SAMPLE DATA PROCESSING METHOD AND SYSTEM, AND STORAGE MEDIUM

说　明　书

样本数据的处理方法、系统及存储介质

技术领域

本说明书涉及人工智能技术领域，尤其涉及一种样本数据的处理方法、系统及存储介质。

背景技术

随着人工智能技术的快速发展，神经网络模型在多个领域（例如图像识别、自然语言处理、语音识别等）有着广泛的应用。神经网络模型在应用前，需要通过大量的样本数据进行训练。样本数据的质量和数量可以直接影响训练后的神经网络模型的性能。因此如何得到质量和数量兼备的样本数据是目前亟需解决的问题。

背景技术部分的内容仅仅是发明人个人所知晓的信息，并不代表上述信息在本公开申请日之前已经进入公共领域，也不代表其可以成为本公开的现有技术。

发明内容

本说明书提供一种样本数据的处理方法、系统及存储介质，可以在增加样本数据的数量和多样化的同时，减少额外生成的样本数据中特征值的随机性，尽量保留原始样本数据的分布特征，以此提高被训练模型的性能、以及在新样本数据上的泛化能力。

第一方面，本说明书提供一种样本数据的处理方法，包括：获得 N 个原始样本数据，每个原始样本数据包括多个特征的特征值，所述 N 为大于 1 的整数；基于所述 N 个原始样本数据生成 K 个增强样本数据，其中，每个增强样本数据中的所述多个特征的特征值来自于所述 N 个原始样本数据的至少两个原始样本数据，所述 K 为大于或等

于 1 的整数；进而基于所述 N 个原始样本数据和所述 K 个增强样本数据对目标模型进行训练。

第二方面，本说明书还提供一种样本数据的处理系统，包括：至少一个存储介质，存储有至少一个指令集，用于处理样本数据；以及至少一个处理器，同所述至少一个存储介质通信连接，其中，所述至少一个处理器运行时读取所述至少一个指令集，并且根据所述至少一个指令集的指示：获得 N 个原始样本数据，每个原始样本数据包括多个特征的特征值，所述 N 为大于 1 的整数，基于所述 N 个原始样本数据生成 K 个增强样本数据，其中，每个增强样本数据中的所述多个特征的特征值来自于所述 N 个原始样本数据的至少两个原始样本数据，所述 K 为大于或等于 1 的整数，以及，基于所述 N 个原始样本数据和所述 K 个增强样本数据对目标模型进行训练 。

第三方面，本说明书提供一种计算机可读非暂时性存储介质，其中，所述计算机可读非暂时性存储介质中存储有至少一个指令集，所述至少一个指令集被至少一个处理器执行时，实现如第一方面所述的样本数据的处理方法。

本说明书提供的样本数据的处理方法、系统及存储介质的其他功能将在以下说明中部分列出。本说明书提供的样本数据的处理方法、系统及存储介质的创造性方面可以通过实践或使用下面详细示例中所述的方法、装置和组合得到充分解释。

附图说明

为了更清楚地说明本说明书实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本说明书的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 示出了根据本说明书的实施例提供的一种处理样本数据的应用场景示意图；

图 2 示出了根据本说明书的实施例提供的一种数据处理系统的硬件结构图；

图 3 示出了根据本说明书的实施例提供的样本数据的一种处理方法的流程图；

图 4 示出了根据本说明书的实施例提供的一种生成增强样本数据的方式示意图；

图 5 示出了根据本说明书的实施例提供的另一种生成增强样本数据的方式示意图；

图 6 示出了根据本说明书的实施例提供的样本数据的另一种处理方法的流程图；

图 7 示出了根据本说明书的实施例提供的图 6 所示方法中 S220 的执行步骤示意图；以及

图 8 示出了根据本说明书的实施例提供的数据预处理过程的示意图。

具体实施方式

以下描述提供了本说明书的特定应用场景和要求，目的是使本领域技术人员能够制造和使用本说明书中的内容。对于本领域技术人员来说，对所公开的实施例的各种局部修改是显而易见的，并且在不脱离本说明书的精神和范围的情况下，可以将这里定义的一般原理应用于其他实施例和应用。因此，本说明书不限于所示的实施例，而是与权利要求一致的最宽范围。

这里使用的术语仅用于描述特定示例实施例的目的，而不是限制性的。比如，除非上下文另有明确说明，这里所使用的，单数形式“一”，“一个”和“该”也可以包括复数形式。当在本说明书中使用时，术语“包括”、“包含”和/或“含有”意思是指所关联的整数，步骤、操作、元素和/或组件存在，但不排除一个或多个其他特征、整数、步骤、操作、元素、组件和/或组的存在或在该系统/方法中可以添加其他特征、整数、步骤、操作、元素、组件和/或组。

考虑到以下描述，本说明书的这些特征和其他特征、以及结构的相关元件的操作和功能、以及部件的组合和制造的经济性可以得到明显提高。参考附图，所有这些形成本说明书的一部分。然而，应该清楚地理解，附图仅用于说明和描述的目的，并不旨在限制本说明书的范围。还应理解，附图未按比例绘制。

本说明书中使用的流程图示出了根据本说明书中的一些实施例的系统实现的操作。应该清楚地理解，流程图的操作可以不按顺序实现。相反，操作可以以反转顺序或同时实现。此外，可以向流程图添加一个或多个其他操作。可以从流程图中移除一个或多个操作。

目前，机器学习模型在很多场景中得到广泛应用。在机器学习模型被部署到具体应用场景之前，需要通过大量的样本数据进行训练。在本说明书提供的技术方案中，在模型训练之前，数据处理系统可以对样本数据进行处理，使得处理后的样本数据更加适合训练机器学习模型，以便提高训练后的机器学习模型的性能。

需要说明的是，本说明书对于机器学习模型的类型不做限定。在一些实施例中，上述机器学习模型可以为神经网络模型。神经网络模型是一种受人脑结构和功能启发而来的计算模型，它们通过模拟神经元之间的连接来处理信息。在数学和计算机科学领域，特别是在机器学习和人工智能中，神经网络模型被设计用来识别模式、分类数据、预测结果等。神经网络可以根据设计结构的不同分为多种类型，例如前馈神经网络、卷积神经网络、循环神经网络等。每种类型适用于不同的任务和数据类型。神经网络的能力在于它们可以通过从大量数据中学习，来识别复杂的非线性关系，是当今许多高级机器学习任务的基石。在一些实施例中，机器学习模型还可以为其他模型。后文中以神经网络模型为例进行说明。

然而，在一些应用场景中，能够采集到的样本数据的数量有限。比如，当需要针对一个新的服务平台训练神经网络模型时，由于该服务平台的运行时间比较短，积累的历史服务数据有限，因此能够从历史服务数据中采集的样本数据的数量比较少。采用数量较少的样本数据进行模型训练，可能会导致训练得到的神经网络模型的鲁棒性欠缺，或者存在过拟合问题。

例如，每个样本数据包括多个特征的特征值。特征值可以理解为特征的取值。例如，若特征为交易金额，则特征值为交易的具体数额以及币种（例如 100 元）；若特征

为用户标识，则特征值为用户的具体 ID、名称等；若特征为交易日期，则特征值为具体的年月日（例如 2024 年 1 月 1 日）。

在样本数据的数量较少的情况下，神经网络模型可能会过度学习样本数据中的异常特征值（噪声），而不是重点学习样本数据中正常特征值之间的潜在规律。从而导致神经网络模型鲁棒性不好、有过拟合问题，进而泛化能力较差。

为了克服上述问题，数据处理系统可以采用数据增强的方案。数据增强也可以称为数据扩展。即数据处理系统可以在原始样本数据的基础上，生成更多的样本数据。为了与原始样本数据进行区分，本说明书中将生成的样本数据称为增强样本数据。上述增强样本数据和原始样本数据共同用于模型训练，使得用以模型训练的样本数据的数量达到要求。数据增强可以理解为：对原始样本数据应用一系列变换来生成更多的样本数据，以提高样本数据的数量和多样性。针对不同类型的样本数据，可以应用的变换方式有所不同。以图像数据为例，所述变换包括但不限于：旋转、翻转、缩放、平移、剪切、添加随机值、变换颜色、调整对比度或改变亮度等。所述变换的具体形式可以根据原始样本数据包含的特征值的特点而定。

