(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2017 年 8 月 10 日 (10.08.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/133569 A1

(51) 国际专利分类号:
*G06F 17/30* (2006.01)

(21) 国际申请号: PCT/CN2017/072405

(22) 国际申请日: 2017 年 1 月 24 日 (24.01.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610082141.1 2016 年 2 月 5 日 (05.02.2016) CN

(71) 申请人: **阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED)** [—/CN]; 英属开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人: **姜晓燕 (JIANG, Xiaoyan)**; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 **王少萌 (WANG, Shaomeng)**; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 **杨旭 (YANG, Xu)**; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 **蔡宁 (CAI, Ning)**; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: **北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.)**; 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (CN)。

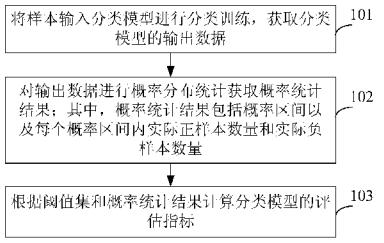
(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO,

[见续页]

(54) **Title**: EVALUATION INDEX OBTAINING METHOD AND DEVICE

(54) **发明名称** : 评估指标获取方法及装置




101 Input samples to a classification model for classification training, and obtain output data of the classification model

102 Collect probability distribution statistics about the output data to obtain a probability statistics result, wherein the probability statistics result comprises probability intervals as well as the actual number of positive samples and the actual number of negative samples in each probability interval

103 Calculate evaluation indexes of the classification model according to a threshold set and the probability statistics result

图 1

(57) **Abstract**: An evaluation index obtaining method and device. The method comprises: inputting samples to a classification model for classification training, and obtaining output data of the classification model (101); collecting probability distribution statistics about the output data to obtain a probability statistics result, wherein the probability statistics result comprises probability intervals as well as the actual number of positive samples and the actual number of negative samples in each probability interval (102); and calculating evaluation indexes of the classification model according to a threshold set and the probability statistics result (103). By collecting probability statistics about the output data of the classification model, and calculating the evaluation indexes on the basis of the obtained probability statistics result, the method and device resolve the problem of repeatedly scanning output data during a calculation process of evaluation indexes, and can improve the calculation efficiency of the evaluation indexes when the output data is large-scale data.

(57) **摘要**: 一种评估指标获取方法及装置，该方法包括：通过将样本输入分类模型进行分类训练，获取分类模型的输出数据（101），对输出数据进行概率分布统计获取概率统计结果，其中概率统计结果包括概率区间以及每个概率区间内实际正样本数量和实际负样本数量（102），根据阈值集和概率统计结果计算分类模型的评估指标（103）。该方法和装置通过对分类模型的输出数据进行概率统计，基于得到的概率统计结果对评估指标进行计算，解决了在评估指标的计算过程中多次扫描输出数据的问题，尤其在输出数据为大规模数据时可以提高评估指标的计算效率。

WO 2017/133569 A1

RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

**本国际公布**:

— 包括国际检索报告(条约第 21 条(3))。

# 评估指标获取方法及装置

本申请要求2016年02月05日递交的申请号为201610082141.1、发明名称为“评估指标获取方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

## 技术领域

本发明属于数据处理领域，尤其涉及一种评估指标获取方法及装置。

## 背景技术

在大数据挖掘的业务场景中，经常需要针对超大规模的数据使用分类算法进行训练分类。当前分类算法有很多，而且不同的分类算法又用很多不同的变种。当根据分类算法建立一个分类模型之后，会考虑该分类模型的性能或准确率，因此需要对该分类模型的优良情况进行评估。目前，二分类算法模型的评估指标包括：混淆矩阵、受试者工作特征曲线（receiver operating characteristic curve，简称ROC）图、ROC图中的曲线下的面积（Area Under RocCurve，简称AUC）值与提升（Lift）图等指标。

现有的对二分类算法对应的分类模型的评估方法或者系统中，在获取评估指标的过程中，每当输入一个阈值点时，在计算与该阈值点对应的评估参数时，就需要对分类模型的输出数据进行一次扫描。经过大量阈值点的输入，然后获取到该分类模型的评估指标。对大规模数据来说，通过多次扫描分类模型的输出数据，获取该分类模型评估指标的方式存在计算效率较低的问题。

