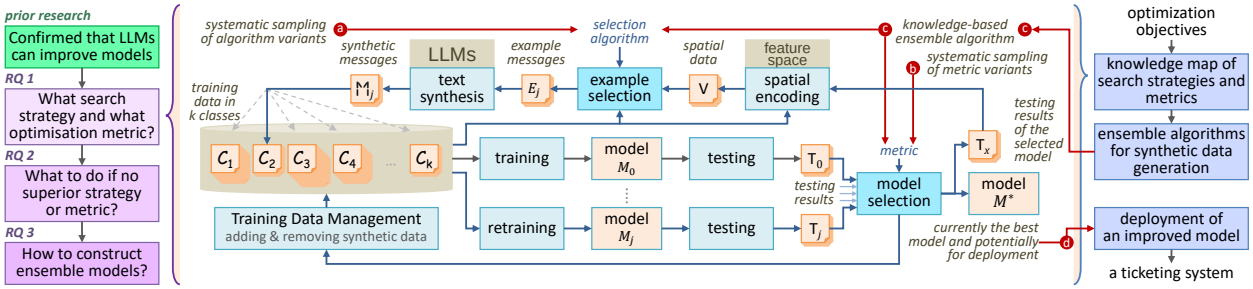
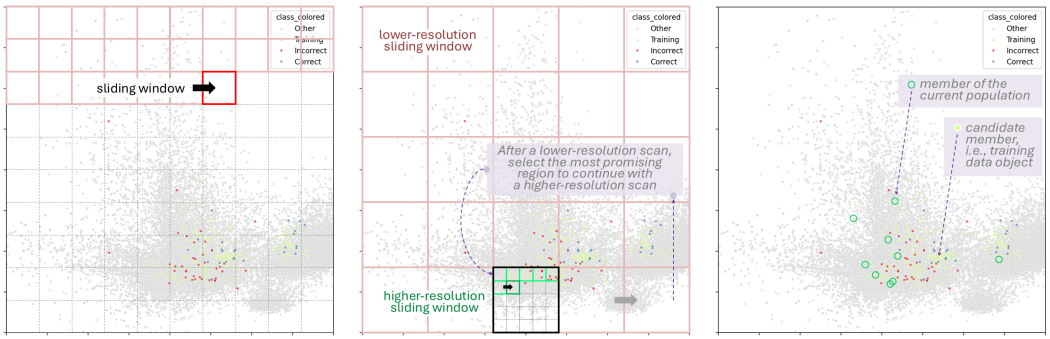


Figure 2: AutoGeTS 架构：（左）研究问题（RQs）引发了其设计。（中）其用于回答研究问题、创建知识图谱和进行模型改进操作的实验工作流程。（右）其用于改善模型的操作流水线。它可以支持多个票务系统的开发和维护。

R3. 某些特定类别的召回率（例如，重要类别）应该尽可能高，并且超过某一阈值，以便尽量减少因该类信息被发送到其他服务而导致的延迟。基于类别的召回率是此要求的明显指标。

这些要求决定了模型优化的目标函数和约束的定义。然而，由于使用 LLM 生成合成数据以辅助 ML 是一种新近的方法 [24]，了解不同的 LLM 示例选择算法如何影响不同指标下的优化性能是必要的。这引出了图 2 中的研究问题 1。



(a) sliding window (b) hierarchical sliding window (c) genetic algorithm

Figure 3: 本文探讨了用于选择示例消息的三种策略，这些示例消息将被大型语言模型用于生成合成数据。这些示例仅从特定类别的训练数据中选择（浅绿色点），而蓝色、红色和灰色点则展示了上下文信息，包括正确和错误的测试结果以及其他类别的数据。

AutoGeTS 架构和工作流程。AutoGeTS 架构包括一个实验环境和一个操作流水线，如图 2 的中部和右部所示。实验环境最初用于回答研究问题 1，后来用于支持操作流水线，操作流水线将在第 **??** 节详细介绍。

给定一个需要改进的模型 M_0 和一组改进需求（整体或特定类别），改进过程是一个优化工作流程，如图 2 中间部分所示。为了回答研究问题 1，我们系统地采样不同的选择算法变体，针对每个算法样本运行工作流程，收集大量结果，并分析这些变体在获得最佳模型时的性能。

正如在 [24] 中报道的那样，使用类似于图 3 中所示的基于类别的可视化图，人们可以从训练数据中选择示例消息，并提供给 LLM 以生成合成数据。为了大规模进行实验，我们在这个过程中替换了人工。对于选定的类别 C，每个类别都自动生成可视化图。所有训练和测试数据都被映射到 n 维特征空间中的数据点（在我们的工作中为 $n = 20$）。每个图描述了两个特征维度中的数据点，其中蓝色表示 C 中正确的测试数据点，红色表示 C 中的错误数据点，浅绿色表示 C 中的训练数据点，灰色表示其他类别的所有数据点。请注意，算法仅从绿色数据点中选择示例消息。

选择示例消息的算法可以有多种变体，其有效性取决于（1）训练数据集，（2）模型的结构和训练参数，（3）优化指标，以及（4）使用的 LLM 及其控制参数。虽然难以对所有算法变体及其相关条件进行细致采样，但我们的实验覆盖了广泛的范围：

0. 我们比较了三种示例选择的搜索策略，即滑动窗口（SW）、层次滑动窗口（HSW）和通用算法（GA）。我们进行了试验研究，以确保每种策略的参数（例如，窗口大小、交叉方法等）得到充分优化。

1. 除了在工业票务系统中收集的真实数据集外，我们还实验了另外两个数据集，TREC-6 [28] 和 Amazon Reviews 2023 [29] 。
2. 我们为我们的基线模型 M_0 优化了结构和训练参数，并在整个实验中对训练数据集始终使用相同的结构和训练参数。
3. 我们使用多种指标比较优化性能，包括每个 15 类的平衡准确率、F1-score 和召回率以及整个数据集的表现。
4. 我们使用了 GPT-3.5 的 API，采用了零样本提示模板。我们进行了试验研究，以确保其控制参数对使用所有三种算法（即 SW、HSW 和 GA）选择的示例进行充分优化。我们比较了使用 GPT-3.5 的 API 生成的数据和 Easy Data Augmentation (EDA) 工具 [30] 生成的数据。

示例选择策略。给定训练数据集中 m 条短信，共有 2^m 种不同的组合可以选择 $1 \leq k \leq m$ 条消息作为输入示例供 LLM 生成合成数据。从概率上看，使用这些 2^m 组合中的任何一种生成的合成数据都可能有助于改善模型。然而，测试所有 2^n 组合属于 NP 类别。

之前的非自动化工作 [24] 发现利用在不同二维特征空间中的负测试结果的视觉聚类（红点）来指导单个类别中从训练数据（浅绿色点）中选择示例。他们发现这种方法相对有效，尽管有些尝试未能导致模型改进。基于这一发现，我们考虑三种示例选择策略：

▶ 滑动窗口 (SW) —如图 3 (a) 所示，这种简单的自动化策略通过逐个扫描二维特征空间的方块进行操作，并在每次尝试中随机选择 k 个训练数据点作为样本。尝试的总次数多于窗口的数量。

▶ 分层滑动窗口 (HSW) —在非自动化工作 [24] 中，机器学习开发者根据他们观察到的视觉模式直观地判断选择示例的区域大小。为了提供比 SW 方法更大的灵活性，这种自动化策略在第一级使用更大的滑动窗口，当遇到一个有潜力或有趣的窗口时，它会如图 3 (b) 所示分层地采用更小的窗口进行更多尝试。k 的规范与 SW 相同。尝试的总次数是每个级别检查的每个窗口中的尝试次数之和。

▶ 遗传算法 (GA) —在非自动化工作 [24] 中，机器学习开发者不仅仅依赖于错误测试数据点的简单视觉聚类，有时他们会从一些分隔的区域选择示例。为了提供比 SW 和 HSW 方法更大的灵活性，这种自动化策略允许从二维特征空间内的任何训练数据中选择示例。给定特定类别 C_i 中的 m_i 个训练数据点，GA 维持一个包含 r 个染色体的种群，每个染色体中有 $1 \leq s \leq m_i$ 个开启的基因，代表选择为示例的 s 个训练数据点。在每次迭代中，GA 进化种群，同时尝试通过尚未测试的染色体改进模型。尝试的总次数是每次迭代中的尝试次数之和。

优化的目标函数。如图 2 所示，上述三种策略为示例选择过程提供了替代算法，其中图中间的工作流程进行了多次尝试以寻找特定类别的不同训练数据子集，作为用于 LLM 生成合成数据的示例。然后将合成数据添加到训练数据中，重新训练和测试分类。因此，该工作流程本质上是一个优化过程，旨在找到一组示例 E^*，使得模型 M^* 被认为是最佳模型，即，给定 $\mathbb{E}$，其中包括多次迭代工作流程中尝试的所有示例集，最佳模型 M^* 是：

$$\begin{aligned} \text{Objective:} \quad & \arg\max\big(f(M_0), \max_{E_j \in \mathbb{E}} f(M_j)\big) \quad j = 1, 2, \ldots, t \\ \text{Subject to:} \quad & M_j \leftarrow \text{TRAIN}\big(M^{\square}, D_{\text{tn}}, \text{LLM}(E_j)\big) \quad j = 1, 2, \ldots, t \\ & f(M_j) \leftarrow \phi\big(\text{TEST}(M_j, D_{\text{tt}})\big) \quad j = 1, 2, \ldots, t \end{aligned}$$

其中 D_{tn} 和 D_{tt} 分别是原始训练和测试数据，M_0 是待改进的基准模型，$t = \|\mathbb{E}\|$ 是尝试的总次数（即，从 D_{tn} 中选择的示例数据点集），$M_j (j \in [1, t])$ 是优化过程中采样的模型，TRAIN() 和 TEST() 是用于训练和测试分类模型 M_j 的过程，f 是使用统计测量 ϕ 对测试结果的评估，$M^{\square}$ 是分类模型的结构，而 LLM() 是 LLM 使用一组示例 E_j 生成合成数据的过程。

在票务系统中，回忆指标对于单个类别 ϕ^{cr} 是客户组织关注的一个度量指标，因为改进 ϕ^{cr} 可以减少在图 1 中的人工分类量。另一方面，对于特定的类别，当我们同时考虑 TP 和 FP（真阳性和假阳性）消息时，对于一个小类别，FP 总数（# FP）可以很容易地超过 TP 总数（# TP）。像准确性和精确度这样的度量可能会过度受到 # FP 的偏向。出于这个原因，我们专注于回忆 ϕ^{cr} 和平衡准确性 ϕ^{cba} 来衡量在特定类别 C_i 的情境下所取得的改进。同时，为了衡量模型的整体性能，我们使用整体的平衡准确性 ϕ^{oba} 和整体 F1-score ϕ^{of1} 。综上所述：

$$\phi_c^{\text{cr}} = \frac{\#\,\text{TP}_c}{\#\,\text{TP}_c + \#\,\text{FN}_c} \qquad \phi_c^{\text{cba}} = \frac{1}{2}\Big(\frac{\#\,\text{TP}_c}{\#\,\text{TP}_c + \#\,\text{FP}_c} + \frac{\#\,\text{TN}_c}{\#\,\text{TN}_c + \#\,\text{FN}_c}\Big)$$

$$\phi^{\text{of1}} = \frac{2\,\#\,\text{TP}}{2\,\#\,\text{TP} + \#\,\text{FP} + \#\,\text{FN}} \qquad \phi^{\text{oba}} = \frac{1}{2}\Big(\frac{\#\,\text{TP}}{\#\,\text{TP} + \#\,\text{FP}} + \frac{\#\,\text{TN}}{\#\,\text{TN} + \#\,\text{FN}}\Big)$$

，其中下标 $_c$ 指示类别 C 中的总值。

在这项工作中，我们首先进行了单阶段实验以回答研究问题 1，如第 2 节所讨论。在这里，“单阶段”意味着通过使用 $\mathbb{E}_a$ 运行工作流，即从单个类别 C_a 的训练数据中选择示例，来改进基线模型 M_0 。如图 2 中的（a）和（b）所示，我们系统地采样了三种算法变体和四种指标变体，以及选择示例信息的类别的变动。在第 **??** 节中，我们将报告系统测试可以为研究问题 2 和 3 提供答案，同时支持多阶段优化，通过在多个阶段中运行工作流，使用不同类别的示例，例如 $\mathbb{E}_a$ 、$\mathbb{E}_b$ 等，从而改进 M_0 。

对于研究问题 1，我们从几个方面寻找答案：

- 这三种策略（SW，HSW，GA）在不同的测试条件下（例如，性能指标，找到 M^* 所允许的时间等）表现如何。
- 对于每个指标 ϕ 作为目标函数，是否存在一种更优的策略?
- 使用一个衡量标准 ϕ 进行优化如何影响用其他衡量标准评估的模型性能?
- 使用一个类别的例子进行改进如何影响模型的整体性能及其在其他类别的性能?
- 大语言模型生成的合成数据是否可以与传统的数据增强方法相媲美？为此，我们还使用 Easy Data Augmentation (EDA) 工具 [30] 生成的数据进行了实验。
- 我们在实验中获得的发现是否依赖于数据？为此，我们还在两个公共数据集上进行了实验，TREC-6 [28] 和 Amazon Reviews 2023 [29] 。

实验设置。我们的主要实验使用在一个工业票务系统中收集的真实数据集进行。数据集包含 39,100 条作为训练和测试的数据对象的消息。它们的标签分为 15 个类别，分别对应不同的服务（图 1 ）。如表 1 所示，数据集高度不平衡，一些类别的数量少于总数据对象的% 的 1。由于这种不平衡，我们将数据集划分为 60 % 用于训练，20 % 用于优化测试，20 % 则用于优化过程之外的性能测试。

在不同组织的几乎所有工单系统中，使用不平衡数据进行模型开发是一种常见现象。

Table 1: 原始 CatBoost 模型 M_0 的性能

Class	Class Size	Balanced Accuracy	Recall	F1-Score
T1	8529	0.986	0.979	0.977
T2	11350	0.950	0.941	0.921
T3	4719	0.952	0.914	0.922
T4	1387	0.899	0.801	0.859
T5	2755	0.889	0.794	0.794
T6	1888	0.821	0.665	0.623
T7	1963	0.883	0.780	0.766
T8	1028	0.828	0.665	0.672
T9	1466	0.861	0.747	0.680
T10	1699	0.761	0.540	0.554
T11	471	0.973	0.947	0.967
T12	358	0.742	0.484	0.608
T13	180	0.666	0.333	0.469
T14	764	0.772	0.548	0.607
T15	543	0.726	0.452	0.596
Overall	39100	0.923	0.856	0.856

参照图 2 中间的工作流程，我们使用 GPT-3.5（版本：2023-03-15-preview）生成合成文本，主要参数为温度 = 0.7，最大 tokens 数 = 550，top p = 0.5，频率惩罚 = 0.3，出现惩罚 = 0.0。原始模型（M_0 ）是在一个工业应用中使用 CatBoost 开发的。为了确保这项研究与工业应用相关，我们使用 CatBoost 一致地训练所有模型，并固定超参数，即迭代次数 = 300，学习率 = 0.2，深度 = 8，L2 叶子正则化 = 1。

策略和指标的比较。我们系统地测试了三种策略（SW, HSW, GA）与四个目标指标（$\phi^{\text{cr}}, \phi^{\text{cba}}, \phi^{\text{oba}}, \phi^{\text{of1}}$ ）的组合。当这些组合应用于选择示例消息的 15 个类中的每一个类时，总共形成了 180 种组合。表格 2 (a) 显示了这 180 种组合的结果。所有结果均来自固定时间实验（1 GPU 小时）。

从表中，我们可以进行以下几个观察：

- 基于类别的客观指标（ϕ^{cr} 和 ϕ^{cba} ）大多对目标类别带来了改进，只有少数例外。例如，当对类别 T1 应用 ϕ^{cr} 时，策略 SW、HSW 和 GA 分别将 T1 的召回率提高了 0.23、0.35 和 0.35。
- 总体目标指标（ϕ^{oba} 和 ϕ^{of1} ）在选择示例消息的每个类别中都得到了持续改善。