原始样本数据可以理解为：在数据收集阶段直接从源头获取的样本数据。例如数据处理系统可以从平台数据库得到多个原始样本数据。每个原始样本数据包括的特征值来自所述平台数据库。增强样本数据可以理解为：在数据增强过程中，基于原始样本数据额外生成的、与原始样本数据不完全相同的样本数据。

数据处理系统获得增强样本数据的方式有多种，下面以其中几种示例说明。

增强方式 1

数据处理系统生成任一原始样本数据的副本，随机选择所述原始样本数据中的一个或多个特征，将副本中的所述一个或多个特征的特征值替换为随机值。数据处理系统将被替换后的所述副本作为增强样本数据。

在增强方式1中，虽然可以增加样本数据的数量，但是随机值的取值不受限制，可能会导致在数据增强过程中引入异常特征值。例如，原始样本数据#1 包括第一特征，第一特征的取值范围为[x_1，x_2]。数据处理系统基于原始样本数据#1 生成的增强样本数据#1 中的第一特征为随机值。所述随机值在[x_1，x_2]范围之外。这就导致增强样本数据#1 中包含了不在第一特征的正常取值范围内的特征值。采用这样的增强样本数据训练，可能导致神经网络模型训练不收敛，或者有过拟合等性能问题。

增强方式 2

数据处理系统生成多个原始样本数据的副本，并从原始样本数据中包括的多个特征中确定目标特征。数据处理系统为多个副本中对应目标特征的多个特征值各自加入不同的随机值。数据处理系统将加了随机值的副本作为增强样本数据。其中，被加入的所述不同的随机值呈正态分布。

例如，目标特征为第一特征。数据处理系统生成原始样本数据#1、原始样本数据#2 和原始样本数据#3 的副本。数据处理系统将原始样本数据#1 的副本中的第一特征的特征值加随机值#1，对其他特征的特征值不做处理，将加了随机值#1 的副本作为增强样本数据#1。以此类推，数据处理系统将原始样本数据#2 的副本中的第一特征的特征值加随机值#2，得到增强样本数据#2。数据处理系统将原始样本数据#3 的副本中的第一特征的特征值加随机值#3，得到增强样本数据#3。其中，随机值#1、随机值#2 和随机值#3 呈正态分布。

在增强方式2中，同样可以增加样本数据的数量。但是，数据处理系统在生成增强样本数据的过程中，引入了特定分布的随机值。这样的方式可能掩盖或者扭曲了原始样本数据中特征值的真实分布，导致神经网络模型在训练过程中学习不到特征值间的真实规律，影响神经网络模型的性能。

鉴于此，本说明书的实施例提供一种样本数据的处理方法、设备及存储介质，可以在增加样本数据的数量和多样化的同时，在一定程度上减少增强样本数据中特征值

的随机性，保留原始样本数据的分布特征，以此提高被训练模型的性能、以及在新样本数据上的泛化能力。

通常情况下，采用样本数据对神经网络模型训练之前，还需要对样本数据进行数据预处理。即，数据处理系统通过一系列技术和步骤的结合，将样本数据中的特征值处理为适合神经网络模型分析的数据，或者处理为适合喂入神经网络模型的范式化的数据。例如，针对样本数据中不同特征的特征值的量纲/尺度不同的情况，数据处理系统可以将各特征的特征值处理到位于相同数值区间内，使得不同特征的特征值之间具有可比性。

对于原始样本数据的数量不足的场景：数据处理系统在对原始样本数据做数据增强后、训练神经网络模型之前，还需要对原始样本数据以及增强样本数据做数据预处理。对于原始样本数据的数量足够的场景：数据处理系统在训练神经网络模型之前，可以不进行数据增强，但需要对原始样本数据进行数据预处理。预处理后的样本数据可以用于训练神经网络模型。

数据处理系统对样本数据（例如原始样本数据，或者，原始样本数据和增强样本数据）进行数据预处理的方式有多种，下面以其中几种示例说明。

预处理方式 1：标准化

数据处理系统从样本数据的多个特征中确定目标特征，并确定所有样本数据中对应所述目标特征的特征值的平均值和标准差。进而，数据处理系统基于所述平均值和标准差将每个样本数据中对应所述目标特征的特征值处理为更新值。从而数据处理系统得到预处理后的样本数据。

例如，对于样本数据#1，预处理后的样本数据#1 中对应所述目标特征的更新值基于以下公式得到：

$$X = \frac{x_i - \mu}{\sigma}$$

其中，X 为所述更新值，x_i 为预处理前的样本数据#1 中对应所述目标特征的特征

值，μ为所述平均值，σ为所述标准差。

预处理方式2：归一化

数据处理系统从样本数据的多个特征中确定目标特征，并确定所有样本数据中对应所述目标特征的特征值中的最大特征值。进而，数据处理系统基于所述最大特征值将每个样本数据中对应所述目标特征的特征值处理为更新值，从而数据处理系统得到预处理后的样本数据。更新值位于区间#1内。区间#1比如[0, 1]。

例如，对于样本数据#1，预处理后的样本数据#1 中对应所述目标特征的更新值基于以下公式得到：

$$X = \frac{x_i}{x_{max}}$$

其中，X为所述更新值，x_i为预处理前的样本数据#1 中对应所述目标特征的特征值，x_{max}为所述最大特征值。

上述两种预处理方式，对于样本数据中的异常特征值比较敏感。比如异常特征值不在目标特征的正常取值范围内，存在过大或者过小的问题，导致平均值或者标准差受到异常特征值的影响。因此数据处理系统基于平均值和标准差得到的更新值也会受到异常特征值的影响。或者，异常特征值过大，导致最大特征值本身是异常特征值，因此数据处理系统基于最大特征值得到的更新值同样会受到异常特征值的影响。从而采用通过这两种预处理方式得到的处理后的样本数据训练神经网络模型，会影响神经网络模型的性能。

鉴于此，本说明书的实施例提供一种样本数据的处理方法、设备及存储介质，可以基于样本数据中与待预处理特征对应的各特征值的排序索引，将与该待预处理特征对应的各特征值更新为目标区间内的值。这样的处理方式，不直接基于特征值，而是基于特征值的排序索引进行预处理，可以减少异常特征值、以及不同特征间的尺度不同的影响，增强了样本数据的鲁棒性，更适用于模型训练，从而提升被训练模型的性能。

下面结合图 1 对本说明书提供的处理样本数据的应用场景进行介绍。

图 1 示出了根据本说明书的实施例提供的一种处理样本数据的应用场景示意图。如图 1 所示，场景 001 可以包括：数据源 100 和样本数据的处理系统 200（后文简称为数据处理系统 200）。

数据源 100 用于存储各种类型的数据。在一些实施例中，数据源 100 可以是提供各种服务的服务器。例如对用户提供支付服务或者交易服务等的后台服务器。数据源 100 可以存储数据源 100 为用户提供服务的过程中产生的服务数据。在一些实施例中，数据源 100 可以为服务平台的数据库，存储服务平台产生的服务数据。

数据处理系统 200 可以是具有一定计算能力的计算系统。在一些实施例中，数据处理系统 200 可以执行本说明书描述的样本数据的处理方法。数据处理系统 200 可以存储有用于实现样本数据的处理方法的数据和指令，并可以执行或用于执行所述数据和指令。在一些实施例中，数据处理系统 200 可以包括具有数据信息处理功能的硬件设备和驱动该硬件设备工作所需必要的程序。

示例性的，数据处理系统 200 可以从数据源 100 获得相关数据。可以理解，数据处理系统 200 获取的数据类型依据学习任务确定。比如学习任务#1 是基于交易信息识别当前交易是否是风险交易。则数据处理系统 200 从数据源 100 获得多个交易的交易信息。比如学习任务#2 是基于商户信息识别当前商户是否是风险商户。则数据处理系统 200 从数据源 100 获得多个商户的商户信息。