## 发明内容

本发明提供一种评估指标获取方法及装置，用于解决通过多次扫描分类模型的输出数据来获取评估指标的方式存在计算效率较低的问题。

为了实现上述目的，本发明提供了一种评估指标获取方法，包括：

将样本输入分类模型进行分类训练，获取分类模型的输出数据；

对所述输出数据进行概率分布统计获取概率统计结果；其中，所述概率统计结果包括概率区间以及每个概率区间内实际正样本数量和实际负样本数量；

根据阈值集和所述概率统计结果计算所述分类模型的评估指标。

为了实现上述目的，本发明提供了一种评估指标获取装置，包括：

分类训练模块，用于将样本输入分类模型进行分类训练，获取分类模型的输出数据；

概率统计模块，用于对所述输出数据进行概率分布统计获取概率统计结果；其中，所述概率统计结果包括概率区间以及每个概率区间内实际正样本数量和实际负样本数量；

计算模块，用于根据阈值集和所述概率统计结果计算所述分类模型的评估指标。

本发明提供的评估指标获取方法及装置，通过对分类模型的输出数据进行概率统计，基于得到的包括概率区间以及对应的实际正样本和实际负样本数量的概率统计结果对评估指标进行计算，解决了在评估指标的计算过程中多次扫描输出数据的问题，尤其在输出数据为大规模数据时可以提高评估指标的计算效率。

## 附图说明

图 1 为本发明实施例一的评估指标获取方法的流程示意图；

图 2 为本发明实施例二的评估指标获取方法的流程示意图；

图 3 为本发明实施例二的评估指标获取方法的应用示例示意图之一；

图 4 为本发明实施例二的评估指标获取方法的应用示例示意图之二；

图 5 为本发明实施例三的评估指标获取装置的结构示意图；

图 6 为本发明实施例四的评估指标获取装置的结构示意图。

## 具体实施方式

下面结合附图对本发明实施例提供的评估指标获取方法及装置进行详细描述。

实施例一

如图 1 所示，其为本发明实施例一的评估指标获取方法的流程示意图。该评估指标获取方法包括以下步骤：

S101、将样本输入分类模型进行分类训练，获取分类模型的输出数据。

二分类算法对应的分类模型将样本分成正样本或者负样本。在分类模型中往往将正样本用“1”表示，将负样本用“0”表示。其中，输入分类模型的每个样本都有一个原始的样本属性。本实施例中，样本属性包括正样本属性和负样本属性。原始的样本属性表示样本实际是正样本还是负样本。

为了对分类模型进行评估，需要将样本输入分类模型中进行分类训练，在训练完成后，分类模型会对每个样本进行分类和概率预测。具体地，分类模型在训练完成后为每个样本输出训练后的样本属性，训练后的样本属性可以指示出样本经过分类模型后是正

样本还是负样本。

进一步地，分类模型在训练完成后还会为每个样本进行概率预测，用户可以根据实际需要选择输出每个样本经过分类模型预测成正样本的概率，或者选择输出每个样本经过分类模型预测成负样本的概率。其中，样本经过分类模型被预测成正样本的概率和被预测成负样本的概率的和为1。

S102、对输出数据进行概率分布统计获取概率统计结果；其中，概率统计结果包括概率区间以及每个概率区间内实际正样本数量和实际负样本数量。

在获取到输出数据后，由于分类模型会对每个样本进行概率预测，这样输出数据中每个样本会有一个预测概率，本实施例中，分类模型输出的每个样本的概率为每个样本被分类模型预测成正样本的预测概率。

进一步地，根据预测概率对输出数据进行概率分布统计，获取概率统计结果。在进行概率统计时首先需要划分概率区间，然后在每个概率区间内基于输出数据中每个样本原始的样本属性统计实际正样本数量和实际负样本数量，得到正样本和负样本的概率分布图，基于正样本的概率分布图获取每个概率区间内实际正样本数量，基于负样本的概率分布图获取每个概率区间内实际负样本数量。

优选地，基于直方图算法对输出数据进行概率分布的统计，获取正样本的直方图和负样本的直方图，基于正样本的直方图和负样本的直方图能够获取到上述概率统计结果。

S103、根据阈值集和概率统计结果计算分类模型的评估指标。

在获取到概率统计结果后，需要获取阈值集，其中阈值集中包括多个阈值点，然后基于每个阈值点和概率统计结果中每个概率区间内实际正样本数据和实际负样本数据，获取每个阈值点对应的评估参数，利用所有阈值点对应的评估参数生成分类模型的评估指标。

本实施例中，在概率统计结果后，可以将概率统计结果中的概率区间的端点值作为阈值点构成阈值集。例如，可以利用每个概率区间的下限值作为阈值点构成阈值集。或者将部分概率区间的下限值作为阈值点构成阈值集。再例如，可以将概率区间的上限值作为阈值点构成阈值集。本实施例中在概率统计的过程中，对概率区间进行划分，概率区间的端点可作为分界点，直接将概率区间的端点值作为阈值点，不需要进行阈值点的重新设定，进而提高了评估指标的计算效率。