Table 2: (a) 系统测试的结果包括在票务数据集中针对 3 种策略、4 种目标指标和 15 个类别的 180 种组合。单元格中的条形图描绘了在范围 $(0\%, 50\%]$ 内的改进，而红色文本表示性能恶化。(b) 结果可以被总结为一张知识图，以显示每个指标-类别组合的最佳策略或策略组合。对于地图的每个区域，与最佳策略相差不超过 0.03 % 的策略也被选中。

(a) the results of systematic testing combinations

	ΔCR %			ΔCBA %			ΔOBA %			ΔOF1 %		
Class	SW	HSW	GA	SW	HSW	GA	SW	HSW	GA	SW	HSW	GA
T1	0.23	0.35	0.35	0.05	0.05	-0.12	0.30	0.31	0.20	0.61	0.63	0.40
T2	0.71	0.85	0.80	0.53	0.50	-0.11	0.33	0.36	0.10	0.66	0.73	0.19
T3	2.01	2.01	2.01	0.61	0.65	0.73	0.33	0.30	0.31	0.66	0.61	0.63
T4	8.00	9.78	8.44	3.39	4.10	3.59	0.32	0.36	0.39	0.64	0.72	0.78
T5	5.67	4.26	2.84	2.12	1.57	0.66	0.34	0.36	0.11	0.69	0.73	0.22
T6	6.85	6.45	4.84	2.32	2.33	1.22	0.38	0.33	0.16	0.76	0.66	0.33
T7	5.84	5.84	-0.32	2.58	2.56	-0.30	0.39	0.36	0.13	0.78	0.73	0.25
T8	10.19	11.46	5.10	3.87	4.33	1.71	0.33	0.32	0.22	0.66	0.64	0.43
T9	4.39	5.37	2.93	1.70	2.22	0.89	0.39	0.30	0.28	0.79	0.60	0.57
T10	12.15	11.05	12.08	3.69	3.25	3.68	0.37	0.48	0.29	0.75	0.96	0.58
T11	1.12	1.12	1.12	0.55	0.55	0.55	0.33	0.33	0.42	0.66	0.66	0.85
T12	29.03	32.26	32.26	9.43	10.41	10.45	0.40	0.35	0.39	0.81	0.70	0.79
T13	26.67	33.33	33.33	6.65	8.22	8.23	0.33	0.40	0.37	0.66	0.81	0.75
T14	13.51	16.22	16.22	4.23	5.12	5.13	0.25	0.31	0.21	0.51	0.63	0.42
T15	20.75	20.75	24.53	6.28	6.14	7.34	0.36	0.36	0.40	0.72	0.73	0.81

(b) a summary of the best strategy-metric

	Measured Performance Metric			
Class	CR	CBA	OBA	OF1
T1	HSW/GA	SW/HSW	SW/HSW	SW/HSW
T2	HSW			HSW
T3	SW/HSW/GA	GA	SW/HSW/GA	SW/GA
T4	HSW	HSW	HSW/GA	GA
T5	SW	SW	SW/HSW	HSW
T6		SW/HSW	SW	SW
T7	SW/HSW		SW/HSW	
T8	HSW	HSW		SW/HSW
T9			SW	SW
T10	SW	SW/GA	HSW	HSW
T11	SW/HSW/GA	SW/HSW/GA	GA	GA
T12	HSW/GA	HSW/GA	SW/GA	SW/GA
T13			HSW/GA	HSW
T14			HSW	
T15	GA	GA	GA	GA

Table 3: 当一个模型在票务数据集中针对类 $T1, T2, \ldots, T15$ 采用 SW 策略和目标度量 ϕ^{cr} （基于类别的召回率）得到改进时，目标类别的直接改进可以在表格左侧的黄色单元格中看到。同时，这对其他类别和其他性能指标也有正面和负面的影响。单元格中的绿色条形表示改进的范围为 $(0\%, 50\%]$ ，而红色文本表示负面影响。最后一行显示的是基准性能。针对三种策略（SW、HSW、GA）和四种目标度量（$\phi_{\text{cr}}, \phi_{\text{cba}}, \phi_{\text{oba}}, \phi_{\text{of1}}$ ），共有 12 张这样的表格。

	Optimization: SW-CR-Class[row]; Performance Metric: ΔOF1%, ΔCR %																Optimization: SW-CR-Class[row]; Performance Metric: ΔOBA%, ΔCBA %														
	ΔOF1	ΔCR %															ΔOBA	ΔCBA %													
Class	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	0.5	0.2	-0.1	0.5	-0.4	0.5	0.8	1.9	3.8	-1.5	8.8	0.0	6.5	0.0	-1.4	-5.7	0.2	0.0	0.2	0.2	-0.2	0.2	0.3	1.0	1.6	-0.5	3.1	0.0	2.1	0.0	-0.5	-1.8
T2	0.6	-0.2	0.7	-0.1	1.8	0.5	0.8	3.2	0.0	2.0	5.5	0.0	16.1	-13.3	-1.4	-7.5	0.3	-0.1	0.5	-0.1	0.8	0.2	0.3	1.5	0.1	0.9	2.0	0.0	5.2	-3.3	-0.5	-2.4
T3	-0.1	-0.2	-0.8	2.0	1.3	-1.2	2.0	-2.9	2.5	0.5	0.0	0.0	3.2	0.0	1.4	-3.8	0.0	-0.1	-0.2	0.6	0.6	-0.5	0.8	-1.3	1.2	0.3	0.0	0.0	1.0	0.0	0.5	-1.2
T4	0.1	-0.4	-0.4	0.0	8.0	-0.9	1.2	2.9	1.3	-2.0	2.8	0.0	-3.2	6.7	-1.4	-9.4	0.1	-0.2	-0.1	0.0	3.4	-0.4	0.5	1.3	0.6	-0.8	0.9	0.0	-1.0	1.7	-0.5	-3.0
T5	0.6	-0.2	0.3	0.2	0.0	5.7	5.2	0.0	3.8	0.0	0.6	0.0	-3.2	0.0	-2.7	-7.5	0.3	-0.1	0.4	0.1	0.0	2.1	2.3	0.1	1.6	0.1	0.3	0.0	-1.1	0.0	-1.0	-2.4
T6	0.2	-0.1	-0.5	0.1	1.8	1.4	6.9	1.6	-5.1	0.5	4.4	0.0	-3.2	-6.7	-2.7	-9.4	0.1	-0.1	0.0	0.0	0.8	0.7	2.3	0.8	-1.9	0.3	1.7	0.0	-1.0	-1.7	-0.9	-3.0
T7	0.8	-0.1	0.6	0.0	1.3	0.9	0.0	5.8	3.8	0.5	3.3	0.0	16.1	0.0	-1.4	-1.9	0.4	0.0	0.5	0.0	0.6	0.5	0.1	2.5	1.6	0.3	1.2	0.0	5.3	0.0	-0.5	-0.6
T8	-0.1	-0.1	-0.7	0.1	0.4	-1.9	-1.2	1.0	10.2	0.5	2.2	0.0	0.0	6.7	-4.1	-5.7	-0.1	0.0	-0.2	0.1	0.2	-0.8	-0.4	0.4	3.9	0.1	0.8	0.0	0.0	1.7	-1.4	-1.8
T9	0.2	-0.5	0.4	0.4	1.8	0.5	-0.4	1.6	-1.3	4.4	2.2	0.0	-3.2	-6.7	-4.1	-5.7	0.1	-0.2	0.3	0.1	0.8	0.2	-0.1	0.8	-0.5	1.7	0.9	0.0	-1.0	-1.7	-1.5	-1.8
T10	-0.1	-0.4	-0.5	0.5	-0.4	-0.9	-0.8	1.9	-3.2	-2.9	12.2	0.0	-6.5	6.7	1.4	-5.7	0.0	-0.2	0.0	0.2	-0.2	-0.3	-0.2	0.8	-1.2	-1.1	3.7	0.0	-2.1	1.6	0.5	-1.8
T11	-0.1	-0.2	0.0	0.5	0.4	-0.9	-3.2	1.3	0.6	-1.0	2.2	1.1	-3.2	-6.7	1.4	-3.8	0.0	-0.1	0.1	0.2	0.2	-0.4	-1.3	0.5	0.3	-0.4	0.8	0.6	-1.1	-1.7	0.5	-1.2
T12	0.4	-0.2	0.5	0.2	0.4	0.7	-1.2	1.6	4.5	-2.4	4.4	0.0	29.0	0.0	-4.1	-7.5	0.2	-0.1	0.4	0.1	0.2	0.3	-0.4	0.7	1.8	-1.0	1.6	0.0	9.4	0.0	-1.5	-2.4
T13	0.2	-0.5	0.0	0.0	1.8	0.2	0.8	3.9	0.0	-2.0	4.4	0.0	-6.5	26.7	-1.4	-9.4	0.1	-0.2	0.1	0.0	0.8	0.1	0.4	1.7	0.0	-0.9	1.6	0.0	-2.1	6.6	-0.5	-3.0
T14	-0.1	-0.2	-0.1	0.0	-0.4	1.2	-1.6	0.0	-1.3	-1.0	-1.1	0.0	3.2	0.0	13.5	-5.7	0.0	-0.1	0.2	0.0	-0.2	0.5	-0.5	0.0	-0.4	-0.4	-0.4	0.0	1.1	0.0	4.2	-1.8
T15	0.1	0.0	0.0	-0.4	1.8	0.2	-1.2	-1.0	1.9	-2.0	3.9	0.0	-3.2	0.0	-2.7	20.8	0.1	0.0	0.1	-0.2	0.8	0.1	-0.4	-0.3	0.8	-0.8	1.3	0.0	-1.0	0.0	-1.0	6.3
M0	*0.86*	*0.98*	*0.94*	*0.91*	*0.80*	*0.79*	*0.66*	*0.78*	*0.67*	*0.74*	*0.54*	*0.95*	*0.48*	*0.33*	*0.55*	*0.45*	*0.92*	*0.99*	*0.95*	*0.95*	*0.90*	*0.89*	*0.82*	*0.88*	*0.83*	*0.86*	*0.76*	*0.97*	*0.74*	*0.67*	*0.77*	*0.73*

- 较小的班级，例如，T12 ∼ T15 有明显改善。
- 表 2 (b) 汇总了被认为是每个类别-指标组合中最佳的策略。尽管每个策略可能出现在形成色块的几个相邻单元格中，但没有一个无可争议的优越策略。这个观察导致了图 2 中的研究问题 2。

超越客观指标的性能。当在实验中将策略-指标组合应用于某一类时，我们还测量了模型的整体性能及其在其他类上的表现。表 3 显示了当将组合 (SW, ϕ^{cr}) 应用于 15 个类中的每一个时的这些测量结果。

例如，在第一行 T1 中，使用从 T1 训练数据中生成的合成数据重新训练的模型不仅提高了 T1 的召回率（即黄色单元格中提高了 0.2 %），还分别提高了两个整体度量 ϕ^{of1} 和 ϕ^{oba} 0.5 % 和 0.2 %，以及许多其他类别的基于类别的度量 ϕ^{cr} 和 ϕ^{cba} 。显著的是，类别 T10 的 ϕ^{cr} 提高了 8.8 %。

从表中我们还可以观察到，一些类别如 T10、T12 和 T13 经常从旨在改善其他类别的工作流程中受益。同时，有许多带有红色数字的单元格表明性能变差。请注意，这些结果仅代表单阶段实验。

通过三种策略和四个目标指标的 12 种组合，还有 11 个类似于表 3 的表格。这些实验的结果可以在附录 A 中找到。因此，如果想要提升类 T1 的 ϕ^{cr} ，可以在所有 12 个表中搜索列 Δ CR % -T1，寻找最佳的性能，从而找到最佳组合策略-指标-类。这 180 个单元格中的任何一个都可能为提升类 T1 的 ϕ^{cr} 提供最佳设置。

这 12 张实验结果表还告诉我们，不应做出简单的假设，例如 “为了改进 ϕ^{cba} 类的 C ，使用度量 ϕ^{cba} 和来自 C 类训练数据的示例消息进行模型改进总是最佳方法。” 这导致了研究问题 3 的解决方案（见 **??** 节）。

进一步实验。为了确保前面所述实验中观察到的现象不是孤立的个例，我们在相同的实验环境中使用不同的数据集（TREC-6 [28] 和 Amazon Reviews 2023 [29] ）以及一种替代的方法生成合成数据（Easy Data Augmentation（EDA）工具 [30] ）进行了进一步的实验。这些实验的结果可以在附录 **??** 中找到。

通常，我们可以观察到一个常见现象，即不存在优越的策略，我们也不能假设有一个优越的目标指标。同时，通过使用 EDA 生成的合成数据进行再训练所带来的改进量通常显著低于使用 GPT-3.5 生成的合成数据。

我们的大规模实验结果表明，没有一种优越的搜索策略或目标度量可以解决在第 2 节开头所述的典型业务需求。这引导我们研究问题 2。我们注意到，当我们重复这些单阶段实验时，结果非常一致。虽然每种 [SW, HSW, GA] $\times[\phi^{\text{cr}}, \phi^{\text{cba}}, \phi^{\text{oba}}, \phi^{\text{of1}}]$ 组合的性能依赖于许多因素（例如，训练数据、类别大小、特征规格、LLM 和模型结构 $M^{\square}$ ），但在为特定工单系统改进模型的过程中，这些因素变化不大。因此，单阶段实验（如第 **??** 节中报告的那些）收集了有关哪些搜索策略和哪些度量可能适用于每个类别的知识。换句话说，表格（如表 3 中的表）本质上是由机器学习训练和测试过程生成的定量知识图。一旦这些表格可用，模型改进过程可以使用它们来识别更有效的策略、度量和类别，以提高性能度量，并使过程更加成本效益。

多阶段模型改进。给定一个需要改进的模型 M_0 ，通过对 $[C_1, C_2, \ldots, C_k]\times$ [SW, HSW, GA] $\times[\phi^{\text{cr}}, \phi^{\text{cba}}, \phi^{\text{oba}}, \phi^{\text{of1}}]$ 的系统采样产生一个知识地图。在这一过程结束时，很可能也会产生最改进的模型 M_1^* 。因为在这个系统采样阶段，每次尝试只涉及某个单一类的训练数据示例信息。自然地，我们可以启动一个新的流程来改进在第一阶段获得的 M_1^* 。正如表格 3 中所示，我们也可以使用从另一个类中选择的示例来实现改进的需求。有了 12 个这样的表格，我们也可以使用不同的策略和不同的目标度量。给定一个单一的业务需求，我们可以调用多个阶段的模型改进，使用策略、目标度量和类的不同组合来提供示例信息。

多目标模型改进。回顾第 2 节中提到的三个典型业务需求，模型改进过程常常需要应对多个需求，例如，提升类别 X 和 Y 的召回率，同时提高整体准确率。第 **??** 节报告的实验也表明，如果仅仅采用一种方法尝试，这些需求可能会相互冲突（例如，表格 3 中的红色数字），而许多其他策略、目标度量和类别的组合可以提供替代的做法。一旦类似于表格 3 的表格可用，模型改进过程就可以利用这些表格中存储的知识来探索不同的成本效益方法以满足多目标需求。

因此，在早期阶段获得的性能测量可以促进基于知识的算法的发展，这为研究问题 3 提供了答案。

知识集成算法。图 4 显示了用于多阶段和多目标模型改进的集成算法流程图。给定一组目标 $O_1, O_2, \ldots, O_l$ 和一组约束 $S_1, S_2, \ldots, S_l$ ，算法从创建一组目标 $\boldsymbol{P_o}$ 开始。算法每次从 $\boldsymbol{P_o}$ 中随机选择一个目标 O_i ，旨在改进当前最佳候选模型，并以 O_i 为重点。随机性还为算法的外循环提供了进一步改进当前最佳候选模型的机会。

如果算法检测到缺少知识图，则会激活创建此类图的过程，同时优化 O_i （即黄色块）。如 2 和 **??** 部分所述，该过程是图 2 的中间部分，其系统地采样所有类别、搜索策略和目标指标的组合。例如，对于票务示例，该过程可能会为 15 个类别 × 3 个策略 × 4 个指标运行 180 个工作流。