示例性的，数据处理系统 200 可以基于所述相关数据生成多个原始样本数据。例如，数据处理系统 200 可以先基于学习任务的需求确定出多个特征，再基于所述多个特征从所述相关数据中生成多个原始样本数据。每个原始样本数据包括所述多个特征各自对应的特征值。

数据处理系统 200 可以适用于原始样本数据的数量不足的场景 1，也可以适用于原始样本数据的数量充足的场景 2。如图 1 所示，数据处理系统 200 在场景 1 和场景 2 下

执行的步骤有所不同。在图 1 中，相较于场景 2，数据处理系统 200 在场景 1 下需要额外执行虚线框对应的执行步骤（例如生成增强样本数据，对增强样本数据进行数据预处理）。

示例性的，对于场景 1：数据处理系统 200 得到多个原始样本数据后，可以基于所述多个原始样本数据中的部分或者全部原始样本数据生成多个增强样本数据。进而，数据处理系统 200 可以基于所述多个原始样本数据和所述多个增强样本数据对目标模型进行训练。目标模型可以是神经网络模型，本说明书实施例对此不做限制。增强样本数据的生成方式通过后文的方法 P100 详细介绍，在此不再赘述。

在一些实施例中，数据处理系统 200 对目标模型训练之前，还可以对样本数据做数据预处理操作。

示例性的，对于场景 2：数据处理系统 200 得到多个原始样本数据后，可以对所述多个原始样本数据进行预处理。进而数据处理系统 200 基于预处理后的所述多个原始样本数据训练目标模型。具体的预处理操作通过后文的方法 P200 详细介绍，在此不再赘述。

示例性的，对于场景 1：数据处理系统 200 基于所述多个原始样本数据生成多个增强样本数据。而后，数据处理系统 200 可以对所述多个原始样本数据和所述多个增强样本数据进行预处理，进而基于预处理后的所述多个原始样本数据和所述多个增强样本数据训练目标模型。具体的预处理操作通过后文的方法 P200 详细介绍，在此不再赘述。

还可以理解，图 1 中的数据源 100 和数据处理系统 200 的数目仅仅是示意性的。根据实现需要，可以具有任意数目的数据源 100 和数据处理系统 200。

需要说明的是，本说明书提供的对样本数据的处理方法包括：数据增强方法（方法 P100）和/或数据预处理方法（方法 P200）。所述样本数据的处理方法可以完全在数据源 100 上执行，可以完全在数据处理系统 200 上执行，也可以部分在数据源 100 上执行、部分在数据处理系统 200 上执行。或者还可以有其他符合逻辑的执行方式，不再一

一列举。

在一些实施例中，数据处理系统200可以对应于单个电子设备。图1中示出的原始样本数据的生成过程、增强样本数据的生成过程、数据预处理过程、模型训练过程可以均由上述单个电子设备执行。

在一些实施例中，数据处理系统200可以对应于由多个电子设备组成的设备集群。图1中示出的原始样本数据的生成过程、增强样本数据的生成过程、数据预处理过程、模型训练过程可以由上述多个电子设备联合执行或者分开执行。例如，电子设备1执行上述原始样本数据和增强样本数据的生成过程，电子设备2执行上述数据预处理过程，电子设备3执行上述模型训练过程；或者，电子设备1执行上述原始样本数据的生成过程、增强样本数据的生成过程和数据预处理过程，电子设备3执行上述模型训练过程；或者上述多个电子设备还可以有其他的执行方式，不再一一列举。

图2示出了根据本说明书的实施例提供的一种数据处理系统200的硬件结构图。

如图2所示，数据处理系统200可以包括至少一个存储介质230和至少一个处理器220。在一些实施例中，数据处理系统200还可以包括通信端口250和内部通信总线210。此外，数据处理系统200还可以包括I/O组件260。

内部通信总线210可以连接不同的系统组件。例如，内部通信总线210可以连接存储介质230、处理器220、通信端口250和I/O组件260。

I/O组件260支持数据处理系统200和其他组件之间的输入/输出。

通信端口250用于数据处理系统200同外界的数据通信。比如，通信端口250可以用于数据处理系统200同网络400之间的数据通信。通信端口250可以是有线通信端口也可以是无线通信端口。

网络400用于在数据源100和数据处理系统200之间提供通信连接的介质。网络400可以促进信息或数据的交换。在一些实施例中，网络400可以是任何类型的有线或无线网络，也可以是其组合。比如，网络400可以包括电缆网络，有线网络、光纤网

络、电信通信网络、内联网、互联网、局域网（LAN）、广域网（WAN）、无线局域网（WLAN）、大都市市区网（MAN）、公用电话交换网（PSTN）、蓝牙网络、ZigBee 网络、近场通信（NFC）网络或类似网络。在一些实施例中，网络 400 可以包括一个或多个网络接入点。例如，网络 400 可以包括有线或无线网络接入点，如基站或互联网交换点，通过该接入点，数据源 100 和数据处理系统 200 的一个或多个组件可以连接到网络 400 以交换数据或信息。

存储介质 230 可以包括数据存储装置。所述数据存储装置可以是非暂时性存储介质，也可以是暂时性存储介质。比如，所述数据存储装置可以包括磁盘 232、只读存储介质（ROM）234 或随机存取存储介质（RAM）236 中的一种或多种。存储介质 230 还包括存储在所述数据存储装置中的至少一个指令集。所述指令集可以包括计算机程序代码，所述计算机程序代码可以包括执行本说明书提供的样本数据的处理方法的程序、例程、对象、组件、数据结构、过程、模块等等。

处理器 220 可以同存储介质 230 通信连接。处理器 220 用以执行上述至少一个指令集。当数据处理系统 200 运行时，处理器 220 读取所述至少一个指令集，并且根据所述至少一个指令集的指示，执行本说明书提供的样本数据的处理方法。

处理器 220 可以是一个或多个处理器的形式，在一些实施例中，处理器 220 可以包括一个或多个硬件处理器，例如微控制器，微处理器，精简指令集计算机（RISC），专用集成电路（ASIC），特定于应用的指令集处理器（ASIP），中心处理单元（CPU），图形处理单元（GPU），物理处理单元（PPU），微控制器单元，数字信号处理器（DSP），现场可编程门阵列（FPGA），高级 RISC 机器（ARM），可编程逻辑器件（PLD），能够执行一个或多个功能的任何电路或处理器等，或其任何组合。

仅仅为了说明问题，在附图中数据处理系统 200 中仅示出一个处理器 220。然而，应当注意，本说明书中数据处理系统 200 还可以包括多个处理器，因此，本说明书中披露的操作和/或方法步骤可以由一个处理器执行，也可以由多个处理器联合执行。例如，

如果在本说明书中描述数据处理系统200的处理器220执行步骤A和步骤B，则应该理解，步骤A和步骤B也可以由两个不同处理器220联合或分开执行（例如，第一处理器执行步骤A，第二处理器执行步骤B，或者第一处理器和第二处理器共同执行步骤A和B）。

图3示出了根据本说明书的实施例提供的一种样本数据的处理方法P100的流程图。如前，数据处理系统200可以执行样本数据的处理方法P100。具体地，数据处理系统200中的处理器可以读取存储在其本地存储介质中的指令集，然后根据指令集的规定，执行样本数据的处理方法P100。

样本数据的处理方法P100可以对样本数据进行数据增强，可应用于交易领域、网络安全领域、金融领域等需要训练目标模型的领域，本说明书实施例对此不做限制。为描述清楚，后文主要以交易领域为例说明。

如图3所示，样本数据的方法P100可以包括如下步骤。

S110：获得N个原始样本数据，每个原始样本数据包括多个特征的特征值，N为大于1的整数。

数据处理系统200可以基于目标模型的学习任务获得所述N个原始样本数据。所述目标模型为待训练模型。目标模型比如神经网络模型，本说明书实施例不限于此。

例如，数据处理系统200可以基于学习任务从数据源100获得相关数据，并基于所述相关数据生成所述N个原始样本数据。例如，数据处理系统200可以基于学习任务的需求确定出多个特征，再针对每个样本（例如一个交易或者一个商户等）从所述相关数据中获得所述多个特征的特征值，从而生成该样本对应的样本数据。