可选地，可以接收用户输入的利用概率区间的端点值作为阈值点构成阈值集。例如，用户可以将每个概率区间的下限值作为阈值点构成阈值集，或者用户选取部分概率区间

的下限值作为阈值点构成阈值集。本实施例中，用户根据反馈的概率统计结果，可以初步对分类模型的效果有一定的了解，从而能够选取合适的阈值点构成阈值集，用户交互较好，而且对分类模型的评估更加准确。

进一步地，在获取到阈值集后，根据阈值集中的阈值点和概率统计结果计算评估指标。其中，评估指标包括混淆矩阵、ROC 曲线、AUC 值和 Lift 图。

其中，混淆矩阵中包括：实际为正样本预测为正样本的数量（True Positives，简称 TP）、实际为负样本预测为正样本的数量（False Positives，简称 FP）、实际为负样本预测为负样本的数量（True Negatives，简称 TN）和实际为正样本预测为负样本的数量（False Negatives，简称 FN）。

在获取到阈值点之后，将阈值点作为分界点，对于正样本的概率分布来说，大于阈值点的所有概率区间内实际正样本被分类模型预测成正样本，对实际正样本被分类模型预测成正样本的数量进行累积，将累积的实际正样本被分类模型预测成正样本的数量作为混淆矩阵的 TP。而小于阈值点的所有概率区间内实际正样本被分类模型预测成负样本，对实际正样本被分类模型预测成负样本的数量进行累计，将累计后的实际正样本被分类模型预测成负样本的数量作为混淆矩阵的 FP。

对于负样本的概率分布来说，大于阈值点的所有概率区间内实际负样本被分类模型预测成正样本，对实际负样本被分类模型预测成正样本的数量进行累积，将累积的实际负样本被分类模型预测成正样本的数量作为混淆矩阵的 FN。而小于阈值点的所有概率区间内实际负样本被分类模型预测成负样本，对实际负样本被分类模型预测成负样本的数量进行累计，将累计后的实际负样本被分类模型预测成负样本的数量作为混淆矩阵的 TN。

在获取到阈值点对应的混淆矩阵后，可以利用混淆矩阵中的 TP、FP、TN 和 FN，计算得到其他评估指标的该阈值点对应的评估参数，当所有阈值点对应的评估参数计算完成后，利用每个阈值点对应的评估参数生成评估指标。例如，根据一个阈值点对应的混淆矩阵可以计算出在该阈值点处 ROC 曲线的坐标，将坐标作为该阈值点 ROC 曲线的评估参数。当所有阈值点对应的评估参数计算完成后，利用每个阈值点对应的 ROC 曲线的坐标绘制 ROC 曲线。

本实施例提供的评估指标获取方法，通过对分类模型的输出数据进行概率统计，基于得到包括概率区间以及每个概率区间内实际正样本数量和实际负样本数量的概率统计结果对评估指标进行计算，解决了在评估指标的计算过程中多次扫描输出数据的问题，

尤其在输出数据为大规模数据时可以提高评估指标的计算效率。

实施例二

如图 2 所示，其为本发明实施例二的评估指标获取方法的流程示意图。该评估指标获取方法包括以下步骤：

S201、将样本输入分类模型进行分类训练，获取分类模型的输出数据。

为了对分类模型进行评估，需要将样本输入分类模型中进行分类训练，在训练完成后，分类模型会对每个样本进行分类和概率预测。具体地，分类模型在训练完成后为每个样本输出训练后的样本属性，训练后的样本属性可以指示出样本经过分类模型后是正样本还是负样本。进一步地，分类模型在训练完成后还会为每个样本进行概率预测，一般分类模型会选择输出每个样本经过分类模型预测成正样本的概率。

本实施例中，分类模型进行分类训练后的输出数据中包括：每个样本原始的样本属性以及每个样本被分类模型预测成正样本的预测概率。本实施例中，样本属性包括正样本属性和负样本属性。在分类模型中往往将正样本用“1”表示，将负样本用“0”表示。

S202、基于直方图算法对输出数据进行概率区间划分，统计每个概率区间内实际正样本数量和实际负样本数量。

具体地，对分类模型的输出数据进行扫描。本实施例中，假设分类器的输出表格式为：原始的样本属性、分类模型的预测后样本属性以及样本被分类模型预测成正样本的预测概率。一般情况下，分类模型可以设置有选择项，可以选择输出样本被分类模型预测成正样本的预测概率或者样本被分类模型预测成正样本的预测概率。相应地，可以选择生成正样本对应的 ROC 曲线和 Lift 图，或者选择生成负样本对应的 ROC 曲线和 Lift 图，本实施例中以正样本为例。