如果知识图谱已经存在，算法将直接转向基于知识的方法（即三个蓝色模块）。算法选择可能最能使 O_i 受益的前 $k > 0$ 个实验设置，这些设置是根据知识图谱选出的。请注意，选出的设置不必使用为 O_i 定义的目标指标 ϕ 。例如，如果 O_i 旨在提高 ϕ^{cr}_{C11} （类 C_{11} 的召回率），只要知识图谱显示 ϕ^{cr}_{C13} 、ϕ^{cba}_{C1} 和 ϕ^{of1}_{C7} 也能够提高 ϕ^{cr}_{C11} 的性能指标，它们就可以作为待选的目标指标之一。

该算法还根据知识图来分配计算资源。考虑到知识图显示之前的前 k 设置改善了 O_i 达到 $x_1\%, x_2\%, \ldots, x_k\%$ ，算法将计算资源按比例分配给第 i 个设置，作为总可用资源的 $(x_i/\sum x_j)\%$ 。

该算法在每次循环中将最佳模型添加到模型池 $\boldsymbol{P_m}$ 中。在过程结束时，可以根据预定义的数值标准（例如，不同目标测量的加权和）从 $\boldsymbol{P_m}$ 中选择最佳模型，或由引入额外操作知识的人工专家选出最佳模型。或者，可以选择几个最佳模型来创建一个集成分类模型。

进一步的实验。我们进行了进一步的实验来研究单目标和多目标多阶段模型改进的场景。图 5 展示了一组具有单目标的实验过程和结果。业务需求将类 T13 视为重要类，并希望提高其准确性。它首先被翻译为一个优化规范，其中调用了五种不同的过程，(a) 基于知识的方法，(b, c, d) 三种直接以三种不同策略针对 T13 的蛮力方法，以及 (e) 一种随机选择策略-度量-类组合的方法。每种方法在五个阶段中进行尝试。对于单一目标，可以根据最高性能指标 $\Delta\text{CBA}_{\text{T13}}$ 选择最佳模型，即由基于知识的方法得出的 11.815 %。我们用不同类型的单一目标进行了类似的实验，结果显示在大多数实验中基于知识的方法表现更好。

图 6 展示了一组实验的过程和结果，这些实验有两个目标。业务需求是提高 T7 和 T12 类的召回率。这些需求首先被转化为最大化 $\phi^{\text{cr}}_{\text{T7}}$ 和 $\phi^{\text{cr}}_{\text{T12}}$ ，以及一个多目标评估函数。对于这一组实验，我们定义了一个简单的函数 $S = \phi^{\text{cr}}_{\text{T7}} + \phi^{\text{cr}}_{\text{T12}}$ ，它将触发将一个模型加入模型池 $\boldsymbol{P_m}$ 中的动作，并为后续阶段将相应的合成数据加入训练数据中。实际上，可以定义一个更复杂的评估函数和触发动作，例如，给予某些目标更高的优先级，或比训练数据更频繁地更新 $\boldsymbol{P_m}$ 。在右边的穷举方法 (b) 中，最初的四个阶段逐步做出了改进，然后进一步的改进变得困难。基于知识的方法 (a) 为每个目标调用了三种不同的设置，并以更分散的方式进行了改进，在第

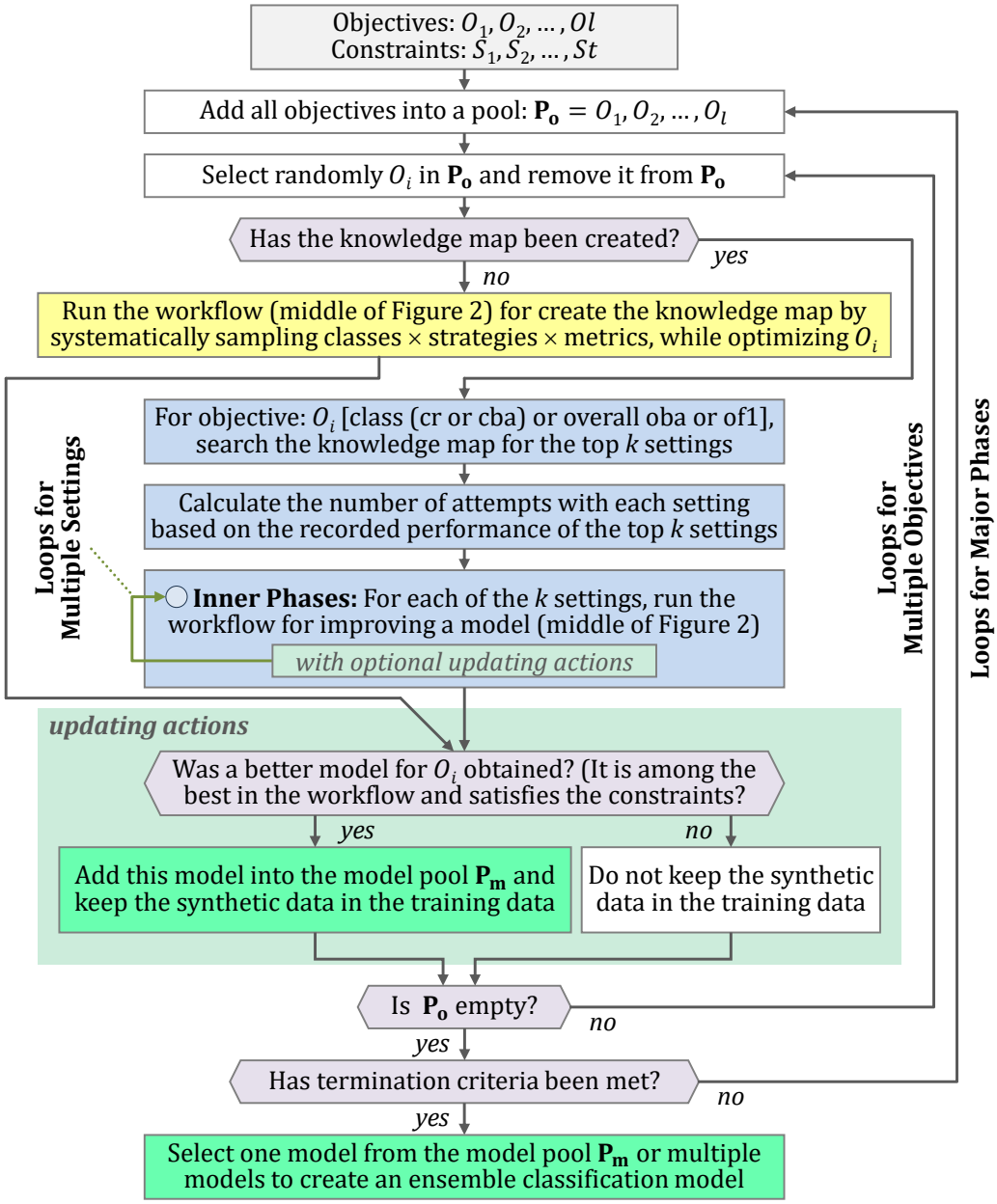


Figure 4: 一种基于知识的集合算法用于多阶段和多目标模型改进。

11 个 th 内阶段达到了最高的 S 值。我们进行了不同类型的多目标设置的类似实验，而基于知识的方法在大多数实验中表现得更好。

3 结论

在这项工作中，我们开发了一个自动化工作流程 AutoGeTS，用于改进文本分类模型，以及一种利用工作流程中发现的知识来指导后续模型改进过程的方法。虽然这项工作是由工业应用中的挑战所激励的，但我们遵循了三个研究问题，进行了大规模实验，使用公共数据集（TREC-6 [28] 和 Amazon Reviews 2023 [29] ）验证了我们的发现，此外还使用了票务数据，并将大型语言模型（LLMs）方法与传统的数据增强方法进行了比较。

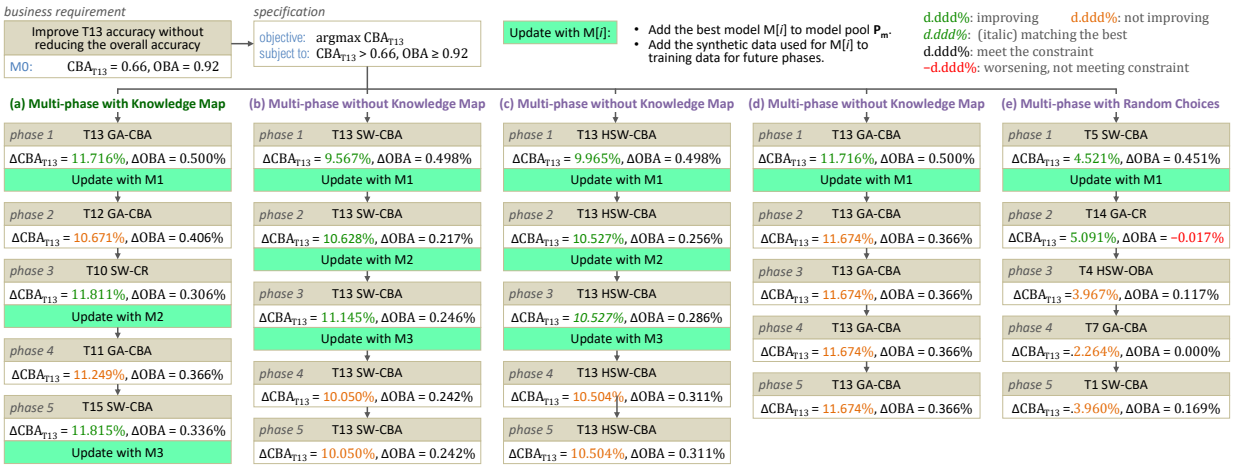


Figure 5: 使用和不使用知识图示的多阶段模型改进示例。

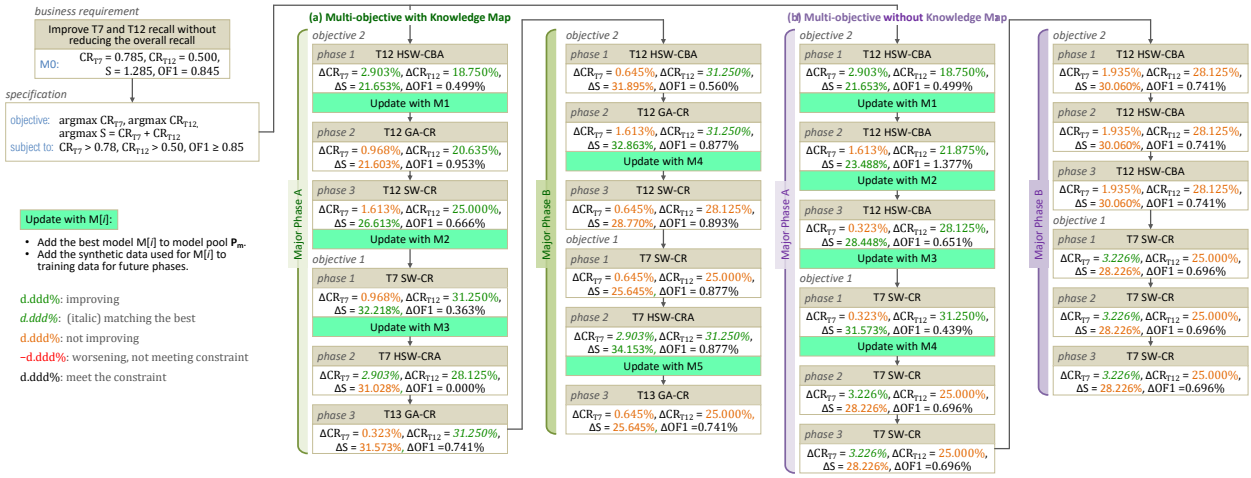


Figure 6: 使用和不使用知识图进行多目标模型改进的示例。

在许多机器学习工作流程中，通常没有单个已知技术是优于其他技术的，尽管大多数人都希望找到最佳解决方案。知识图是一种分而治之的图，显示在每个上下文中的最佳解决方案。因此，基于知识的集成是通过在每个上下文中使用最佳解决方案进行的分段优化。我们预计这种方法可以适用于许多其他实际应用。

这项工作还证实，使用大型语言模型生成合成数据可以解决在不同组织中部署的工单系统模型改进工作流中的一个常见挑战。由于大型语言模型不是这些分类模型的一部分，它们不会给模型增加额外的负担。随着大型语言模型通过版本更新得到改进，它们为 AutoGeTS 工作流提供了持续的技术支持。

在未来的工作中，我们计划进行更大规模的实验并分析结果，以深入理解不同因素如何使某些示例信息比其他信息更有效。我们希望利用这样的分析作为有用的知识，为开发更智能和更有效的技术来选择示例以生成合成数据提供指导。

4

致谢 这项工作得以实现要归功于欧洲数据科学家网络。我们要感谢那些为此项目提供便利的人们，特别是来自哥本哈根商学院的 Dolores Romero Morales。

References

[1] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41, 2022.

[2] Phani Krishna Kollapur Gandla, Rajesh Kumar Verma, Chhabi Rani Panigrahi, and Bibudhendu Pati. *Ticket Classification Using Machine Learning*, pages 487–501. Springer Nature Singapore, 2024.

[3] Alan K Meier, John Busch, and Craig C Conner. Testing the accuracy of a measurement-based building energy model with synthetic data. *Energy and buildings*, 12(1):77–82, 1988.

[4] Tom Bersano, Brad Clement, and Leonid Shilkrot. Synthetic data for testing in databases. *University of Michigan*, 1997.

[5] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

[6] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[7] Clifton D Sutton. Classification and regression trees, bagging, and boosting. *Handbook of statistics*, 24:303–329, 2005.

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[9] M Jaderberg, K Simonyan, A Vedaldi, and A Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. In *NIPS Deep Learning Workshop*. Neural Information Processing Systems, 2014.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[11] Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial feature matching for text generation. In *International conference on machine learning*, pages 4006–4015. PMLR, 2017.

[12] Andoni Cortés, Clemente Rodríguez, Gorka Vélez, Javier Barandiarán, and Marcos Nieto. Analysis of classifier training on synthetic data for cross-domain datasets. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):190–199, 2020.

[13] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[14] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416:244–255, 2020.

[15] Danilo Croce, Giuseppe Castellucci, and Roberto Basili. Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2114–2119, 2020.

[16] Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. Generate, annotate, and learn: Nlp with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842, 2022.

[17] D Garcia Torres. Generation of synthetic data with generative adversarial networks. *Unpublished doctoral dissertation). Ph. D. Thesis, Royal Institute of Technology, Stockholm, Sweden*, 26, 2018.

[18] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[20] Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*, 2021.

[21] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. Gpt3mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, 2021.

[22] Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric Xing, and Zhiting Hu. Promptagent: Strategic planning with language models enables expert-level prompt optimization. In *The Twelfth International Conference on Learning Representations*, 2023.

[23] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[24] Yuanzhe Jin, Adrian Carrasco-Revilla, and Min Chen. igaiva: Integrated generative ai and visual analytics in a machine learning workflow for text classification. *arXiv preprint arXiv:2409.15848*, 2024.

[25] Yuhong Xu, Zhiwen Yu, Wenming Cao, and C. L. Philip Chen. Adaptive dense ensemble model for text classification. *IEEE Transactions on Cybernetics*, 52(8):7513–7526, 2022.

[26] Fengwei Zhou, Jiawei Li, Chuanlong Xie, Fei Chen, Lanqing Hong, Rui Sun, and Zhenguo Li. Metaaugment: Sample-aware data augmentation policy learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11097–11105, 2021.

[27] Victor Kwaku Agbesi, Wenyu Chen, Sophyani Banaamwini Yussif, Chiagoziem C Ukwuoma, Yeong Hyeon Gu, and Mugahed A Al-Antari. Mutcelm: An optimal multi-textcnn-based ensemble learning for text classification. *Heliyon*, 10(19), 2024.

[28] Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[29] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

[30] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, 2019.