比如，目标模型的学习任务为基于交易信息识别当前交易是否是风险交易。所述多个特征包括但不限于：用户、商户、交易时间、交易金额、交易物品以及交易平台等。数据处理系统200可以从一个交易的交易信息中提取所述多个特征的特征值，并基于提取出的多个特征值生成一个原始样本数据。以一个原始样本数据为例，原始样本

数据包括用户 a、商户 b、2013 年 12 月 15 日 13:14:52、x_1 交易币（币种）、数码产品以及 xx 平台等。

再比如，目标模型的学习任务为基于商户信息识别当前商户是否是风险商户。所述多个特征包括但不限于：商户名称、注册时间、注册地址、经营范围以及交易平台等。数据处理系统 200 可以从一个商户的商户信息中提取所述多个特征的特征值，并基于提取出的多个特征值生成一个原始样本数据。以一个原始样本数据为例，原始样本数据包括商户 c、2015 年 8 月 6 日、xx 市 xx 区 xx、日用品经营以及 xx 平台等。

S120：基于 N 个原始样本数据生成 K 个增强样本数据，K 为大于或等于 1 的整数。其中，每个增强样本数据中的所述多个特征的特征值来自于 N 个原始样本数据的至少两个原始样本数据。

数据处理系统 200 确定原始样本数据的数量 N 达不到要求（例如模型训练所需的数量要求）的情况下，数据处理系统 200 需要基于 N 个原始样本数据生成 K 个增强样本数据。例如，当原始样本数据的数量 N 大于或等于第一阈值，则数据处理系统 200 确定原始样本数据的数量 N 达到要求。或者，当原始样本数据的数量 N 小于第一阈值，则数据处理系统 200 确定原始样本数据的数量 N 达不到要求。

在一些实施例中，每个原始样本数据对应有标签信息，生成每个增强样本数据的所述至少两个原始样本数据对应的所述标签信息相同。

例如，目标模型用于学习分类。每个原始样本数据可以包含标签信息。比如，目标模型的学习任务为基于交易信息识别交易是否是风险交易，则标签信息包括有风险或者无风险。有风险表示该交易是风险交易。无风险表示该交易不是风险交易。

可以理解，不同增强样本数据可以基于不同原始样本数据生成。或者，不同增强样本数据可以基于不同数量的原始样本数据生成。

例如，数据处理系统 200 基于原始样本数据①和原始样本数据②生成增强样本数据#1，即增强样本数据#1 中各特征的特征值来自于原始样本数据①和原始样本数据②；

数据处理系统200基于原始样本数据①、原始样本数据②和原始样本数据③生成增强样本数据#2，即增强样本数据#2 中各特征的特征值来自于原始样本数据①、原始样本数据②和原始样本数据③；数据处理系统200基于原始样本数据④和原始样本数据⑤生成增强样本数据#3，即增强样本数据#3 中各特征的特征值来自于原始样本数据④和原始样本数据⑤。

其中，原始样本数据①、原始样本数据②和原始样本数据③对应的标签信息相同；原始样本数据④和原始样本数据⑤对应的标签信息相同。若原始样本数据①和原始样本数据②对应的标签信息都是无风险，则增强样本数据#1 对应的标签信息也是无风险。以此类推，增强样本数据#2 对应的标签信息也是无风险。若原始样本数据④和原始样本数据⑤对应的标签信息都是有风险，则增强样本数据#3 对应的标签信息也是有风险。

在该实施例中，数据处理系统200生成一个增强样本数据的过程中，所依据的是相同标签信息的原始样本数据。这样可以保障生成的增强样本数据相对可靠，避免依据不同标签信息的原始样本数据生成增强样本数据后，扰乱样本数据的潜在规律，不利于目标模型的训练。

在一些实施例中，生成每个增强样本数据的所述至少两个原始样本数据包括第一原始样本数据和至少一个第二原始样本数据。其中，每个增强样本数据是通过如下方式生成的：数据处理系统200生成第一原始样本数据的副本，将所述副本中与至少一个目标特征对应的特征值替换为至少一个第二原始样本数据中与所述至少一个目标特征对应的特征值。所述至少一个目标特征为所述多个特征中的部分特征。数据处理系统200将替换特征值后的所述副本作为所述增强样本数据。

下面以增强样本数据 i 的生成过程举例说明，其他增强样本数据的生成过程类似，可参考增强样本数据 i 的相关描述。

增强样本数据 i 中对应所述多个特征的特征值来自于所述 N 个原始样本数据的至少两个原始样本数据。所述至少两个原始样本数据包括第一原始样本数据和至少一个第二原始样本数据。

例如，数据处理系统 200 可以从所述 N 个原始样本数据中确定所述至少两个原始样本数据。其中，所述至少两个原始样本数据的数量可以是随机数量，也可以是预设的固定数量，本说明书实施例对此不做限定。数据处理系统 200 从所述 N 个原始样本数据中确定所述至少两个原始样本数据时，可以采用随机选择的方式，也可以是按照某个顺序选择，本说明书实施例对此不做限定。

数据处理系统200确定所述至少两个原始样本数据之后，将其中一个原始样本数据作为第一原始样本数据，将其他原始样本数据作为第二原始样本数据。而后，数据处理系统200生成第一原始样本数据的副本。所述副本的内容与所述第一原始样本数据相同，即所述副本中所述多个特征的特征值与所述第一原始样本数据中所述多个特征的特征值相同。

可以理解，数据处理系统200得到第一原始样本数据的副本后，可以将副本中的部分特征（即所述至少一个目标特征）的特征值替换为所述第二原始样本数据中对应特征的特征值。其中，上述被替换的特征（即所述至少一个目标特征）的数量可以是随机数量，也可以是预设的固定数量，本说明书实施例对此不做限定。上述被替换的特征是数据处理系统200在所述多个特征中随机选择，也可以是基于所述多个特征的顺序按序选择，还可以是由操作人员预先指定，本说明书实施例对此不做限定。

在一些实施例中，所述至少一个目标特征在所述多个特征中的占比小于或等于预设占比。

在该实施例中，第一原始样本数据的副本被替换的特征值的数量的占比不超过预设占比。从而保障被替换的特征值是少部分，使得副本中的大部分特征值能够保留，尽量避免改变原始样本数据的潜在规律，影响训练得到的目标模型的性能。基于这样

的替换方式，数据处理系统200生成的增强样本数据中的特征值有一定随机性，但随机性是有限度的。因此有助于目标模型学习真实世界数据中可能会出现的小幅度变化，提高目标模型的泛化能力。

示例性的，数据处理系统200基于所述预设占比确定目标特征的数量，基于所述数量从所述多个特征中随机选择所述数量的特征，并将被选择的特征作为所述至少一个目标特征。

例如，所述预设占比可以为小于或等于 10%的值。比如，所述预设占比可以为 5%，6%，7%，8%，9%或者 10%等，本说明书实施例对此不做限制。假设所述原始样本数据包括的所述多个特征中所有特征的数量为m，则数据处理系统200确定所述至少一个目标特征的数量小于或者等于（m×预设占比）。其中，×表示乘法。

数据处理系统200从所述多个特征中随机选择目标特征可以有多种方式，下面以其中几种示例，本说明书实施例对此不做限制。

例如，数据处理系统 200 生成一个随机的掩码，每个码位对应一个特征，码位以True或False呈现。数据处理系统200可以将 True对应的特征选择为所述目标特征。

例如，数据处理系统200生成一个随机的二进制数，每个数位对应一个特征，数位以0或1呈现。数据处理系统200可以将1对应的特征选择为所述目标特征。

由上文描述可知，针对一个增强样本数据，以所述增强样本数据中的各特征的特征值来自于第一原始样本和第二原始样本为例，所述增强样本数据可以视为通过将第一原始样本数据中的部分特征的特征值替换为第二原始样本数据的对应特征的特征值得到的，或者说，所述增强样本数据是通过对第一原始样本数据和第二原始样本数据中的部分特征的特征值进行交换得到的，或者说，所述增强样本数据是通过将第一原始样本数据中的一部分特征的特征值与第二原始样本数据中的另一部分特征的特征值进行拼接/组合得到的。