APPENDICES

AutoGeTS: Knowledge-based Automated Generation of Text Synthetics for Improving Text Classification

在以下附录中，我们提供了通过可视化图来展现的更多实验结果。在双盲评审过程结束后，实验数据将在 GitHub 上公布。这些附录包括：

- 附录 **??** 提供了与两个额外数据集及传统数据扩增工具的进一步实验结果。结果证实了在第 **??** 节中关于研究问题 1 的发现。
- 附录 A 中提供了进一步的实验结果，共包含 11 个表格。总共有 12 个表格，只有一个表格（表 3）在论文主体中给出。这 12 个表格中的结果帮助我们回答了研究问题 2 和研究问题 3，并为第 **??** 节中介绍的集成算法提供了知识图。

在本附录中，我们报告了进一步的实验，以验证关于票务数据实验中缺乏优越搜索策略的观察（表 2）。我们使用了两个公共数据集进行了这些进一步的实验：TREC-6 [28] 和亚马逊评论 2023 [29]，并结合使用了 GPT-3.5 的 API 和 Easy Data Augmentation（EDA）工具 [30]。

TREC-6 数据集 [28] 包含 5542 个基于事实的问题，这些问题被分类为六个语义类别，每类的问题数量从 86 到 1250 不等。在表 4 中展示了在 TREC-6 上训练的原始 CatBoost 分类模型（没有使用合成数据）的性能。

Table 4: 在 TREC-6 数据集上训练的原始 CatBoost 模型 M_0 的性能（没有使用合成数据）。

Class	Class Size	Balanced Accuracy	Recall	F1-Score
ENTY	1250	0.861	0.825	0.757
HUM	1223	0.903	0.850	0.846
DESC	1162	0.881	0.802	0.820
NUM	896	0.908	0.836	0.866
LOC	835	0.882	0.789	0.819
ABBR	86	0.761	0.522	0.686
Overall	5542	0.889	0.816	0.816

与票务数据类似，我们使用 EDA 工具和 GPT-3.5 来生成合成数据，并进行了包含三种策略、四个目标指标和六个类别的所有组合的实验，其中例子问题被选择作为提示。EDA 的结果如表 5 所示，GPT 3.5 的结果如表 6 所示。

Table 5: 对 TREC-6 数据集和由 EDA 工具生成的合成数据进行系统实验。(a) 对票务数据集中 3 种策略、4 个目标指标和 6 个类别的 72 种组合进行系统测试的结果。单元格中的条形图展示了在范围 $(0\%, 50\%]$ 内的提升，而红色文本则表示性能下降。(b) 结果可以总结为一张知识地图，显示每个指标-类别组合的最佳策略或策略组合。对于每个地图区域，与最佳策略差距在 0.03 % 之内的策略也被选中。

(a) the results of systematic testing combinations

	ΔCR %			ΔCBA %			ΔOBA %			ΔOF1 %		
Class	SW	HSW	GA	SW	HSW	GA	SW	HSW	GA	SW	HSW	GA
ENTY	3.62	3.62	1.95	0.97	0.96	0.52	0.51	0.19	0.19	0.93	0.34	0.34
HUM	4.31	8.33	4.31	1.50	2.37	1.64	0.63	0.49	0.37	1.14	0.90	0.67
DESC	9.34	14.84	8.79	2.15	2.97	2.29	0.62	0.49	0.06	1.12	0.90	0.11
NUM	8.86	9.49	8.23	4.02	4.99	3.60	0.88	0.68	0.25	1.59	1.24	0.45
LOC	20.10	18.62	8.40	5.12	5.56	3.02	0.63	0.43	0.06	1.14	0.79	0.11
ABBR	20.50	20.73	25.84	8.34	8.57	8.57	0.64	0.66	0.62	1.16	1.20	1.12

(b) a summary of the best strategy-metric

	Measured Performance Metric			
Class	CR	CBA	OBA	OF1
ENTY	SW/HSW	SW	SW	SW
HUM	HSW	HSW		
DESC				
NUM				
LOC	SW			
ABBR	GA	SW/GA	HSW	HSW

从这两个表中，我们可以观察到：

- 表格 6 中的改进幅度通常高于表格 5 中的改进幅度，这证实了使用 LLMs 生成合成数据的优势。
- 使用 EDA 工具生成的合成数据时，以分类度量作为目标度量，策略 HSW 在多个类别上表现更好。而当使用整体度量作为目标度量时，策略 SW 在大多数情况下表现更好。
- 利用 GPT 3.5 API 生成的合成数据，策略 SW 在绝大多数情况下都表现得更好。然而，有五个单元格中 SW 并不是首选策略。
- 这两个表格证实了基于知识的方法能够为单个类别和目标指标选择更有效的策略。

Table 6: 使用 TREC-6 数据集和由 GPT-3.5 API 生成的合成数据进行系统实验。(a) 在工单数据集中，测试了 3 种策略、4 个目标度量和 6 个类别的 72 种组合的系统测试结果。单元格中的条形图显示了在范围 $(0\%, 50\%]$ 内的改进，而红色文本表示性能下降。(b) 可以将结果总结为一个知识图，显示每个度量-类别组合的最佳策略。对于每个地图区域，与最佳策略相差不到 0.03 % 的策略也被选中。

(a) the results of systematic testing combinations

	ΔCR %			ΔCBA %			ΔOBA %			ΔOF1 %		
Class	SW	HSW	GA	SW	HSW	GA	SW	HSW	GA	SW	HSW	GA
ENTY	7.39	5.91	3.16	2.40	1.45	0.65	0.91	0.37	0.25	1.64	0.67	0.45
HUM	14.10	13.59	7.35	3.20	3.14	2.26	0.97	0.65	0.75	1.76	1.17	1.37
DESC	25.62	21.47	11.23	4.79	4.01	2.95	0.91	0.66	0.65	1.64	1.20	1.19
NUM	15.70	15.70	15.70	5.85	5.64	5.64	1.09	0.78	0.80	1.98	1.42	1.46
LOC	18.60	20.67	14.82	5.93	6.17	4.43	1.05	0.93	0.62	1.91	1.69	1.12
ABBR	32.38	39.40	30.08	10.31	10.31	10.31	0.91	0.99	0.99	1.64	1.80	1.79

(b) a summary of the best strategy-metric

	Measured Performance Metric			
Class	CR	CBA	OBA	OF1
ENTY	SW	SW	SW	SW
HUM				
DESC				
NUM	SW/HSW/GA			
LOC	HSW	HSW		
ABBR		SW/HSW/GA	HSW/GA	HSW/GA

.1 使用 EDA 和 GPT-3.5 的亚马逊子集

在 Amazon Reviews'2023 数据集 [29] 中，随机选择了礼品卡主题的 10000 条评论。这个子集有五个不同的评分类别，类别大小从 121 到 8389 条评论不等。在该 Amazon Reviews 子集（不含合成数据）上训练的原始 CatBoost 分类模型的性能如表 7 所示。

Table 7: 在亚马逊评论子集上训练的原始 CatBoost 模型 M_0 的性能（没有使用合成数据）。

Class	Class Size	Balanced Accuracy	Recall	F1-Score
R1	807	0.834	0.705	0.650
R2	121	0.499	0	0
R3	206	0.524	0.049	0.089
R4	477	0.586	0.174	0.288
R5	8389	0.758	0.977	0.950
Overall	10000	0.934	0.894	0.894

我们可以注意到，对于类别 R2，没有真正的正结果，因此召回率和 F1 得分都是 0。对于 R3 和 R4，召回率和 F1 得分的度量都非常低。由于这些类别相比于 R5 而言是较小的类别，因此整体度量看起来相当合理却具有误导性。

与票务数据和 TREC-6 数据类似，我们同时使用了 EDA 工具和 GPT-3.5 生成合成数据，并且进行了结合三种策略、四种目标指标和五个类别的所有组合的实验，其中示例评论被选择作为提示。使用 EDA 工具的实验的直接测试结果显示在表 **??** 中，使用 GPT 3.5 API 的直接测试结果显示在表 **??** 中。注意，在这两个表中，类别 R2 的召回值（CR）现在都高于零。显然，合成数据提供了帮助。

类似地，我们分别在表 8 和 9 中展示了改进的数量。从这两个表中，我们可以观察到：

- 表 9 中的改进幅度通常比表 8 中的要高，证实了使用 LLMs 生成合成数据的优势。
- 使用合成数据显著提高了 R2、R3 和 R4 类的召回率。
- 没有明显的优越策略。
- 这两张表格证实了基于知识的方法可以为各个类别和目标指标选择更有效的策略。

Table 8: 使用亚马逊评论数据集的子集和由 EDA 工具生成的合成数据进行系统实验。(a) 展示了 3 种策略、4 种目标指标和票务数据集中的 5 个类别的 60 种组合的系统测试结果。单元格中的条形显示了 $(0\%, 50\%]$ 范围内的改进，而红色文本表示性能下降。对于左侧的六个深绿色单元格，原始召回率为零或接近零。因此，改进幅度远超过 50 个%。Δ CR % 值以白色文字显示在深绿色背景上。###### 表示无穷。(b) 结果可以总结为一种知识地图，显示每个指标-类别组合的最佳策略或多种策略。对于每个地图区域，与最佳策略相差小于 0.03 % 的策略也会被选中。

(a) the results of systematic testing combinations

	ΔCR %			ΔCBA %			ΔOBA %			ΔOF1 %		
Class	SW	HSW	GA	SW	HSW	GA	SW	HSW	GA	SW	HSW	GA
R1	3.81	8.23	3.81	1.38	2.25	1.48	0.27	0.27	0.40	0.45	0.45	0.67
R2	######	######	######	5.37	5.37	5.37	0.44	0.40	0.40	0.73	0.67	0.67
R3	150.0	100.0	100.0	6.98	6.89	6.64	0.37	0.33	0.33	0.62	0.56	0.56
R4	26.67	33.33	33.33	2.54	2.80	2.76	0.23	0.40	0.27	0.39	0.67	0.45
R5	0.66	0.72	0.66	3.10	3.40	3.43	0.40	0.44	0.44	0.67	0.73	0.73

(b) a summary of the best strategy-metric combinations

	Measured Performance Metric			
Class	CR	CBA	OBA	OF1
R1	HSW	HSW	GA	GA
R2	SW/GA	SW/HSW/GA	SW	SW
R3	SW	SW		
R4	HSW/GA	HSW	HSW	HSW
R5	HSW	GA	HSW/GA	HSW/GA

Table 9: 对亚马逊评论数据集的一个子集以及由 GPT-3.5 API 生成的合成数据进行系统化实验。(a) 对票务数据集中 3 个策略、4 个目标指标和 5 个类的 60 种组合进行系统测试的结果。单元格中的条形显示了在 $(0\%, 50\%]$ 范围内的改善，而红色文本表示性能下降。对于左侧的九个深绿色单元格来说，原始召回值为 0 或接近零。因此，改进幅度远超 50 % 。Δ CR % 值显示为深绿色背景上的白色文本。###### 表示无穷大。(b) 结果可以总结为一个知识图，展示了每个指标-类别组合的最佳策略或策略。对于每个地图区域，距离最佳策略 0.03 % 之内的策略也被选中。

(a) the results of systematic testing combinations

	ΔCR %			ΔCBA %			ΔOBA %			ΔOF1 %		
Class	SW	HSW	GA	SW	HSW	GA	SW	HSW	GA	SW	HSW	GA
R1	5.71	13.04	5.71	2.58	3.85	2.38	0.47	0.50	0.47	0.78	0.84	0.78
R2	######	######	######	15.86	15.91	15.86	0.64	0.54	0.47	1.06	0.89	0.78
R3	200.0	284.4	200.0	9.31	8.82	8.92	0.44	0.50	0.47	0.73	0.84	0.78
R4	99.7	99.7	86.8	4.82	4.65	3.70	0.47	0.47	0.44	0.78	0.78	0.73
R5	1.20	2.04	1.81	9.21	8.44	8.85	0.50	0.60	0.70	0.84	1.01	1.17

(b) a summary of the best strategy-metric combinations

	Measured Performance Metric			
Class	CR	CBA	OBA	OF1
R1	HSW	HSW	HSW	HSW
R2	SW/HSW/GA		SW	SW
R3	HSW	SW	HSW	HSW
R4	SW/HSW		SW/HSW	SW/HSW
R5	HSW		GA	GA

Table 10: 类似于表格 3 ，当一个模型在使用 HSW 策略和目标指标 ϕ^{cr} （基于类别的召回率）来改进票务数据集中的类别 $T1, T2, \ldots, T15$ 时，可以在表格左侧部分的黄色单元格中看到目标类别的直接改进。同时，对其他类别和其他性能指标也有正面和负面的影响。单元格中的绿色条表示改进范围 $(0\%, 50\%]$ ，而红色文本表示负面影响。最后一行显示了基准性能。

	Optimization: HSW-CR-Class[row]; Performance Metric: ΔOF1%, ΔCR %																Optimization: HSW-CR-Class[row]; Performance Metric: ΔOBA%, ΔCBA %														
	ΔOF1	ΔCR %															ΔOBA	ΔCBA %													
Class	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	-0.4	0.4	-0.4	0.6	1.3	-0.9	-2.8	1.6	-5.1	0.0	1.1	0.0	9.7	-20.0	-6.8	-28.3	-0.2	-0.2	-0.1	0.2	0.5	-0.5	-1.1	0.7	-1.9	0.1	0.3	0.0	3.1	-5.0	-2.4	-8.9
T2	0.5	-0.2	0.8	0.4	0.4	0.0	1.6	1.6	2.5	-2.4	2.8	0.0	3.2	0.0	0.0	-1.9	0.2	-0.1	0.4	0.2	0.2	0.0	0.7	0.8	1.1	-1.0	0.9	0.0	1.1	0.0	0.1	-0.6
T3	-0.1	-0.2	-0.8	2.0	1.3	-1.2	2.0	-2.9	2.5	0.5	0.0	0.0	3.2	0.0	1.4	-3.8	0.0	-0.1	-0.2	0.6	0.6	-0.5	0.8	-1.3	1.2	0.3	0.0	0.0	1.0	0.0	0.5	-1.2
T4	0.1	-0.2	-0.5	0.0	9.8	-0.7	-2.8	1.6	1.9	-2.9	5.5	0.0	9.7	-13.3	-2.7	-7.5	0.0	-0.1	0.0	0.0	4.1	-0.3	-1.1	0.7	0.8	-1.2	1.9	0.0	3.2	-3.3	-0.9	-2.4
T5	0.3	-0.5	0.4	0.4	0.9	4.3	0.0	1.3	0.0	-2.9	0.6	0.0	6.5	-6.7	1.4	-3.8	0.2	-0.2	0.3	0.1	0.4	1.6	0.2	0.7	0.1	-1.2	0.1	0.0	2.1	-1.7	0.5	-1.2
T6	0.1	-0.4	0.1	-0.2	1.8	1.7	6.5	0.3	-1.9	-2.4	0.6	0.0	-3.2	0.0	-1.4	-9.4	0.1	-0.2	0.2	-0.1	0.8	0.8	2.4	0.2	-0.7	-1.0	0.2	0.0	-1.1	0.0	-0.5	-3.0
T7	0.7	-0.2	0.7	0.2	2.7	1.7	1.2	5.8	3.2	-3.4	3.9	0.0	6.5	-6.7	-1.4	-5.7	0.4	-0.1	0.5	0.1	1.2	0.8	0.6	2.6	1.3	-1.4	1.4	0.0	2.1	-1.7	-0.5	-1.8
T8	-0.2	-0.2	-0.6	-0.7	1.3	-0.9	-1.2	-0.6	11.5	0.5	1.7	0.0	-6.5	-6.7	-4.1	-7.5	-0.1	-0.1	-0.3	-0.4	0.6	-0.3	-0.5	-0.3	4.3	0.3	0.6	0.0	-2.1	-1.7	-1.4	-2.3
T9	0.4	-0.2	0.4	0.1	1.3	1.9	-2.0	1.9	-2.5	5.4	4.4	0.0	9.7	0.0	-6.8	-5.7	0.2	-0.1	0.2	0.1	0.6	0.9	-0.7	0.9	-1.0	2.2	1.7	0.0	3.2	0.0	-2.4	-1.8
T10	0.1	-0.1	-0.1	0.4	-1.3	0.0	0.8	0.0	-2.5	0.5	11.0	0.0	-6.5	0.0	-6.8	-7.5	0.0	0.0	0.2	0.2	-0.6	-0.1	0.5	0.1	-0.9	0.2	3.2	0.0	-2.1	0.0	-2.4	-2.3
T11	0.6	-0.3	0.1	0.9	2.7	1.9	0.4	2.6	2.5	-2.9	9.4	1.1	3.2	-6.7	2.7	-9.4	0.3	-0.2	0.4	0.4	1.2	0.9	0.3	1.2	1.1	-1.1	3.2	0.5	1.0	-1.7	0.9	-3.0
T12	0.3	-0.2	0.0	0.4	3.6	0.9	-2.4	1.3	-0.6	0.0	3.9	0.0	32.3	-6.7	-1.4	-7.5	0.1	-0.1	0.2	0.1	1.5	0.4	-0.8	0.6	-0.1	0.0	1.3	0.0	10.4	-1.7	-0.4	-2.4
T13	0.0	-0.3	-0.2	0.0	0.9	0.7	-3.2	0.3	3.2	0.5	2.2	0.0	-3.2	33.3	0.0	-1.9	0.0	-0.1	-0.1	0.0	0.4	0.3	-1.3	0.2	1.2	0.2	0.8	0.0	-1.0	8.2	0.0	-0.6
T14	-1.0	-0.2	-1.1	-0.5	2.2	-1.2	-3.6	0.0	-6.4	-4.4	-5.0	0.0	-12.9	-6.7	16.2	-11.3	-0.5	-0.1	-0.4	-0.3	1.0	-0.5	-1.5	0.0	-2.5	-1.9	-1.8	0.0	-4.2	-1.7	5.1	-3.5
T15	0.2	-0.3	-0.3	0.5	1.8	0.0	-4.4	2.6	1.9	-1.0	3.3	0.0	9.7	-13.3	1.4	20.8	0.1	-0.1	0.1	0.2	0.8	-0.1	-1.7	1.1	0.8	-0.3	1.1	0.0	3.1	-3.3	0.5	6.1
M0	*0.86*	*0.98*	*0.94*	*0.91*	*0.80*	*0.79*	*0.66*	*0.78*	*0.67*	*0.74*	*0.54*	*0.95*	*0.48*	*0.33*	*0.55*	*0.45*	*0.92*	*0.99*	*0.95*	*0.95*	*0.90*	*0.89*	*0.82*	*0.88*	*0.83*	*0.86*	*0.76*	*0.97*	*0.74*	*0.67*	*0.77*	*0.73*

Table 11: 与表 3 和表 10 类似，此表显示了 GA 策略和目标指标 ϕ^{cr} （基于类别的召回率）对票务数据集中的类 $T1, T2, \ldots, T15$ 的影响。

	Optimization: GA-CR-Class[row]; Performance Metric: ΔOF1%, ΔCR %																Optimization: GA-CR-Class[row]; Performance Metric: ΔOBA%, ΔCBA %														
	ΔOF1	ΔCR %															ΔOBA	ΔCBA %													
Class	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	0.2	0.4	-0.6	0.8	1.3	-0.7	-0.4	1.3	-2.5	0.5	10.5	0.0	3.2	0.0	-4.1	-13.2	0.1	-0.1	0.0	0.4	0.5	-0.3	-0.1	0.6	-1.0	0.3	3.7	0.0	1.0	0.0	-1.5	-4.1
T2	0.0	0.4	0.8	-0.5	-0.9	0.9	-2.8	-0.6	-5.7	1.0	5.0	0.0	3.2	-13.3	-2.7	-17.0	0.0	0.2	-0.1	-0.2	-0.4	0.4	-0.9	-0.2	-2.3	0.6	1.8	0.0	1.0	-3.3	-1.0	-5.3
T3	0.3	-0.2	0.0	2.0	1.3	0.5	0.8	1.0	-0.6	-1.5	3.9	0.0	-6.5	-6.7	-5.4	-5.7	0.1	-0.1	-0.1	0.7	0.6	0.2	0.3	0.5	-0.2	-0.5	1.6	0.0	-2.1	-1.7	-1.9	-1.7
T4	0.4	0.0	0.0	1.1	8.4	-0.7	0.8	1.0	-2.5	0.5	3.3	0.0	-3.2	0.0	-2.7	-3.8	0.2	0.1	0.2	0.5	3.6	-0.3	0.2	0.5	-0.9	0.2	1.3	0.0	-1.1	0.0	-0.9	-1.2
T5	0.2	-0.3	-0.3	0.9	0.4	2.8	1.2	1.0	-2.5	2.4	3.3	0.0	3.2	-20.0	-1.4	-9.4	0.1	-0.1	0.3	0.6	0.2	0.7	0.6	0.4	-0.9	1.1	1.2	0.0	1.0	-5.0	-0.5	-3.0
T6	-0.1	0.0	0.0	0.2	0.0	0.9	4.8	1.0	-2.5	-3.4	0.0	0.0	-3.2	-13.3	-4.1	-15.1	0.0	0.0	0.2	0.1	0.0	0.3	1.2	0.5	-0.8	-1.4	0.2	0.0	-1.0	-3.3	-1.4	-4.7
T7	0.0	0.0	0.0	0.4	0.0	0.5	2.4	-0.3	1.9	-2.0	1.7	0.0	-3.2	0.0	-6.8	-5.7	0.0	0.0	0.1	0.2	0.1	0.1	0.9	-0.3	0.8	-0.8	0.6	0.0	-1.1	0.0	-2.4	-1.8
T8	-0.4	-0.2	-0.5	0.1	-0.9	1.2	-3.2	0.6	5.1	-2.0	-4.4	0.0	3.2	-6.7	0.0	-9.4	-0.2	-0.1	-0.2	0.1	-0.4	0.5	-1.3	0.3	1.7	-0.8	-1.5	0.0	1.0	-1.7	0.0	-3.0
T9	-0.1	-0.2	0.1	0.2	-0.4	-0.2	0.0	1.3	-3.2	2.9	-2.2	0.0	6.5	0.0	-1.4	-9.4	0.0	-0.1	0.2	0.1	-0.2	-0.1	0.0	0.6	-1.2	0.9	-0.7	0.0	2.1	0.0	-0.4	-3.0
T10	0.1	-0.5	-0.2	0.2	-0.4	1.2	-0.4	1.0	-3.2	0.5	12.1	-1.1	3.2	-6.7	-1.4	-11.3	0.0	-0.2	0.2	0.1	-0.2	0.5	0.0	0.5	-1.2	0.3	3.7	-0.5	1.1	-1.6	-0.4	-3.5
T11	0.2	-0.1	0.1	0.5	0.0	-1.9	3.2	2.6	1.9	-2.4	3.3	1.1	-3.2	0.0	0.0	-11.3	0.1	0.0	0.1	0.2	0.0	-0.9	1.2	1.1	0.8	-0.9	1.1	0.5	-1.0	0.0	0.0	-3.6
T12	0.1	-0.4	-0.2	0.2	1.8	1.7	-0.4	1.0	0.0	-2.0	1.7	0.0	32.3	6.7	-1.4	-5.7	0.1	-0.2	0.1	0.1	0.8	0.7	-0.2	0.4	0.1	-0.9	0.6	0.0	10.4	1.7	-0.5	-1.8
T13	0.1	-0.6	0.1	0.2	1.3	-1.4	-2.0	5.2	1.9	-2.4	1.1	0.0	3.2	33.3	-1.4	-3.8	0.0	-0.3	0.3	0.1	0.5	-0.6	-0.8	2.2	0.7	-1.0	0.4	0.0	1.0	8.2	-0.4	-1.2
T14	0.0	-0.2	0.1	0.1	0.4	-0.9	-0.8	0.6	-2.5	0.0	-1.7	0.0	0.0	-13.3	16.2	-5.7	0.0	-0.1	0.2	0.1	0.2	-0.3	-0.2	0.3	-1.0	0.1	-0.5	0.0	0.0	-3.3	5.1	-1.8
T15	0.1	-0.4	0.0	0.4	1.3	0.2	-1.6	1.0	-3.8	-1.5	2.8	0.0	3.2	-6.7	0.0	24.5	0.1	-0.1	0.2	0.1	0.6	0.1	-0.5	0.4	-1.5	-0.6	0.9	0.0	1.1	-1.7	0.0	7.3
M0	*0.86*	*0.98*	*0.94*	*0.91*	*0.80*	*0.79*	*0.66*	*0.78*	*0.67*	*0.74*	*0.54*	*0.95*	*0.48*	*0.33*	*0.55*	*0.45*	*0.92*	*0.99*	*0.95*	*0.95*	*0.90*	*0.89*	*0.82*	*0.88*	*0.83*	*0.86*	*0.76*	*0.97*	*0.74*	*0.67*	*0.77*	*0.73*

A 对其他性能指标的影响

如第 **??** 节所述，使用票务数据，有 180 种组合的策略、四个目标指标和 15 个类别，其中示例是从它们的训练数据中选择出来作为生成合成数据的提示。给定这些组合中的每一种，除了特定类别的目标指标外，我们还测量其他性能指标，包括两个总体指标（总体平衡精度（OBA）和总体 F1 分数（OF1）），以及两个基于类别的指标（召回率（CR）和平衡精度（CBA））在所有类别上的表现。表格 3 显示了 180 种组合中 15 种组合的交叉影响结果，即 SW 策略、ϕ^{cr} （基于类别的召回率）和 15 个类别。在本附录中，我们在表格 10 ∼ 20 中提供了其他 15×11 种组合的交叉影响结果。

此外，从这些表格中，我们还可以观察到许多有趣的现象，例如：

- 大多数被针对的类别，即这些表中的黄色单元格，受益于改进过程，这些单元格中的数值为正（黑色数字）。黄色单元格中的红色数字通常不常见。
- T12 的召回率通常会显著提高，即使它不是目标类，例如，在表格 3 、12 、13 、15 、16 、17 、18 、19 和 20 中，Δ CR % T12 列中的绿色条很明显。

最重要的是，这些表格中的数据为第 **??** 节中的集成算法提供了知识图谱。

Table 12: 类似于表 3 和表 10 、 ∼ 、 11 ， 本文展示了 SW 策略和目标指标 ϕ^{cba} （基于类别的平衡准确率）对于票务数据集中类别 $T1, T2, \ldots, T15$ 的影响。

	Optimization: SW-CBA-Class[row]; Performance Metric: ΔOF1%, ΔCR %																Optimization: SW-CBA-Class[row]; Performance Metric: ΔOBA%, ΔCBA %														
	ΔOF1	ΔCR %															ΔOBA	ΔCBA %													
Class	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	-0.1	-0.1	0.1	0.4	1.3	-0.5	-1.2	1.3	0.0	-1.0	-1.7	0.0	0.0	-6.7	-2.7	-3.8	0.0	0.0	0.1	0.2	0.6	-0.3	-0.5	0.6	0.1	-0.4	-0.6	0.0	0.0	-1.7	-1.0	-1.2
T2	0.6	-0.2	0.7	-0.1	1.8	0.5	0.8	3.2	0.0	2.0	5.5	0.0	16.1	-13.3	-1.4	-7.5	0.3	-0.1	0.5	-0.1	0.8	0.2	0.3	1.5	0.1	0.9	2.0	0.0	5.2	-3.3	-0.5	-2.4
T3	-0.4	0.1	-0.5	1.9	0.0	-2.6	-1.6	-1.6	0.0	-0.5	-1.1	0.0	0.0	-6.7	-5.4	-3.8	-0.2	0.0	-0.1	0.6	0.0	-1.3	-0.8	-0.6	0.1	-0.2	-0.4	0.0	0.0	-1.7	-1.9	-1.2
T4	0.1	-0.4	-0.4	0.0	8.0	-0.9	1.2	2.9	1.3	-2.0	2.8	0.0	-3.2	6.7	-1.4	-9.4	0.1	-0.2	-0.1	0.0	3.4	-0.4	0.5	1.3	0.6	-0.8	0.9	0.0	-1.0	1.7	-0.5	-3.0
T5	0.6	-0.2	0.3	0.2	0.0	5.7	5.2	0.0	3.8	0.0	0.6	0.0	-3.2	0.0	-2.7	-7.5	0.3	-0.1	0.4	0.1	0.0	2.1	2.3	0.1	1.6	0.1	0.3	0.0	-1.1	0.0	-1.0	-2.4
T6	0.3	-0.4	0.0	-0.2	1.8	0.9	5.6	1.0	0.6	-1.0	1.1	0.0	12.9	-6.7	1.4	-5.7	0.1	-0.2	0.1	-0.1	0.8	0.4	2.3	0.4	0.3	-0.4	0.5	0.0	4.2	-1.7	0.5	-1.8
T7	-0.3	-0.2	-0.4	-0.6	0.4	-0.2	-2.4	5.8	0.6	-0.5	-2.8	0.0	-6.5	0.0	-5.4	-7.5	-0.1	0.0	-0.2	-0.3	0.2	-0.2	-1.1	2.6	0.3	-0.2	-1.0	0.0	-2.1	0.0	-1.9	-2.3
T8	-0.1	-0.1	-0.7	0.1	0.4	-1.9	-1.2	1.0	10.2	0.5	2.2	0.0	0.0	6.7	-4.1	-5.7	-0.1	0.0	-0.2	0.1	0.2	-0.8	-0.4	0.4	3.9	0.1	0.8	0.0	0.0	1.7	-1.4	-1.8
T9	0.2	-0.5	0.4	0.4	1.8	0.5	-0.4	1.6	-1.3	4.4	2.2	0.0	-3.2	-6.7	-4.1	-5.7	0.1	-0.2	0.3	0.1	0.8	0.2	-0.1	0.8	-0.5	1.7	0.9	0.0	-1.0	-1.7	-1.5	-1.8
T10	-0.1	-0.4	-0.5	0.5	-0.4	-0.9	-0.8	1.9	-3.2	-2.9	12.2	0.0	-6.5	6.7	1.4	-5.7	0.0	-0.2	0.0	0.2	-0.2	-0.3	-0.2	0.8	-1.2	-1.1	3.7	0.0	-2.1	1.6	0.5	-1.8
T11	-0.1	-0.2	0.0	0.5	0.4	-0.9	-3.2	1.3	0.6	-1.0	2.2	1.1	-3.2	-6.7	1.4	-3.8	0.0	-0.1	0.1	0.2	0.2	-0.4	-1.3	0.5	0.3	-0.4	0.8	0.6	-1.1	-1.7	0.5	-1.2
T12	0.4	-0.2	0.5	0.2	0.4	0.7	-1.2	1.6	4.5	-2.4	4.4	0.0	29.0	0.0	-4.1	-7.5	0.2	-0.1	0.4	0.1	0.2	0.3	-0.4	0.7	1.8	-1.0	1.6	0.0	9.4	0.0	-1.5	-2.4
T13	0.2	-0.5	0.0	0.0	1.8	0.2	0.8	3.9	0.0	-2.0	4.4	0.0	-6.5	26.7	-1.4	-9.4	0.1	-0.2	0.1	0.0	0.8	0.1	0.4	1.7	0.0	-0.9	1.6	0.0	-2.1	6.6	-0.5	-3.0
T14	-0.1	-0.2	-0.1	0.0	-0.4	1.2	-1.6	0.0	-1.3	-1.0	-1.1	0.0	3.2	0.0	13.5	-5.7	0.0	-0.1	0.2	0.0	-0.2	0.5	-0.5	0.0	-0.4	-0.4	-0.4	0.0	1.1	0.0	4.2	-1.8
T15	0.1	0.0	0.0	-0.4	1.8	0.2	-1.2	-1.0	1.9	-2.0	3.9	0.0	-3.2	0.0	-2.7	20.8	0.1	0.0	0.1	-0.2	0.8	0.1	-0.4	-0.3	0.8	-0.8	1.3	0.0	-1.0	0.0	-1.0	6.3
M0	*0.86*	*0.98*	*0.94*	*0.91*	*0.80*	*0.79*	*0.66*	*0.78*	*0.67*	*0.74*	*0.54*	*0.95*	*0.48*	*0.33*	*0.55*	*0.45*	*0.92*	*0.99*	*0.95*	*0.95*	*0.90*	*0.89*	*0.82*	*0.88*	*0.83*	*0.86*	*0.76*	*0.97*	*0.74*	*0.67*	*0.77*	*0.73*

Table 13: 与表格 3 和表格 10 ∼ 12 类似，本表展示了 HSW 策略和目标度量 ϕ^{cba} （基于类别的平衡精度）对票务数据集中类别 $T1, T2, \ldots, T15$ 的影响。