可以理解，不同的增强样本数据所对应的生成方式（所基于的原始样本的数量、被替换特征的数量、被替换的特征等）可以相同，也可以不同。本说明书实施例对此不做限制。

数据处理系统200确定所述至少一个目标特征后，将第一原始样本数据的副本中对应所述至少一个目标特征的特征值替换掉，将替换后的副本作为增强样本数据 i。其中，替换方式可以有多种，下面以其中几种示例，本说明书实施例对此不做限制。

方式 1

示例性的，若第二原始样本数据的数量为 1，则数据处理系统 200 将所述第一原始样本数据的副本中对应所述至少一个目标特征的特征值都替换为所述第二原始样本数据中对应所述至少一个目标特征的特征值。

例如，图 4 所示，目标特征包括特征 e 和特征 g。所述第一原始样本数据的副本中特征 e 对应的特征值为特征值 e_1，特征 g 对应的特征值为特征值 g_1。所述第二原始样本数据中特征 e 对应的特征值为特征值 e_2，特征 g 对应的特征值为特征值 g_2。数据处理系统 200 将第一原始样本数据的副本中的特征值 e_1 替换为特征值 e_2，将特征值 g_1 替换为特征值 g_2。替换后的所述副本即增强样本数据 i。

方式 2

示例性的，第二原始样本数据的数量大于 1。所述至少一个目标特征包括第一特征和第二特征。第二原始样本数据包括第三原始样本数据和第四原始样本数据。数据处理系统200将所述第一原始样本数据的副本中与所述第一特征对应的特征值替换为所述第三原始样本数据中与所述第一特征对应的特征值，并将所述副本中与所述第二特征对应的特征值替换为所述第四原始样本数据中与所述第二特征对应的特征值。

所述第三原始样本数据可以理解为：向所述第一原始样本数据的副本提供第一特征对应的特征值的替换值的原始样本数据。第三原始样本数据可以是一个原始样本数据，也可以是多个原始样本数据，本说明书实施例对此不做限制。

所述第四原始样本数据可以理解为：向所述第一原始样本数据的副本提供第二特征对应的特征值的替换值的原始样本数据。第四原始样本数据可以是一个原始样本数据，也可以是多个原始样本数据，本说明书实施例对此不做限制。

下面以所述第三原始样本数据是一个原始样本数据，所述第四原始样本数据是一个原始样本数据为例，说明替换过程。

例如，图 5 所示，目标特征包括第一特征（特征 e）和第二特征（特征 g）。所述第一原始样本数据的副本中特征 e 对应的特征值为特征值 e_1，特征 g 对应的特征值为特征值 g_1。所述第三原始样本数据中特征 e 对应的特征值为特征值 e_2。所述第四原始样本数据中特征 g 对应的特征值为特征值 g_3。数据处理系统 200 将副本中特征值 e_1 替换为特征值 e_2，将特征值 g_1 替换为特征值 g_3。替换后的所述副本即增强样本数据 i。

可以理解，若所述至少一个目标特征包括多个第一特征，且第三原始样本数据是多个原始样本数据。则数据处理系统 200 可以将所述副本中与部分第一特征对应的特征值替换为部分第三原始样本数据中与所述部分第一特征对应的特征值，并将所述副本中与另一部分第一特征对应的特征值替换为另一部分第三原始样本数据中与所述另一部分第一特征对应的特征值。

例如，所述至少一个目标特征包括第一特征#1、第一特征#2 和第一特征#3。第三原始样本数据包括原始样本数据#1，原始样本数据#2 和原始样本数据#3。数据处理系统 200 可以将所述副本中与第一特征#1 对应的特征值替换为原始样本数据#1 中与第一特征#1 对应的特征值， 将所述副本中与第一特征#2 对应的特征值替换为原始样本数据#2 中与第一特征#2 对应的特征值，将所述副本中与第一特征#3 对应的特征值替换为原始样本数据#3 中与第一特征#3 对应的特征值。

再例如，所述至少一个目标特征包括第一特征#1、第一特征#2 和第一特征#3。第三原始样本数据包括原始样本数据#1 和原始样本数据#3。数据处理系统 200 可以将所述副本中与第一特征#1 对应的特征值替换为原始样本数据#1 中与第一特征#1 对应的特

征值，将所述副本中与第一特征#2 对应的特征值替换为原始样本数据#1 中与第一特征#2 对应的特征值，将所述副本中与第一特征#3 对应的特征值替换为原始样本数据#3 中与第一特征#3 对应的特征值。

可以理解，当第一特征的数量和第三原始样本数据的数量是其他组合时，数据处理系统 200 可以按照其他的替换方式替换副本中的替换值。不再一一列举。

还可以理解，将副本中对应第二特征的特征值替换为第四原始样本数据中的特征值的方式，与上文示例的将副本中对应第一特征的特征值替换为第三原始样本数据中的特征值的方式类似，不再赘述。

还可以理解，所述方式 2 中，数据处理系统 200 将所述目标特征分了 2 类，分别是第一特征和第二特征。数据处理系统 200 将所述第二原始样本数据也分了 2 类，分别是第三原始样本数据和第四原始样本数据。除此之外，数据处理系统 200 还可以将所述目标特征分为 Q 类，Q 为大于 2 的整数。数据处理系统 200 还可以将所述第二原始样本数据分为 Q 类。在替换过程中，数据处理系统 200 将所述副本中与第一类特征对应的特征值替换为第一类原始样本数据中与所述第一类特征对应的特征值。以此类推，直至数据处理系统 200 将所述副本中所有目标特征的特征值替换掉。除上文示例的替换方式之外，数据处理系统 200 还可以以其他符合逻辑、能够实现的方式替换所述副本中的特征值。不再一一列举。

在上述示例的替换方式中，副本中的特征值被替换为原始样本数据中的特征值。替换过程相对随机，但随机程度有限，尽量避免改变原始样本数据中某特征的特征值的分布，影响训练后的目标模型的性能。

上文是以增强样本数据 i 为例描述其生成方式。按照类似的方式，数据处理系统 200 可以生成所述 K 个增强样本数据。

S130：基于 N 个原始样本数据和 K 个增强样本数据对目标模型进行训练。

示例性的，数据处理系统 200 可以将所述 N 个原始样本数据和 K 个增强样本数据

分为训练集和验证集，并将训练集中的样本数据分为多个 epoch。数据处理系统 200 在每轮迭代训练中使用一个或者多个 epoch。如此重复进行多轮迭代训练，直到数据处理系统 200 基于验证集、目标函数确定训练得到的目标模型达到训练结束条件。

综上所述，本说明书提供的样本数据的处理方法 P100、设备及存储介质，通过数据增强的方式生成多个增强样本数据，使得用以训练目标模型的样本数据的数量能够满足要求，避免目标模型的过拟合问题。并且增强样本数据的特征值均来自于原始样本数据，避免在增强样本数据中引入原始样本数据中没有出现过的特征值，减少随机性。在保障样本数据的多样性的同时，尽量保留原始样本数据中的真实特征值，避免增强样本数据过于失真，从而提升目标模型的性能。

图 6 示出了根据本说明书的实施例提供的一种样本数据的处理方法 P200 的流程图。如前，数据处理系统 200 可以执行样本数据的处理方法 P200。具体地，数据处理系统 200 中的处理器可以读取存储在其本地存储介质中的指令集，然后根据指令集的规定，执行样本数据的处理方法 P200。

如图 6 所示，样本数据的处理方法 P200 可以包括如下步骤。

S210：获得 M 个样本数据，每个样本数据包括多个特征（同方法 P100 中的多个特征）的特征值，M 为大于 1 的整数。

在一些实施例中，数据处理系统 200 获得的所述 M 个样本数据包括所述 N 个原始样本数据和所述 K 个增强样本数据。

在一些实施例中，数据处理系统 200 获得的所述 M 个样本数据包括所述 N 个原始样本数据。

S220：基于所述 M 个样本数据，针对所述多个特征中的至少部分特征分别执行预处理操作，得到预处理后的所述 M 个样本数据。

在一些实施例中，数据处理系统 200 基于所述 N 个原始样本数据，针对所述多个特征中的至少部分特征分别执行预处理操作，得到预处理后的所述 N 个原始样本数据。

在一些实施例中，数据处理系统 200 基于所述 N 个原始样本数据和所述 K 个增强样本数据，针对所述多个特征中的至少部分特征分别执行预处理操作，得到预处理后的所述 N 个原始样本数据和所述 K 个增强样本数据。