	Optimization: HSW-CBA-Class[row]; Performance Metric: ΔOF1%, ΔCR %																Optimization: HSW-CBA-Class[row]; Performance Metric: ΔOBA%, ΔCBA %														
	ΔOF1	ΔCR %															ΔOBA	ΔCBA %													
Class	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	-0.1	-0.1	0.1	0.4	1.3	-0.5	-1.2	1.3	0.0	-1.0	-1.7	0.0	0.0	-6.7	-2.7	-3.8	0.0	0.0	0.1	0.2	0.6	-0.3	-0.5	0.6	0.1	-0.4	-0.6	0.0	0.0	-1.7	-1.0	-1.2
T2	0.7	0.1	0.6	0.2	2.7	0.9	0.4	2.9	3.8	-1.5	5.0	0.0	12.9	-13.3	1.4	-3.8	0.4	0.0	0.5	0.1	1.1	0.4	0.3	1.4	1.5	-0.4	1.7	0.0	4.2	-3.3	0.5	-1.2
T3	0.3	-0.2	-0.2	1.4	-0.4	0.0	1.6	2.3	1.9	-0.5	1.7	0.0	3.2	0.0	0.0	-3.8	0.1	-0.1	0.0	0.6	-0.2	0.1	0.6	1.1	0.7	-0.2	0.6	0.0	1.0	0.0	0.0	-1.2
T4	0.1	-0.2	-0.5	0.0	9.8	-0.7	-2.8	1.6	1.9	-2.9	5.5	0.0	9.7	-13.3	-2.7	-7.5	0.0	-0.1	0.0	0.0	4.1	-0.3	-1.1	0.7	0.8	-1.2	1.9	0.0	3.2	-3.3	-0.9	-2.4
T5	0.3	-0.5	0.4	0.4	0.9	4.3	0.0	1.3	0.0	-2.9	0.6	0.0	6.5	-6.7	1.4	-3.8	0.2	-0.2	0.3	0.1	0.4	1.6	0.2	0.7	0.1	-1.2	0.1	0.0	2.1	-1.7	0.5	-1.2
T6	-0.3	-0.1	-0.6	-0.4	-0.4	-1.4	4.8	0.6	-1.3	-1.5	-0.6	0.0	0.0	-6.7	-1.4	-7.5	-0.2	0.0	-0.3	-0.2	-0.2	-0.6	1.8	0.3	-0.5	-0.6	-0.2	0.0	0.0	-1.7	-0.5	-2.4
T7	0.7	-0.2	0.7	0.2	2.7	1.7	1.2	5.8	3.2	-3.4	3.9	0.0	6.5	-6.7	-1.4	-5.7	0.4	-0.1	0.5	0.1	1.2	0.8	0.6	2.6	1.3	-1.4	1.4	0.0	2.1	-1.7	-0.5	-1.8
T8	-0.2	-0.2	-0.6	-0.7	1.3	-0.9	-1.2	-0.6	11.5	0.5	1.7	0.0	-6.5	-6.7	-4.1	-7.5	-0.1	-0.1	-0.3	-0.4	0.6	-0.3	-0.5	-0.3	4.3	0.3	0.6	0.0	-2.1	-1.7	-1.4	-2.3
T9	0.4	-0.2	0.4	0.1	1.3	1.9	-2.0	1.9	-2.5	5.4	4.4	0.0	9.7	0.0	-6.8	-5.7	0.2	-0.1	0.2	0.1	0.6	0.9	-0.7	0.9	-1.0	2.2	1.7	0.0	3.2	0.0	-2.4	-1.8
T10	0.1	-0.1	-0.1	0.4	-1.3	0.0	0.8	0.0	-2.5	0.5	11.0	0.0	-6.5	0.0	-6.8	-7.5	0.0	0.0	0.2	0.2	-0.6	-0.1	0.5	0.1	-0.9	0.2	3.2	0.0	-2.1	0.0	-2.4	-2.3
T11	0.6	-0.3	0.1	0.9	2.7	1.9	0.4	2.6	2.5	-2.9	9.4	1.1	3.2	-6.7	2.7	-9.4	0.3	-0.2	0.4	0.4	1.2	0.9	0.3	1.2	1.1	-1.1	3.2	0.5	1.0	-1.7	0.9	-3.0
T12	0.3	-0.2	0.0	0.4	3.6	0.9	-2.4	1.3	-0.6	0.0	3.9	0.0	32.3	-6.7	-1.4	-7.5	0.1	-0.1	0.2	0.1	1.5	0.4	-0.8	0.6	-0.1	0.0	1.3	0.0	10.4	-1.7	-0.4	-2.4
T13	0.0	-0.3	-0.2	0.0	0.9	0.7	-3.2	0.3	3.2	0.5	2.2	0.0	-3.2	33.3	0.0	-1.9	0.0	-0.1	-0.1	0.0	0.4	0.3	-1.3	0.2	1.2	0.2	0.8	0.0	-1.0	8.2	0.0	-0.6
T14	-1.0	-0.2	-1.1	-0.5	2.2	-1.2	-3.6	0.0	-6.4	-4.4	-5.0	0.0	-12.9	-6.7	16.2	-11.3	-0.5	-0.1	-0.4	-0.3	1.0	-0.5	-1.5	0.0	-2.5	-1.9	-1.8	0.0	-4.2	-1.7	5.1	-3.5
T15	0.2	-0.3	-0.3	0.5	1.8	0.0	-4.4	2.6	1.9	-1.0	3.3	0.0	9.7	-13.3	1.4	20.8	0.1	-0.1	0.1	0.2	0.8	-0.1	-1.7	1.1	0.8	-0.3	1.1	0.0	3.1	-3.3	0.5	6.1
M0	*0.86*	*0.98*	*0.94*	*0.91*	*0.80*	*0.79*	*0.66*	*0.78*	*0.67*	*0.74*	*0.54*	*0.95*	*0.48*	*0.33*	*0.55*	*0.45*	*0.92*	*0.99*	*0.95*	*0.95*	*0.90*	*0.89*	*0.82*	*0.88*	*0.83*	*0.86*	*0.76*	*0.97*	*0.74*	*0.67*	*0.77*	*0.73*

Table 14: 与表格 3 和表格 10 ∼ 13 类似，此表格展示了 GA 策略和目标指标 ϕ^{cba} （基于类别的平衡准确率）对票务数据集中类别 $T1, T2, \ldots, T15$ 的影响。

	Optimization: GA-CBA-Class[row]; Performance Metric: ΔOF1%, ΔCR %																Optimization: GA-CBA-Class[row]; Performance Metric: ΔOBA%, ΔCBA %														
	ΔOF1	ΔCR %															ΔOBA	ΔCBA %													
Class	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	0.3	0.2	-0.5	0.9	0.0	-0.5	0.0	2.3	1.9	2.9	3.9	0.0	12.9	-20.0	-4.1	-1.9	0.1	-0.2	0.0	0.4	0.0	-0.1	0.1	1.0	0.8	1.3	1.3	0.0	4.2	-5.0	-1.4	-0.6
T2	0.0	0.4	0.8	-0.5	-0.9	0.9	-2.8	-0.6	-5.7	1.0	5.0	0.0	3.2	-13.3	-2.7	-17.0	0.0	0.2	-0.1	-0.2	-0.4	0.4	-0.9	-0.2	-2.3	0.6	1.8	0.0	1.0	-3.3	-1.0	-5.3
T3	0.3	-0.2	0.0	2.0	1.3	0.5	0.8	1.0	-0.6	-1.5	3.9	0.0	-6.5	-6.7	-5.4	-5.7	0.1	-0.1	-0.1	0.7	0.6	0.2	0.3	0.5	-0.2	-0.5	1.6	0.0	-2.1	-1.7	-1.9	-1.7
T4	0.4	0.0	0.0	1.1	8.4	-0.7	0.8	1.0	-2.5	0.5	3.3	0.0	-3.2	0.0	-2.7	-3.8	0.2	0.1	0.2	0.5	3.6	-0.3	0.2	0.5	-0.9	0.2	1.3	0.0	-1.1	0.0	-0.9	-1.2
T5	0.2	-0.3	-0.3	0.9	0.4	2.8	1.2	1.0	-2.5	2.4	3.3	0.0	3.2	-20.0	-1.4	-9.4	0.1	-0.1	0.3	0.6	0.2	0.7	0.6	0.4	-0.9	1.1	1.2	0.0	1.0	-5.0	-0.5	-3.0
T6	-0.2	-0.1	-0.4	0.5	-0.4	0.5	4.0	0.3	-3.8	-2.4	-3.3	-1.1	19.4	-13.3	-2.7	-7.5	-0.1	0.0	0.0	0.2	-0.2	0.2	0.9	0.1	-1.4	-0.9	-1.1	-0.5	6.3	-3.3	-0.9	-2.4
T7	0.0	0.0	0.0	0.4	0.0	0.5	2.4	-0.3	1.9	-2.0	1.7	0.0	-3.2	0.0	-6.8	-5.7	0.0	0.0	0.1	0.2	0.1	0.1	0.9	-0.3	0.8	-0.8	0.6	0.0	-1.1	0.0	-2.4	-1.8
T8	-0.4	-0.2	-0.5	0.1	-0.9	1.2	-3.2	0.6	5.1	-2.0	-4.4	0.0	3.2	-6.7	0.0	-9.4	-0.2	-0.1	-0.2	0.1	-0.4	0.5	-1.3	0.3	1.7	-0.8	-1.5	0.0	1.0	-1.7	0.0	-3.0
T9	-0.1	-0.2	0.1	0.2	-0.4	-0.2	0.0	1.3	-3.2	2.9	-2.2	0.0	6.5	0.0	-1.4	-9.4	0.0	-0.1	0.2	0.1	-0.2	-0.1	0.0	0.6	-1.2	0.9	-0.7	0.0	2.1	0.0	-0.4	-3.0
T10	0.6	-0.2	-0.3	0.6	-0.4	4.0	2.0	1.6	-3.8	3.9	9.4	0.0	3.2	0.0	2.7	-7.5	0.3	-0.1	0.2	0.3	-0.2	1.8	0.9	0.8	-1.4	1.8	2.7	0.0	1.1	0.0	1.0	-2.3
T11	0.2	-0.1	0.1	0.5	0.0	-1.9	3.2	2.6	1.9	-2.4	3.3	1.1	-3.2	0.0	0.0	-11.3	0.1	0.0	0.1	0.2	0.0	-0.9	1.2	1.1	0.8	-0.9	1.1	0.5	-1.0	0.0	0.0	-3.6
T12	0.1	-0.4	-0.2	0.2	1.8	1.7	-0.4	1.0	0.0	-2.0	1.7	0.0	32.3	6.7	-1.4	-5.7	0.1	-0.2	0.1	0.1	0.8	0.7	-0.2	0.4	0.1	-0.9	0.6	0.0	10.4	1.7	-0.5	-1.8
T13	0.1	-0.6	0.1	0.2	1.3	-1.4	-2.0	5.2	1.9	-2.4	1.1	0.0	3.2	33.3	-1.4	-3.8	0.0	-0.3	0.3	0.1	0.5	-0.6	-0.8	2.2	0.7	-1.0	0.4	0.0	1.0	8.2	-0.4	-1.2
T14	0.0	-0.2	0.1	0.1	0.4	-0.9	-0.8	0.6	-2.5	0.0	-1.7	0.0	0.0	-13.3	16.2	-5.7	0.0	-0.1	0.2	0.1	0.2	-0.3	-0.2	0.3	-1.0	0.1	-0.5	0.0	0.0	-3.3	5.1	-1.8
T15	0.1	-0.4	0.0	0.4	1.3	0.2	-1.6	1.0	-3.8	-1.5	2.8	0.0	3.2	-6.7	0.0	24.5	0.1	-0.1	0.2	0.1	0.6	0.1	-0.5	0.4	-1.5	-0.6	0.9	0.0	1.1	-1.7	0.0	7.3
M0	*0.86*	*0.98*	*0.94*	*0.91*	*0.80*	*0.79*	*0.66*	*0.78*	*0.67*	*0.74*	*0.54*	*0.95*	*0.48*	*0.33*	*0.55*	*0.45*	*0.92*	*0.99*	*0.95*	*0.95*	*0.90*	*0.89*	*0.82*	*0.88*	*0.83*	*0.86*	*0.76*	*0.97*	*0.74*	*0.67*	*0.77*	*0.73*

Table 15: 与表格 3 和表格 10 ∼ 14 类似，本表展示了在票务数据集中，SW 策略和目标指标 ϕ^{oba} （总体平衡准确率）对类别 $T1, T2, \ldots, T15$ 的影响。

	Optimization: SW-OBA-Class[row]; Performance Metric: ΔOF1%, ΔCR %																Optimization: SW-OBA-Class[row]; Performance Metric: ΔOBA%, ΔCBA %														
	ΔOF1	ΔCR %															ΔOBA	ΔCBA %													
Class	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	0.6	-0.1	0.6	0.2	2.7	0.2	0.0	1.3	4.5	-1.0	6.1	0.0	16.1	0.0	0.0	-7.5	0.3	-0.1	0.5	0.1	1.2	0.1	0.1	0.6	1.9	-0.4	2.1	0.0	5.3	0.0	0.0	-2.4
T2	0.7	0.1	0.0	0.0	4.9	2.4	2.0	2.3	-0.6	0.0	3.3	0.0	9.7	0.0	2.7	-1.9	0.3	0.1	0.1	0.0	2.2	1.1	0.8	1.1	-0.2	0.1	1.1	0.0	3.2	0.0	1.0	-0.6
T3	0.7	-0.1	0.1	0.9	3.6	2.1	-0.4	3.6	1.9	0.0	5.5	0.0	3.2	-6.7	-2.7	-3.8	0.3	0.0	0.3	0.4	1.6	0.9	0.0	1.6	0.8	0.0	1.9	0.0	1.1	-1.7	-1.0	-1.2
T4	0.6	-0.1	0.3	0.1	7.6	-0.5	1.6	3.2	2.5	-1.5	3.9	0.0	6.5	-6.7	0.0	-3.8	0.3	0.0	0.3	0.0	3.2	-0.1	0.7	1.5	1.1	-0.5	1.3	0.0	2.1	-1.7	0.0	-1.2
T5	0.7	-0.4	0.6	0.1	3.6	3.3	-1.2	3.2	1.3	0.0	6.6	0.0	3.2	-13.3	-1.4	-3.8	0.3	-0.1	0.4	0.1	1.6	1.4	-0.4	1.5	0.6	0.0	2.4	0.0	1.1	-3.3	-0.5	-1.2
T6	0.8	0.0	0.7	0.6	1.8	1.9	2.4	3.2	1.3	-2.0	2.8	0.0	9.7	-13.3	0.0	-1.9	0.4	0.0	0.5	0.3	0.8	0.9	1.0	1.5	0.5	-0.7	1.0	0.0	3.1	-3.3	0.0	-0.6
T7	0.8	-0.1	0.6	0.0	1.3	0.9	0.0	5.8	3.8	0.5	3.3	0.0	16.1	0.0	-1.4	-1.9	0.4	0.0	0.5	0.0	0.6	0.5	0.1	2.5	1.6	0.3	1.2	0.0	5.3	0.0	-0.5	-0.6
T8	0.7	-0.2	0.7	0.1	0.4	0.9	-2.0	3.6	2.5	1.5	7.7	0.0	0.0	-13.3	2.7	0.0	0.3	-0.1	0.5	0.1	0.2	0.4	-0.7	1.6	1.0	0.6	2.8	0.0	0.0	-3.3	1.0	0.0
T9	0.8	-0.1	0.7	0.7	2.2	2.1	3.6	1.0	3.8	-2.0	3.3	0.0	3.2	0.0	4.1	-7.5	0.4	0.0	0.6	0.4	1.0	0.9	1.6	0.5	1.6	-1.0	1.1	0.0	1.1	0.0	1.5	-2.4
T10	0.7	0.0	0.4	-0.1	3.1	1.7	3.2	1.3	0.6	0.5	5.5	0.0	16.1	0.0	0.0	-1.9	0.4	0.0	0.5	-0.1	1.4	0.8	1.5	0.6	0.3	0.3	1.7	0.0	5.3	0.0	0.0	-0.6
T11	0.7	-0.1	0.7	0.6	3.1	0.0	2.4	2.3	1.9	-1.0	3.3	0.0	12.9	0.0	-2.7	-5.7	0.3	0.0	0.5	0.3	1.4	0.0	0.9	1.1	0.8	-0.4	1.2	0.0	4.2	0.0	-1.0	-1.8
T12	0.8	-0.1	0.6	0.5	4.0	1.9	-0.4	3.6	1.3	0.0	1.1	0.0	25.8	-6.7	1.4	-1.9	0.4	0.0	0.5	0.2	1.8	0.8	0.0	1.6	0.6	0.1	0.4	0.0	8.4	-1.7	0.4	-0.6
T13	0.7	-0.4	0.4	0.2	3.1	1.4	-1.2	3.6	1.3	1.5	3.3	0.0	9.7	20.0	1.4	1.9	0.3	-0.2	0.4	0.1	1.3	0.7	-0.3	1.6	0.6	0.6	1.2	0.0	3.2	4.9	0.4	0.6
T14	0.5	-0.2	0.2	0.7	3.6	2.1	0.8	1.0	1.9	-2.0	3.9	0.0	3.2	-6.7	5.4	-7.5	0.3	-0.1	0.2	0.3	1.6	1.0	0.4	0.5	0.8	-0.8	1.4	0.0	1.1	-1.7	1.8	-2.4
T15	0.7	-0.1	0.7	0.7	1.8	0.0	0.8	1.6	0.6	-0.5	6.1	0.0	19.4	-6.7	0.0	3.8	0.4	0.0	0.5	0.4	0.8	0.0	0.4	0.8	0.3	-0.2	2.2	0.0	6.3	-1.7	0.0	1.1
M0	*0.86*	*0.98*	*0.94*	*0.91*	*0.80*	*0.79*	*0.66*	*0.78*	*0.67*	*0.74*	*0.54*	*0.95*	*0.48*	*0.33*	*0.55*	*0.45*	*0.92*	*0.99*	*0.95*	*0.95*	*0.90*	*0.89*	*0.82*	*0.88*	*0.83*	*0.86*	*0.76*	*0.97*	*0.74*	*0.67*	*0.77*	*0.73*

Table 16: 与表 3 和表 10 ∼ 15 类似，这个表格展示了 HSW 策略和目标度量 ϕ^{oba} （总体平衡准确率）对于票务数据集中类 $T1, T2, \ldots, T15$ 的影响。

	Optimization: HSW-OBA-Class[row]; Performance Metric: ΔOF1%, ΔCR %																Optimization: HSW-OBA-Class[row]; Performance Metric: ΔOBA%, ΔCBA %														
	ΔOF1	ΔCR %															ΔOBA	ΔCBA %													
Class	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	0.6	-0.3	0.6	0.4	2.2	1.7	0.4	1.9	3.8	-3.4	3.9	0.0	19.4	-6.7	2.7	-1.9	0.3	-0.1	0.5	0.2	1.0	0.7	0.3	0.9	1.5	-1.4	1.4	0.0	6.3	-1.7	0.9	-0.6
T2	0.7	0.1	0.6	0.2	2.7	0.9	0.4	2.9	3.8	-1.5	5.0	0.0	12.9	-13.3	1.4	-3.8	0.4	0.0	0.5	0.1	1.1	0.4	0.3	1.4	1.5	-0.4	1.7	0.0	4.2	-3.3	0.5	-1.2
T3	0.6	-0.1	0.3	0.4	3.1	2.6	2.0	1.6	-1.9	1.0	2.8	0.0	6.5	0.0	0.0	-5.7	0.3	0.0	0.3	0.1	1.3	1.1	0.8	0.8	-0.7	0.6	1.1	0.0	2.1	0.0	0.0	-1.8
T4	0.7	-0.4	0.4	-0.1	4.4	1.9	-0.4	4.9	0.6	2.0	6.1	0.0	12.9	-6.7	-1.4	-5.7	0.4	-0.1	0.5	0.0	1.9	0.9	0.0	2.2	0.2	0.9	2.2	0.0	4.2	-1.7	-0.5	-1.8
T5	0.7	-0.1	0.8	0.2	0.4	1.9	0.8	3.2	0.6	0.5	6.1	0.0	3.2	-6.7	-2.7	1.9	0.4	0.0	0.6	0.1	0.2	0.8	0.4	1.5	0.3	0.2	2.3	0.0	1.1	-1.7	-0.9	0.6
T6	0.7	-0.4	0.6	0.2	0.9	-0.5	5.6	4.2	1.9	0.5	3.9	0.0	9.7	6.7	-5.4	-3.8	0.3	-0.1	0.4	0.2	0.4	-0.2	2.3	1.8	0.8	0.2	1.4	0.0	3.2	1.7	-1.9	-1.2
T7	0.7	-0.2	0.7	0.2	2.7	1.7	1.2	5.8	3.2	-3.4	3.9	0.0	6.5	-6.7	-1.4	-5.7	0.4	-0.1	0.5	0.1	1.2	0.8	0.6	2.6	1.3	-1.4	1.4	0.0	2.1	-1.7	-0.5	-1.8
T8	0.6	0.1	0.3	0.7	2.2	1.7	-0.4	1.9	5.7	1.5	2.8	0.0	3.2	-6.7	0.0	-7.5	0.3	0.0	0.5	0.4	1.0	0.8	-0.1	0.9	2.3	0.7	0.8	0.0	1.1	-1.7	0.0	-2.3
T9	0.6	-0.3	0.8	0.6	2.2	0.7	-1.2	3.9	4.5	-1.0	3.3	0.0	3.2	-6.7	0.0	-9.4	0.3	-0.1	0.7	0.3	1.0	0.3	-0.4	1.8	1.8	-0.4	1.2	0.0	1.1	-1.7	-0.1	-2.9
T10	1.0	-0.1	0.5	1.1	4.0	2.1	0.4	3.9	3.8	-1.5	3.3	0.0	9.7	0.0	1.4	1.9	0.5	-0.1	0.5	0.5	1.8	1.0	0.3	1.8	1.6	-0.6	1.2	0.0	3.1	0.0	0.5	0.6
T11	0.7	-0.1	0.5	1.1	2.2	-0.2	-1.2	2.3	4.5	-1.0	8.3	0.0	3.2	0.0	0.0	-5.7	0.3	-0.1	0.5	0.5	1.0	-0.1	-0.4	1.1	1.8	-0.4	2.9	0.0	1.1	0.0	0.0	-1.8
T12	0.7	-0.2	0.6	0.1	3.6	0.5	-1.2	2.3	4.5	0.5	7.7	0.0	19.4	0.0	-1.4	-7.5	0.3	-0.1	0.5	0.1	1.6	0.2	-0.3	1.0	1.9	0.2	2.8	0.0	6.3	0.0	-0.5	-2.4
T13	0.8	-0.2	0.4	0.7	2.7	1.7	1.6	3.6	3.2	-2.0	4.4	0.0	9.7	20.0	-1.4	0.0	0.4	0.0	0.4	0.4	1.2	0.7	0.8	1.6	1.3	-0.8	1.6	0.0	3.2	4.8	-0.5	0.0
T14	0.6	-0.4	0.2	0.5	3.1	1.7	-2.4	3.9	3.2	0.5	5.0	0.0	16.1	0.0	0.0	0.0	0.3	-0.2	0.3	0.2	1.4	0.6	-0.9	1.7	1.4	0.2	1.8	0.0	5.3	0.0	0.0	0.0
T15	0.7	-0.1	0.9	0.4	1.3	-0.9	-2.0	2.6	1.9	2.0	4.4	0.0	16.1	-6.7	0.0	11.3	0.4	0.0	0.7	0.2	0.6	-0.4	-0.6	1.2	0.8	0.9	1.6	0.0	5.2	-1.7	0.0	3.3
M0	*0.86*	*0.98*	*0.94*	*0.91*	*0.80*	*0.79*	*0.66*	*0.78*	*0.67*	*0.74*	*0.54*	*0.95*	*0.48*	*0.33*	*0.55*	*0.45*	*0.92*	*0.99*	*0.95*	*0.95*	*0.90*	*0.89*	*0.82*	*0.88*	*0.83*	*0.86*	*0.76*	*0.97*	*0.74*	*0.67*	*0.77*	*0.73*

Table 17: 类似于表 3 和表 10 ∼ 16 ，本表展示了 GA 策略和目标度量标准 ϕ^{oba} （整体平衡准确度）对票务数据集中类别 $T1, T2, \ldots, T15$ 的影响。

	Optimization: GA-OBA-Class[row]; Performance Metric: ΔOF1%, ΔCR %																Optimization: GA-OBA-Class[row]; Performance Metric: ΔOBA%, ΔCBA %														
	ΔOF1	ΔCR %															ΔOBA	ΔCBA %													
Class	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	0.3	0.2	-0.5	0.9	0.0	-0.5	0.0	2.3	1.9	2.9	3.9	0.0	12.9	-20.0	-4.1	-1.9	0.1	-0.2	0.0	0.4	0.0	-0.1	0.1	1.0	0.8	1.3	1.3	0.0	4.2	-5.0	-1.4	-0.6
T2	0.2	0.1	0.7	-0.4	0.9	1.2	-3.2	2.3	-4.5	-2.4	4.4	0.0	9.7	-13.3	-2.7	-1.9	0.1	0.1	-0.2	-0.1	0.4	0.5	-1.1	1.0	-1.8	-0.8	1.6	0.0	3.1	-3.3	-1.0	-0.6
T3	0.6	0.1	0.4	1.2	3.6	1.2	0.4	1.9	-2.5	-0.5	7.2	0.0	3.2	-6.7	-5.4	-9.4	0.3	0.1	0.3	0.4	1.5	0.5	0.2	0.9	-0.9	-0.1	2.7	0.0	1.0	-1.6	-1.9	-2.9
T4	0.8	0.0	0.6	0.2	6.2	1.4	2.4	0.3	-0.6	2.0	5.0	0.0	6.5	-6.7	-1.4	-3.8	0.4	0.0	0.5	0.2	2.6	0.7	1.0	0.2	-0.1	0.8	1.9	0.0	2.1	-1.7	-0.5	-1.2
T5	0.2	-0.3	-0.3	0.9	0.4	2.8	1.2	1.0	-2.5	2.4	3.3	0.0	3.2	-20.0	-1.4	-9.4	0.1	-0.1	0.3	0.6	0.2	0.7	0.6	0.4	-0.9	1.1	1.2	0.0	1.0	-5.0	-0.5	-3.0
T6	0.3	0.3	0.0	0.6	2.2	2.6	2.0	1.6	-1.3	-2.0	0.0	0.0	12.9	-13.3	-2.7	-15.1	0.2	0.1	0.2	0.3	1.0	1.1	0.2	0.9	-0.4	-0.8	0.2	0.0	4.2	-3.3	-0.9	-4.7
T7	0.3	0.0	0.3	0.1	2.7	0.9	2.0	-4.5	0.6	2.4	1.7	0.0	9.7	-13.3	-4.1	1.9	0.1	0.0	0.5	0.1	1.2	0.4	0.8	-2.2	0.3	1.1	0.5	0.0	3.1	-3.3	-1.5	0.5
T8	0.4	0.2	-0.1	0.6	1.8	1.7	-2.4	2.3	1.3	2.0	4.4	0.0	3.2	-6.7	0.0	-7.5	0.2	0.1	0.2	0.3	0.8	0.8	-0.8	1.1	0.3	0.9	1.6	0.0	1.0	-1.7	0.0	-2.4
T9	0.5	-0.1	0.4	0.8	-0.4	1.4	0.0	1.6	0.0	-1.5	2.2	0.0	9.7	0.0	4.1	1.9	0.2	0.0	0.4	0.5	-0.2	0.6	0.1	0.9	0.0	-0.8	0.8	0.0	3.1	0.0	1.4	0.5
T10	0.6	-0.2	-0.3	0.6	-0.4	4.0	2.0	1.6	-3.8	3.9	9.4	0.0	3.2	0.0	2.7	-7.5	0.3	-0.1	0.2	0.3	-0.2	1.8	0.9	0.8	-1.4	1.8	2.7	0.0	1.1	0.0	1.0	-2.3
T11	0.9	-0.1	0.7	0.1	1.8	2.8	3.6	2.3	0.0	2.4	6.6	0.0	0.0	0.0	-1.4	-9.4	0.4	0.0	0.4	0.1	0.8	1.2	1.5	1.2	0.1	1.1	2.4	0.0	0.0	0.0	-0.5	-2.9
T12	0.7	0.0	0.1	-0.4	4.4	0.0	0.0	4.9	1.9	1.0	5.5	0.0	22.6	0.0	2.7	-1.9	0.3	0.0	0.1	-0.1	1.9	0.1	0.1	2.2	0.8	0.4	2.0	0.0	7.3	0.0	1.0	-0.6
T13	0.7	-0.1	0.6	0.6	3.1	1.2	-0.8	2.3	1.3	2.0	1.7	0.0	9.7	20.0	4.1	-7.5	0.3	0.0	0.5	0.3	1.4	0.5	-0.2	1.1	0.5	0.9	0.6	0.0	3.2	4.9	1.5	-2.4
T14	0.4	0.0	0.8	0.8	1.3	-1.4	0.4	2.3	-1.9	-2.4	1.7	0.0	6.5	-6.7	8.1	-3.8	0.2	0.0	0.5	0.4	0.6	-0.5	0.3	1.1	-0.7	-1.0	0.6	0.0	2.1	-1.7	2.4	-1.2
T15	0.7	0.1	0.0	0.5	2.7	1.9	0.8	0.6	5.1	-1.5	5.0	0.0	6.5	0.0	0.0	18.9	0.4	0.1	0.3	0.2	1.2	0.8	0.5	0.4	2.1	-0.5	1.8	0.0	2.1	0.0	0.0	5.6
M0	*0.86*	*0.98*	*0.94*	*0.91*	*0.80*	*0.79*	*0.66*	*0.78*	*0.67*	*0.74*	*0.54*	*0.95*	*0.48*	*0.33*	*0.55*	*0.45*	*0.92*	*0.99*	*0.95*	*0.95*	*0.90*	*0.89*	*0.82*	*0.88*	*0.83*	*0.86*	*0.76*	*0.97*	*0.74*	*0.67*	*0.77*	*0.73*

Table 18: 与表格 3 和表格 10 ∼ 17 类似，这个表展示了 SW 策略的影响以及在客票数据集中的类别 $T1, T2, \ldots, T15$ 的目标指标 ϕ^{of1} （总体 F1-分数）。