下面主要以数据处理系统 200 对所述 N 个原始样本数据和所述 K 个增强样本数据执行预处理操作为例，说明预处理操作的具体步骤。数据处理系统 200 对所述 N 个原始样本数据执行预处理操作的具体内容，可参考下文描述，不再赘述。

可以理解，数据处理系统 200 对所述 N 个原始样本数据和所述 K 个增强样本数据执行预处理操作时，针对的是所述多个特征中的部分特征或者全部特征。对于不同特征，数据处理系统 200 执行的预处理操作类似。下面以所述至少部分特征中的第三特征为例，介绍数据处理系统 200 针对所述第三特征执行预处理操作的具体步骤。如图 7 所示，S220 可以包括如下步骤：S2202、S2204 和 S2206。

S2202：从所述 N 个原始样本数据和所述 K 个增强样本数据中获得与所述第三特征对应的 N+K 个特征值。

例如，图 8 所示，原始样本数据和增强样本数据均包括特征 a、特征 b、特征 c、特征 e、特征 f、特征 g、特征 h 和特征 j 各自对应的特征值。若所述第三特征为特征 e，则数据处理系统 200 从原始样本数据 1 中获取特征 e 对应的特征值 e_1。以此类推，直至数据处理系统 200 将所有原始样本数据和增强样本数据中与特征 e 对应的特征值获取完成。即数据处理系统 200 获得的所述 N+K 个特征值包括特征值 e_1、…、特征值 e_N、…、特征值 e_3、…以及特征值 e_6。

S2204：确定所述 N+K 个特征值对应的排序索引。

在一些实施例中，数据处理系统 200 对所述 N+K 个特征值进行排序，并基于排序后的所述 N+K 个特征值的顺序得到所述 N+K 个特征值对应的排序索引。

例如，数据处理系统 200 按照从大到小的顺序，将所述 N+K 个特征值进行排序，得到所述 N+K 个特征值对应的排序索引。比如，特征值 e_1 最大，则特征值 e_1 排序第一。

数据处理系统 200 确定特征值 e_1 对应的排序索引为 0。特征值 e_N 仅次于特征值 e_1，则特征值 e_N 排序第二。数据处理系统 200 确定特征值 e_N 对应的排序索引为 1。以此类推。

例如，数据处理系统 200 按照从小到大的顺序，将所述 N+K 个特征值进行排序，得到所述 N+K 个特征值对应的排序索引。比如，特征值 e_3 最小，则特征值 e_3 排序第一。数据处理系统 200 确定特征值 e_3 对应的排序索引为 0。特征值 e_6 仅次于特征值 e_3，则特征值 e_6 排序第二。数据处理系统 200 确定特征值 e_6 对应的排序索引为 1。以此类推。

可以理解，若所述 N+K 个特征值中的部分特征值相同，则所述部分特征值中每个特征值对应的排序索引可以相同，也可以不同。不同的情况：数据处理系统 200 可以针对所述部分特征值随机生成不同的、连续的排序索引。或者，数据处理系统 200 还可以有其他的确定排序索引的方式，本说明书实施例对此不做限制。

如图 8 所示，按照上述排序方式，数据处理系统 200 确定 N+K 个排序索引。其中特征值 e_1 对应的排序索引为 $I(e_1)$，…，特征值 e_N 对应的排序索引为 $I(e_N)$，特征值 e_3 对应的排序索引为 $I(e_3)$，…，特征值 e_6 对应的排序索引为 $I(e_6)$。

S2206：基于所述 N+K 个特征值对应的排序索引确定所述 N+K 个特征值对应的更新值，所述 N+K 个特征值对应的更新值均位于目标区间内。

更新值可以理解为对特征值进行预处理操作后被更新的值。比如，特征值 e_6 在预处理操作后被更新为特征值 e_6 对应的更新值。目标区间为预处理操作后得到的 N+K 个更新值需要满足的区间范围。针对不同的目标模型，目标区间可以不同。在一些实施例中，目标区间可以为[-1, 1]。在一些示例中，目标区间可以为[0，1]。在一些实施例中，目标区间可以为[-2，2]。

数据处理系统 200 基于特征值的排序索引对样本数据做预处理的这种方案，可以避免异常特征值对预处理操作的影响。比如异常特征值为超过第三特征的特征值的取值范围的超大值。如果通过标准化或者归一化等预处理操作，则该异常特征值被预处理后，与其他特征值对应的更新值的差距依然比较大。异常特征值对应的更新值形成目

标模型训练过程中的噪音，不利于目标模型的训练。但是基于排序索引进行预处理，可以使得该异常特征值被预处理后，与其他特征值对应的更新值之间的差距变小。从而可以提高被处理后的样本数据的鲁棒性，提升目标模型的训练效果。

在一些实施例中，针对所述 N+K 个特征值中的每个特征值，数据处理系统 200 基于所述特征值对应的排序索引与最大索引确定所述特征值对应的更新值，所述最大索引为所述 N+K 个特征值对应的排序索引中的最大值。

例如，以所述 N+K 个特征值中的特征值 i 为例，数据处理系统 200 基于以下公式得到所述特征值 i 对应的更新值。

$$X=(\frac{x_i}{x_{max}}-0.5)\times 2$$

其中，X表示所述特征值 i 对应的更新值，x_i表示所述特征值 i 对应的排序索引，x_{max}表示所述最大索引。

数据处理系统 200 基于上述公式对所述 N+K 个特征值中的每个特征值分别进行计算，得到所述 N+K 个特征值对应的 N+K 个更新值。

除了上述公式，数据处理系统 200 还可以通过其他方式计算更新值，比如$X=\frac{x_i}{x_{max}}$。本说明书实施例对此不做限制。

如图 8 所示，数据处理系统 200 按照上述方式得到 N+K 个更新值。其中特征值 e_1 对应的更新值为更新值 e_1'，…，特征值 e_N 对应的更新值为更新值 e_N'，特征值 e_3 对应的更新值为更新值 e_3'，…，特征值 e_6 对应的更新值为更新值 e_6'。

在一些实施例中，在数据处理系统 200 确定所述 N+K 个特征值对应的更新值之后，S220 还包括：

数据处理系统 200 通过预设的分布调整函数对所述 N+K 个特征值对应的更新值进行处理，使得处理后的所述 N+K 个特征值对应的更新值呈目标分布。

所述分布调整函数可以理解为：可以对第三特征的所述 N+K 个特征值的分布进行调整的函数。所述目标分布可以基于目标模型确定。不同目标模型对应的目标分布不

同。比如有些目标模型（例如线性回归模型、逻辑回归模型或神经网络模型）在训练过程中的输入数据需要满足正态分布（或者高斯分布），则所述分布调整函数需将所述 N+K 个特征值对应的更新值处理为呈正态分布的数据。

在该实施例中，数据处理系统 200 采用了分布调整函数，将所述 N+K 个特征值对应的更新值处理后呈目标分布。使得预处理后的所述 N 个原始样本数据和所述 K 个增强样本数据更适合输入目标模型。便于进行目标模型的训练，有利于改善目标模型的训练效果，提高目标模型的泛化能力。

示例性的，所述分布调整函数为逆误差函数(erfinv)。逆误差函数可由麦克劳林级数表示：

$$\mathrm{erf}^{-1}(z)=\sum_{k=0}^{\infty}\frac{c_k}{2k+1}\left(\frac{\sqrt{\pi}}{2}z\right)^{2k+1}$$

其中， $\mathrm{erf}^{-1}(z)$ 表示所述逆误差函数，z 表示所述更新值，$C_0=1$，π 表示圆周率，$C_k=\sum_{m=0}^{k-1}\frac{c_m c_{k-1-m}}{(m+1)(2m+1)}$，即 $\{C_0,C_1,C_2,C_3,...\}=\left\{1,1,\frac{7}{6},\frac{127}{90},...\right\}$，k 和 m 为整数。