	Optimization: SW-OF1-Class[row]; Performance Metric: ΔOF1%, ΔCR %																Optimization: SW-OF1-Class[row]; Performance Metric: ΔOBA%, ΔCBA %														
	ΔOF1	ΔCR %															ΔOBA	ΔCBA %													
Class	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	0.6	-0.1	0.6	0.2	2.7	0.2	0.0	1.3	4.5	-1.0	6.1	0.0	16.1	0.0	0.0	-7.5	0.3	-0.1	0.5	0.1	1.2	0.1	0.1	0.6	1.9	-0.4	2.1	0.0	5.3	0.0	0.0	-2.4
T2	0.7	0.1	0.0	0.0	4.9	2.4	2.0	2.3	-0.6	0.0	3.3	0.0	9.7	0.0	2.7	-1.9	0.3	0.1	0.1	0.0	2.2	1.1	0.8	1.1	-0.2	0.1	1.1	0.0	3.2	0.0	1.0	-0.6
T3	0.7	-0.1	0.1	0.9	3.6	2.1	-0.4	3.6	1.9	0.0	5.5	0.0	3.2	-6.7	-2.7	-3.8	0.3	0.0	0.3	0.4	1.6	0.9	0.0	1.6	0.8	0.0	1.9	0.0	1.1	-1.7	-1.0	-1.2
T4	0.6	-0.1	0.3	0.1	7.6	-0.5	1.6	3.2	2.5	-1.5	3.9	0.0	6.5	-6.7	0.0	-3.8	0.3	0.0	0.3	0.0	3.2	-0.1	0.7	1.5	1.1	-0.5	1.3	0.0	2.1	-1.7	0.0	-1.2
T5	0.7	-0.4	0.6	0.1	3.6	3.3	-1.2	3.2	1.3	0.0	6.6	0.0	3.2	-13.3	-1.4	-3.8	0.3	-0.1	0.4	0.1	1.6	1.4	-0.4	1.5	0.6	0.0	2.4	0.0	1.1	-3.3	-0.5	-1.2
T6	0.8	0.0	0.7	0.6	1.8	1.9	2.4	3.2	1.3	-2.0	2.8	0.0	9.7	-13.3	0.0	-1.9	0.4	0.0	0.5	0.3	0.8	0.9	1.0	1.5	0.5	-0.7	1.0	0.0	3.1	-3.3	0.0	-0.6
T7	0.8	-0.1	0.6	0.0	1.3	0.9	0.0	5.8	3.8	0.5	3.3	0.0	16.1	0.0	-1.4	-1.9	0.4	0.0	0.5	0.0	0.6	0.5	0.1	2.5	1.6	0.3	1.2	0.0	5.3	0.0	-0.5	-0.6
T8	0.7	-0.2	0.7	0.1	0.4	0.9	-2.0	3.6	2.5	1.5	7.7	0.0	0.0	-13.3	2.7	0.0	0.3	-0.1	0.5	0.1	0.2	0.4	-0.7	1.6	1.0	0.6	2.8	0.0	0.0	-3.3	1.0	0.0
T9	0.8	-0.1	0.7	0.7	2.2	2.1	3.6	1.0	3.8	-2.0	3.3	0.0	3.2	0.0	4.1	-7.5	0.4	0.0	0.6	0.4	1.0	0.9	1.6	0.5	1.6	-1.0	1.1	0.0	1.1	0.0	1.5	-2.4
T10	0.7	0.0	0.4	-0.1	3.1	1.7	3.2	1.3	0.6	0.5	5.5	0.0	16.1	0.0	0.0	-1.9	0.4	0.0	0.5	-0.1	1.4	0.8	1.5	0.6	0.3	0.3	1.7	0.0	5.3	0.0	0.0	-0.6
T11	0.7	-0.1	0.7	0.6	3.1	0.0	2.4	2.3	1.9	-1.0	3.3	0.0	12.9	0.0	-2.7	-5.7	0.3	0.0	0.5	0.3	1.4	0.0	0.9	1.1	0.8	-0.4	1.2	0.0	4.2	0.0	-1.0	-1.8
T12	0.8	-0.1	0.6	0.5	4.0	1.9	-0.4	3.6	1.3	0.0	1.1	0.0	25.8	-6.7	1.4	-1.9	0.4	0.0	0.5	0.2	1.8	0.8	0.0	1.6	0.6	0.1	0.4	0.0	8.4	-1.7	0.4	-0.6
T13	0.7	-0.4	0.4	0.2	3.1	1.4	-1.2	3.6	1.3	1.5	3.3	0.0	9.7	20.0	1.4	1.9	0.3	-0.2	0.4	0.1	1.3	0.7	-0.3	1.6	0.6	0.6	1.2	0.0	3.2	4.9	0.4	0.6
T14	0.5	-0.2	0.2	0.7	3.6	2.1	0.8	1.0	1.9	-2.0	3.9	0.0	3.2	-6.7	5.4	-7.5	0.3	-0.1	0.2	0.3	1.6	1.0	0.4	0.5	0.8	-0.8	1.4	0.0	1.1	-1.7	1.8	-2.4
T15	0.7	-0.1	0.7	0.7	1.8	0.0	0.8	1.6	0.6	-0.5	6.1	0.0	19.4	-6.7	0.0	3.8	0.4	0.0	0.5	0.4	0.8	0.0	0.4	0.8	0.3	-0.2	2.2	0.0	6.3	-1.7	0.0	1.1
M0	*0.86*	*0.98*	*0.94*	*0.91*	*0.80*	*0.79*	*0.66*	*0.78*	*0.67*	*0.74*	*0.54*	*0.95*	*0.48*	*0.33*	*0.55*	*0.45*	*0.92*	*0.99*	*0.95*	*0.95*	*0.90*	*0.89*	*0.82*	*0.88*	*0.83*	*0.86*	*0.76*	*0.97*	*0.74*	*0.67*	*0.77*	*0.73*

Table 19: 与表格 3 和表格 10 ∼ 18 类似，此表显示了 HSW 策略和目标指标 ϕ^{of1} （总体 F1-得分）对工单数据集中的 $T1, T2, \ldots, T15$ 类的影响。

	Optimization: HSW-OF1-Class[row]; Performance Metric: ΔOF1%, ΔCR %																Optimization: HSW-OF1-Class[row]; Performance Metric: ΔOBA%, ΔCBA %														
	ΔOF1	ΔCR %															ΔOBA	ΔCBA %													
Class	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	0.6	-0.3	0.6	0.4	2.2	1.7	0.4	1.9	3.8	-3.4	3.9	0.0	19.4	-6.7	2.7	-1.9	0.3	-0.1	0.5	0.2	1.0	0.7	0.3	0.9	1.5	-1.4	1.4	0.0	6.3	-1.7	0.9	-0.6
T2	0.6	-0.1	0.4	0.1	3.1	0.9	3.2	1.3	-0.6	-2.0	7.2	0.0	9.7	-6.7	-1.4	0.0	0.3	-0.1	0.2	0.0	1.4	0.5	1.4	0.7	-0.2	-0.8	2.5	0.0	3.2	-1.7	-0.5	0.0
T3	0.6	-0.1	0.3	0.4	3.1	2.6	2.0	1.6	-1.9	1.0	2.8	0.0	6.5	0.0	0.0	-5.7	0.3	0.0	0.3	0.1	1.3	1.1	0.8	0.8	-0.7	0.6	1.1	0.0	2.1	0.0	0.0	-1.8
T4	0.7	-0.4	0.4	-0.1	4.4	1.9	-0.4	4.9	0.6	2.0	6.1	0.0	12.9	-6.7	-1.4	-5.7	0.4	-0.1	0.5	0.0	1.9	0.9	0.0	2.2	0.2	0.9	2.2	0.0	4.2	-1.7	-0.5	-1.8
T5	0.7	-0.1	0.8	0.2	0.4	1.9	0.8	3.2	0.6	0.5	6.1	0.0	3.2	-6.7	-2.7	1.9	0.4	0.0	0.6	0.1	0.2	0.8	0.4	1.5	0.3	0.2	2.3	0.0	1.1	-1.7	-0.9	0.6
T6	0.7	-0.4	0.6	0.2	0.9	-0.5	5.6	4.2	1.9	0.5	3.9	0.0	9.7	6.7	-5.4	-3.8	0.3	-0.1	0.4	0.2	0.4	-0.2	2.3	1.8	0.8	0.2	1.4	0.0	3.2	1.7	-1.9	-1.2
T7	0.7	-0.2	0.7	0.2	2.7	1.7	1.2	5.8	3.2	-3.4	3.9	0.0	6.5	-6.7	-1.4	-5.7	0.4	-0.1	0.5	0.1	1.2	0.8	0.6	2.6	1.3	-1.4	1.4	0.0	2.1	-1.7	-0.5	-1.8
T8	0.6	0.1	0.3	0.7	2.2	1.7	-0.4	1.9	5.7	1.5	2.8	0.0	3.2	-6.7	0.0	-7.5	0.3	0.0	0.5	0.4	1.0	0.8	-0.1	0.9	2.3	0.7	0.8	0.0	1.1	-1.7	0.0	-2.3
T9	0.6	-0.3	0.8	0.6	2.2	0.7	-1.2	3.9	4.5	-1.0	3.3	0.0	3.2	-6.7	0.0	-9.4	0.3	-0.1	0.7	0.3	1.0	0.3	-0.4	1.8	1.8	-0.4	1.2	0.0	1.1	-1.7	-0.1	-2.9
T10	1.0	-0.1	0.5	1.1	4.0	2.1	0.4	3.9	3.8	-1.5	3.3	0.0	9.7	0.0	1.4	1.9	0.5	-0.1	0.5	0.5	1.8	1.0	0.3	1.8	1.6	-0.6	1.2	0.0	3.1	0.0	0.5	0.6
T11	0.7	-0.1	0.5	1.1	2.2	-0.2	-1.2	2.3	4.5	-1.0	8.3	0.0	3.2	0.0	0.0	-5.7	0.3	-0.1	0.5	0.5	1.0	-0.1	-0.4	1.1	1.8	-0.4	2.9	0.0	1.1	0.0	0.0	-1.8
T12	0.7	-0.2	0.6	0.1	3.6	0.5	-1.2	2.3	4.5	0.5	7.7	0.0	19.4	0.0	-1.4	-7.5	0.3	-0.1	0.5	0.1	1.6	0.2	-0.3	1.0	1.9	0.2	2.8	0.0	6.3	0.0	-0.5	-2.4
T13	0.8	-0.2	0.4	0.7	2.7	1.7	1.6	3.6	3.2	-2.0	4.4	0.0	9.7	20.0	-1.4	0.0	0.4	0.0	0.4	0.4	1.2	0.7	0.8	1.6	1.3	-0.8	1.6	0.0	3.2	4.8	-0.5	0.0
T14	0.6	-0.4	0.2	0.5	3.1	1.7	-2.4	3.9	3.2	0.5	5.0	0.0	16.1	0.0	0.0	0.0	0.3	-0.2	0.3	0.2	1.4	0.6	-0.9	1.7	1.4	0.2	1.8	0.0	5.3	0.0	0.0	0.0
T15	0.7	-0.1	0.9	0.4	1.3	-0.9	-2.0	2.6	1.9	2.0	4.4	0.0	16.1	-6.7	0.0	11.3	0.4	0.0	0.7	0.2	0.6	-0.4	-0.6	1.2	0.8	0.9	1.6	0.0	5.2	-1.7	0.0	3.3
M0	*0.86*	*0.98*	*0.94*	*0.91*	*0.80*	*0.79*	*0.66*	*0.78*	*0.67*	*0.74*	*0.54*	*0.95*	*0.48*	*0.33*	*0.55*	*0.45*	*0.92*	*0.99*	*0.95*	*0.95*	*0.90*	*0.89*	*0.82*	*0.88*	*0.83*	*0.86*	*0.76*	*0.97*	*0.74*	*0.67*	*0.77*	*0.73*

Table 20: 与表 3 以及表 10 、∼ 、19 相似，此表展示了 GA 策略和目标指标 ϕ^{of1} （总体 F1 分数）对票务数据集中类别 $T1, T2, \ldots, T15$ 的影响。

	Optimization: GA-OF1-Class[row]; Performance Metric: ΔOF1%, ΔCR %																Optimization: GA-OF1-Class[row]; Performance Metric: ΔOBA%, ΔCBA %														
	ΔOF1	ΔCR %															ΔOBA	ΔCBA %													
Class	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	%	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	0.1	-0.1	-0.6	1.3	0.9	-0.5	-1.6	0.3	0.6	2.4	7.7	0.0	6.5	-20.0	-4.1	-3.8	0.1	-0.3	-0.1	0.6	0.4	-0.1	-0.6	0.1	0.3	1.1	2.6	0.0	2.1	-5.0	-1.4	-1.2
T2	0.2	-0.3	0.5	0.1	1.8	0.2	-2.4	2.9	-2.5	1.5	0.0	0.0	9.7	-33.3	1.4	-1.9	0.1	-0.1	-0.3	0.1	0.8	0.1	-0.8	1.3	-0.9	0.7	0.1	0.0	3.2	-8.3	0.5	-0.6
T3	0.6	-0.1	-0.2	1.5	2.2	2.4	4.8	1.0	0.6	-2.0	8.8	0.0	6.5	0.0	-8.1	-9.4	0.3	0.0	0.0	0.5	1.0	1.1	2.0	0.5	0.3	-0.7	3.2	0.0	2.1	0.0	-2.8	-2.9
T4	0.8	-0.1	0.1	0.8	4.4	0.9	2.8	2.3	3.2	1.5	6.1	0.0	0.0	6.7	-1.4	-5.7	0.4	0.0	0.3	0.4	1.8	0.5	1.2	1.1	1.4	0.7	2.2	0.0	0.0	1.7	-0.5	-1.8
T5	0.2	-0.1	0.2	0.8	-1.3	-1.2	1.6	1.6	-5.7	3.4	3.9	0.0	6.5	-13.3	-4.1	-5.7	0.1	0.0	0.4	0.4	-0.6	-1.1	0.7	0.9	-2.2	1.5	1.4	0.0	2.1	-3.3	-1.4	-1.7
T6	0.1	0.0	0.1	0.9	0.9	1.9	3.2	2.3	-4.5	-3.9	2.8	-1.1	3.2	-13.3	-4.1	-20.8	0.1	0.0	0.4	0.5	0.4	0.8	0.6	1.1	-1.8	-1.6	1.1	-0.5	1.1	-3.3	-1.4	-6.5
T7	0.2	0.1	0.5	0.9	3.6	1.2	2.8	-1.6	-6.4	0.5	0.0	0.0	-6.5	-13.3	-2.7	-7.5	0.1	0.0	0.5	0.5	1.6	0.6	1.2	-0.9	-2.5	0.3	0.0	0.0	-2.1	-3.3	-0.9	-2.4
T8	0.4	0.0	0.3	0.7	1.8	1.2	-3.2	2.6	0.6	1.5	3.9	0.0	3.2	0.0	-4.1	-3.8	0.2	0.0	0.4	0.3	0.8	0.5	-1.2	1.2	-0.1	0.7	1.4	0.0	1.0	-1.5	-1.5	-1.2
T9	0.6	-0.2	0.7	0.4	1.8	3.1	0.4	2.3	2.5	-2.9	-2.2	0.0	12.9	0.0	4.1	-3.8	0.3	-0.1	0.4	0.2	0.8	1.4	0.3	1.1	1.1	-1.3	-0.7	0.0	4.2	0.0	1.4	-1.2
T10	0.4	-0.2	0.0	-0.1	0.0	2.4	2.8	2.9	-3.8	2.9	4.4	0.0	6.5	0.0	-1.4	-5.7	0.2	-0.1	0.2	0.0	0.0	1.0	1.3	1.3	-1.4	1.3	1.1	0.0	2.1	0.0	-0.4	-1.7
T11	0.8	-0.2	0.7	0.7	2.7	2.4	-0.4	3.9	3.2	-1.5	3.9	0.0	3.2	-6.7	1.4	-5.7	0.4	-0.1	0.5	0.4	1.2	1.1	-0.1	1.8	1.3	-0.5	1.4	0.0	1.1	-1.7	0.5	-1.8
T12	0.8	-0.1	0.3	0.8	1.3	-0.2	1.2	3.2	5.1	-1.0	7.2	0.0	22.6	-6.7	1.4	0.0	0.4	0.0	0.4	0.4	0.6	-0.1	0.6	1.5	2.1	-0.3	2.6	0.0	7.3	-1.6	0.4	0.0
T13	0.7	-0.2	0.7	0.2	2.7	0.9	2.4	3.9	3.2	0.5	2.8	0.0	9.7	13.3	-9.5	0.0	0.4	0.0	0.6	0.1	1.1	0.5	1.1	1.7	1.3	0.2	1.0	0.0	3.2	3.2	-3.3	0.0
T14	0.3	0.0	0.6	-0.2	-0.4	0.0	0.4	1.0	-1.9	0.5	1.7	0.0	0.0	-6.7	10.8	-3.8	0.1	0.0	0.5	-0.1	-0.2	0.1	0.2	0.5	-0.7	0.3	0.7	0.0	0.0	-1.7	3.3	-1.2
T15	0.0	-0.1	-0.4	0.1	-1.3	1.4	-3.6	1.9	-1.9	0.5	4.4	0.0	3.2	-13.3	-4.1	15.1	0.0	0.0	0.0	0.0	-0.6	0.6	-1.4	0.9	-0.7	0.3	1.6	0.0	1.1	-3.3	-1.4	4.3
M0	*0.86*	*0.98*	*0.94*	*0.91*	*0.80*	*0.79*	*0.66*	*0.78*	*0.67*	*0.74*	*0.54*	*0.95*	*0.48*	*0.33*	*0.55*	*0.45*	*0.92*	*0.99*	*0.95*	*0.95*	*0.90*	*0.89*	*0.82*	*0.88*	*0.83*	*0.86*	*0.76*	*0.97*	*0.74*	*0.67*	*0.77*	*0.73*