在一些实施例中，在数据处理系统 200 通过预设的分布调整函数对所述 N+K 个特征值对应的更新值进行处理之前，S220 还包括：

数据处理系统 200 将所述 N+K 个特征值对应的更新值调整至所述分布调整函数要求的输入区间。

示例性的，所述 N+K 个特征值对应的更新值在闭区间[-1, 1]内。但是，逆误差函数的输入区间为（-1,1）即，逆误差函数的输入区间不包括-1 和 1。如果数据处理系统 200 将-1 和 1 输入到逆误差函数中，会导致处理后的更新值为无穷值，无法进行后续的模型训练。为了避免这一问题，数据处理系统 200 需要将所述 N+K 个特征值对应的所有更新值中的端值处理一下。

例如，数据处理系统 200 确定所述 N+K 个特征值对应的更新值是否包含所述目标区间的端值。若包含，则需要将所述端值对应的更新值进行处理。

比如，目标区间为[-1, 1]。若数据处理系统 200 确定某个特征值对应的更新值为-1，则将该特征值对应的更新值处理为-1+ε。若数据处理系统200确定某个特征值对应的更新值为1，则将该特征值对应的更新值处理为1-ε。ε可以为预设的接近0的数。比如，ε=1e-6=0.000001。

数据处理系统 200 进行端值处理后，所述 N+K 个特征值对应的更新值位于所述逆误差函数要求的输入区间(-1, 1)内。

在该实施例中，数据处理系统 200 对所述 N+K 个特征值对应的更新值中不符合所述输入区间的更新值进行处理，使其满足所述输入区间。避免数据处理系统200将不满足所述输入区间的更新值输入到所述分布调整函数之后，得到所述目标区间之外的值。从而影响后续的目标模型训练，或者导致不能后续的目标模型训练。

数据处理系统200按照第三特征的预处理操作，将所述多个特征中的部分或者全部特征分别进行预处理操作。而后，数据处理系统 200 得到预处理后的所述 N 个原始样本数据和所述 K 个增强样本数据。

S230：基于所述预处理后的 M 个样本数据，对目标模型进行训练。

在一些实施例中，数据处理系统 200 基于所述预处理后的所述 N 个原始样本数据和所述 K 个增强样本数据，对目标模型进行训练。

示例性的，数据处理系统 200 可以将所述预处理后的所述 N 个原始样本数据和 K 个增强样本数据分为训练集和验证集，并将训练集中的样本数据分为多个 epoch。数据处理系统 200 在每轮迭代训练中使用一个或者多个 epoch。如此重复进行多轮迭代训练，直到数据处理系统200基于验证集、目标函数确定训练得到的目标模型达到训练结束条件。

在一些实施例中，数据处理系统200基于所述预处理后的所述N个原始样本数据，对目标模型进行训练。具体的训练过程参考上述示例，不再赘述。

综上所述，本说明书提供的样本数据的处理方法 P200、设备及存储介质，数据处

理系统200基于特征值的排序索引进行预处理。在预处理过程中更多考虑的是特征值的相对排序，在异常特征值和正常特征值的差距较大时，降低了异常特征值对预处理的影响。并且由于不同特征的特征值的量纲/尺度可能不同，基于排序索引预处理的这种方式，可以减少特征值的量纲/尺度之间的差异对目标模型的训练影响，从而提升目标模型的训练效果和性能。

可以理解，方法 P100 和方法 P200 主要以交易领域为例进行描述。当方法 P100 和方法 P200 用于其他领域时，相关术语可以适应性调整。比如上文列举的学习任务、特征、特征值的举例等。

本说明书另一方面提供一种计算机可读非暂时性存储介质，存储有至少一组用来处理样本数据的可执行指令。当所述可执行指令被处理器执行时，所述可执行指令指导所述处理器实施本说明书所述的方法 P100 和/或 P200 的步骤。在一些可能的实施方式中，本说明书的各个方面还可以实现为一种程序产品的形式，其包括程序代码。当所述程序产品在数据处理系统200上运行时，所述程序代码用于使数据处理系统200执行本说明书描述的方法 P100 和/或 P200 的步骤。用于实现上述方法的程序产品可以采用便携式紧凑盘只读存储器（CD-ROM）包括程序代码，并可以在数据处理系统200上运行。然而，本说明书的程序产品不限于此，在本说明书中，可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统使用或者与其结合使用。所述程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以为但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。可读存储介质的更具体的例子包括：具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、光纤、便携式紧凑盘只读存储器（CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。所述计算机可读存储介质可以包括在基带中或者作为载波一部分传播的数据信号，其

中承载了可读程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。可读存储介质还可以是可读存储介质以外的任何可读介质，该可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。可读存储介质上包含的程序代码可以用任何适当的介质传输，包括但不限于无线、有线、光缆、RF 等等，或者上述的任意合适的组合。可以以一种或多种程序设计语言的任意组合来编写用于执行本说明书操作的程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如 Java、C++等，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在数据处理系统 200 上执行、部分地在数据处理系统 200 上执行、作为一个独立的软件包执行、部分在数据处理系统 200 上部分在远程计算设备上执行、或者完全在远程计算设备上执行。

本说明书提及的“多个”，如果没有特殊的声明，应当理解为两个或者两个以上。

上述对本说明书特定实施例进行了描述。其他实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者是可能有利的。

综上所述，在阅读本详细公开内容之后，本领域技术人员可以明白，前述详细公开内容可以仅以示例的方式呈现，并且可以不是限制性的。尽管这里没有明确说明，本领域技术人员可以理解本说明书需求囊括对实施例的各种合理改变，改进和修改。这些改变，改进和修改旨在由本说明书提出，并且在本说明书的示例性实施例的精神和范围内。

此外，本说明书中的某些术语已被用于描述本说明书的实施例。例如，“一个实施例”，“实施例”和/或“一些实施例”意味着结合该实施例描述的特定特征，结构或特性可

以包括在本说明书的至少一个实施例中。因此，可以强调并且应当理解，在本说明书的各个部分中对“实施例”或“一个实施例”或“替代实施例”的两个或更多个引用不一定都指代相同的实施例。此外，特定特征，结构或特性可以在本说明书的一个或多个实施例中适当地组合。

应当理解，在本说明书的实施例的前述描述中，为了帮助理解一个特征，出于简化本说明书的目的，本说明书将各种特征组合在单个实施例、附图或其描述中。然而，这并不是说这些特征的组合是必须的，本领域技术人员在阅读本说明书的时候完全有可能将其中一部分设备标注出来作为单独的实施例来理解。也就是说，本说明书中的实施例也可以理解为多个次级实施例的整合。而每个次级实施例的内容在于少于单个前述公开实施例的所有特征的时候也是成立的。

本文引用的每个专利，专利申请，专利申请的出版物和其他材料，例如文章、书籍、说明书、出版物、文件、物品等，除了与本文件不一致或相冲突的、或者对权利要求的最宽范围具有限制性影响的文件，均可以通过引用结合于此，并用于现在或以后与本文件相关联的所有目的。此外，在任何材料对相关术语的描述、定义和/或使用与本文件对相关术语的描述、定义和/或使用之间存在任何不一致或冲突时，以本文件中的术语为准。

最后，应理解，本文公开的申请的实施方案是对本说明书的实施方案的原理的说明。其他修改后的实施例也在本说明书的范围内。因此，本说明书披露的实施例仅仅作为示例而非限制。本领域技术人员可以根据本说明书中的实施例采取替代配置来实现本说明书中的申请。因此，本说明书的实施例不限于申请中被精确地描述过的实施例。

权 利 要 求 书

1. 一种样本数据的处理方法，包括：

获得 N 个原始样本数据，每个原始样本数据包括多个特征的特征值，所述 N 为大于 1 的整数；

基于所述 N 个原始样本数据生成 K 个增强样本数据，其中，每个增强样本数据中的所述多个特征的特征值来自于所述 N 个原始样本数据的至少两个原始样本数据，所述 K 为大于或等于 1 的整数；以及

基于所述 N 个原始样本数据和所述 K 个增强样本数据对目标模型进行训练。

2. 根据权利要求 1 所述的方法，其中，每个原始样本数据对应有标签信息，所述至少两个原始样本数据对应的所述标签信息相同。

3. 根据权利要求 1 所述的方法，其中，所述至少两个原始样本数据包括第一原始样本数据和至少一个第二原始样本数据，每个增强样本数据是通过如下方式生成的：

生成所述第一原始样本数据的副本；

将所述副本中与至少一个目标特征对应的特征值替换为所述至少一个第二原始样本数据中与所述至少一个目标特征对应的特征值，所述至少一个目标特征为所述多个特征中的部分特征；以及

将替换后的所述副本作为所述增强样本数据。

4. 根据权利要求 3 所述的方法，其中，所述至少一个目标特征在所述多个特征中的占比小于或等于预设占比。

5. 根据权利要求 4 所述的方法，其中，在所述将所述副本中与至少一个目标特征对应的特征值替换为所述至少一个第二原始样本数据中与所述至少一个目标特征对应的特征值之前，所述方法还包括：

基于所述预设占比，在所述多个特征中随机选择部分特征作为所述至少一个目标特征。

6. 根据权利要求 3 所述的方法，其中，所述至少一个目标特征包括第一特征和第二特征，所述至少一个第二原始样本数据包括第三原始样本数据和第四原始样本数据；

所述将所述副本中与至少一个目标特征对应的特征值替换为所述至少一个第二原始样本数据中与所述至少一个目标特征对应的特征值，包括：

将所述副本中与所述第一特征对应的特征值替换为所述第三原始样本数据中与所述第一特征对应的特征值，以及

将所述副本中与所述第二特征对应的特征值替换为所述第四原始样本数据中与所述第二特征对应的特征值。

7. 根据权利要求 1 所述的方法，其中，所述基于所述 N 个原始样本数据和所述 K 个增强样本数据，对目标模型进行训练，包括：

基于所述 N 个原始样本数据和所述 K 个增强样本数据，针对所述多个特征中的至少部分特征分别执行预处理操作，得到预处理后的 N 个原始样本数据和 K 个增强样本数据，其中，所述至少部分特征包括第三特征，所述第三特征对应的所述预处理操作包括：

从所述 N 个原始样本数据和所述 K 个增强样本数据中获得与所述第三特征对应的 N+K 个特征值，

确定所述 N+K 个特征值对应的排序索引，以及

基于所述 N+K 个特征值对应的排序索引确定所述 N+K 个特征值对应的更新值，所述 N+K 个特征值对应的更新值均位于目标区间内；以及

基于所述预处理后的 N 个原始样本数据和所述 K 个增强样本数据，对所述目标模型进行训练。

8. 根据权利要求 7 所述的方法，其中，所述确定所述 N+K 个特征值对应的排序索引，包括：

对所述 N+K 个特征值进行排序，并基于排序后的所述 N+K 个特征值的顺序得到所述 N+K 个特征值对应的排序索引。

9. 根据权利要求 7 所述的方法，其中，所述基于所述 N+K 个特征值对应的排序索引确定所述 N+K 个特征值对应的更新值，包括：

针对所述 N+K 个特征值中的每个特征值，基于所述特征值对应的排序索引与最大索引确定所述特征值对应的更新值，所述最大索引为所述 N+K 个特征值对应的排序索引中的最大值。

10. 根据权利要求 9 所述的方法，其中，所述基于所述特征值对应的排序索引与最大索引确定所述特征值对应的更新值，包括：

基于以下公式得到所述特征值对应的更新值：

$$X=(\frac{x_i}{x_{max}}-0.5)\times 2$$

其中，X表示所述特征值对应的更新值，x_i表示所述特征值对应的排序索引，x_{max}表示所述最大索引。

11. 根据权利要求 7 所述的方法，其中，在基于所述 N+K 个特征值对应的排序索引确定所述 N+K 个特征值对应的更新值之后，所述方法还包括：

通过预设的分布调整函数对所述 N+K 个特征值对应的更新值进行处理，使得处理后的所述 N+K 个特征值对应的更新值呈目标分布。

12. 根据权利要求 11 所述的方法，其中，在所述通过预设的分布调整函数对所述 N+K 个特征值对应的更新值进行处理之前，所述方法还包括：

将所述 N+K 个特征值对应的更新值调整至所述分布调整函数要求的输入区间。

13.一种样本数据的处理系统，包括：

至少一个存储介质，存储有至少一个指令集，用于处理样本数据；以及

至少一个处理器，同所述至少一个存储介质通信连接，其中，所述至少一个处理器运行时读取所述至少一个指令集，并且根据所述至少一个指令集的指示：

获得 N 个原始样本数据，每个原始样本数据包括多个特征的特征值，所述 N 为大于 1 的整数，

基于所述 N 个原始样本数据生成 K 个增强样本数据，其中，每个增强样本数据中的所述多个特征的特征值来自于所述 N 个原始样本数据的至少两个原始样本数据，所述 K 为大于或等于 1 的整数，以及

基于所述 N 个原始样本数据和所述 K 个增强样本数据对目标模型进行训练。

14. 根据权利要求 13 所述的处理系统，其中，每个原始样本数据对应有标签信息，所述至少两个原始样本数据对应的所述标签信息相同。

15. 根据权利要求 13 所述的处理系统，其中，所述至少两个原始样本数据包括第一原始样本数据和至少一个第二原始样本数据，每个增强样本数据是通过如下方式生成的：

生成所述第一原始样本数据的副本；

将所述副本中与至少一个目标特征对应的特征值替换为所述至少一个第二原始样本数据中与所述至少一个目标特征对应的特征值，所述至少一个目标特征为所述多个特征中的部分特征；以及

将替换后的所述副本作为所述增强样本数据。

16. 根据权利要求 15 所述的处理系统，其中，所述至少一个目标特征在所述多个特征中的占比小于或等于预设占比。

17. 根据权利要求 16 所述的处理系统，其中，在所述将所述副本中与至少一个目标特征对应的特征值替换为所述至少一个第二原始样本数据中与所述至少一个目标特征对应的特征值之前，所述至少一个处理器还：

基于所述预设占比，在所述多个特征中随机选择部分特征作为所述至少一个目标特征。

18. 根据权利要求 15 所述的处理系统，其中，所述至少一个目标特征包括第一特征和第二特征，所述至少一个第二原始样本数据包括第三原始样本数据和第四原始样本数据；为了将所述副本中与至少一个目标特征对应的特征值替换为所述至少一个第二原始样本数据中与所述至少一个目标特征对应的特征值，所述至少一个处理器：

将所述副本中与所述第一特征对应的特征值替换为所述第三原始样本数据中与所述第一特征对应的特征值，以及

将所述副本中与所述第二特征对应的特征值替换为所述第四原始样本数据中与所述第二特征对应的特征值。

19. 根据权利要求 13 所述的处理系统，其中，为了对目标模型进行训练，所述至少一个处理器：

基于所述 N 个原始样本数据和所述 K 个增强样本数据，针对所述多个特征中的至少部分特征分别执行预处理操作，得到预处理后的 N 个原始样本数据和 K 个增强样本数据，其中，所述至少部分特征包括第三特征，所述第三特征对应的所述预处理操作包括：

从所述 N 个原始样本数据和所述 K 个增强样本数据中获得与所述第三特征对应的 N+K 个特征值，

确定所述 N+K 个特征值对应的排序索引，以及

基于所述 N+K 个特征值对应的排序索引确定所述 N+K 个特征值对应的更新值，所述 N+K 个特征值对应的更新值均位于目标区间内；以及

基于所述预处理后的 N 个原始样本数据和所述 K 个增强样本数据，对所述目标模型进行训练。

20. 一种计算机可读非暂时性存储介质，其中，所述计算机可读非暂时性存储介质中存储有至少一个指令集，所述至少一个指令集被至少一个处理器执行时，实现如权利要求 1-12 中任一项所述的方法。