进一步地，根据每个样本被预测成正样本的预测概率和输出数据中每个样本原始的样本属性生成正样本对应的第一直方图和负样本对应的第二直方图。其中，第一直方图的横轴是预测概率，第一直方图的纵轴是实际正样本数量，第二直方图的横轴是预测概率，第二直方图的纵轴是实际负样本数量。

在生成第一直方图和第二直方图的过程中，两个直方图的概率区间可能不同步，为了获取到一致的概率区间，需要调整横轴步长使第一直方图和第二直方图的概率区间一致，在概率区间调整一致后，可以获取到概率统计结果中的概率区间。

在获取到概率区间后，可以从第一直方图中统计获取每个概率区间内实际正样本的数量，以及可以从第二直方图中统计获取每个概率区间内实际负样本的数量。

S203、获取阈值点构成的阈值集。

在生成了概率区间后，可以将概率区间的端点值作为阈值点，构成阈值集，可选地，将部分概率区间的下限值或者上限值作为阈值点构成阈值集，例如，选取每隔一个概率区间选取一个下限值作为阈值点构成阈值集。本实施例中，在概率统计的过程，完成概率区间的划分，概率区间的端点值能够作为分界点，从而可将概率区间的端点值作为阈值点构成阈值集，不需要在对阈值进行重新设定，进而提高了评估指标的计算效率。

可选地，在获取到概率区间后，可以将概率统计结果反馈给用户，以使用户利用概率区间的端点值作为阈值点构成阈值集。例如，用户可以将每个概率区间的下限值作为阈值点作为阈值集，或者用户选取部分概率区间的下限值作为阈值点构成阈值集可以选取部分概率区间的端点值作为阈值点构成阈值集。在获取到阈值集后，用户输入阈值集进行计算评估指标。本实施例中，通过直方图的统计过程，用户根据反馈的概率统计结果，可以初步对分类模型的效果有一定的了解，从而能够选取合适的阈值点构成阈值集，用户交互较好，而且对分类模型的评估更加准确。

S204、按照由大到小的顺序获取阈值集中每个阈值点对应的混淆矩阵。

其中，混淆矩阵包括实际为正样本被预测为正样本的数量 TP、实际为正样本被预测为负样本的数量 FP、实际为负样本被预测为负样本的数量 TN、实际为负样本被预测为正样本的数量 FN，如下表 1 所示。

表 1 为混淆矩阵的示意表

| | | 预测的样本属性 | |
|---|---|---|---|
| | | 1 | 0 |
| 原始的样本属性 | 1 | TP | FP |
| | 0 | FN | TN |

具体地，对于正样本对应的第一直方图，按照阈值点的大小顺序逐次对大于阈值点的所有概率区间内实际正样本数量进行累积得到 TP，以及对小于阈值点的所有概率区间内实际正样本数量进行累积得到 FN。

对于负样本对应的第二直方图，按照阈值点的大小顺序逐次对大于阈值点的所有概率区间内负样本数量进行累积得到 FP，以及对小于阈值点的所有概率区间内负样本数量进行累积得到 TN。

S205、将每个阈值点对应的混淆矩阵作为评估指标。

S206、针对每一个阈值点，根据混淆矩阵获取对应的 ROC 坐标。

S207、利用每个阈值点的 ROC 坐标绘制 ROC 曲线。

S208、获取每个由相邻阈值点对应的 ROC 坐标与 ROC 曲线构成的曲边梯形的面积。

S209、将所有曲边梯形的面积相加得到 ROC 曲线的 AUC 值。

在获取到每个阈值点的混淆矩阵后，根据混淆矩阵可以获取到分类模型其他的评估指标，例如 ROC 曲线、ROC 曲线下面积 AUC 值以及 Lift 图。

具体地，针对每一个阈值点，将 FP 与实际负样本总量的比值作为 ROC 的横坐标，以及将 TP 与实际正样本总量的比值作为 ROC 的纵坐标。在获取到每个阈值点对应的 ROC 坐标后，对所有阈值点对应的 ROC 坐标进行描点绘制 ROC 曲线。

进一步地，在绘制出 ROC 曲线后，由相邻阈值点对应的 ROC 坐标与 ROC 曲线可以构成一个曲边梯形，根据相邻的 ROC 坐标能够计算一个曲边梯形的面积。在获取到所有的曲边梯形的面积后，将所有面积相加得到该 ROC 曲线的 AUC 值。

S210、针对每一个阈值点，根据混淆矩阵获取对应的 Lift 坐标。

具体地，针对每一个阈值点，将 TP 和 FP 的和值与样本总量的比值作为 Lift 图的横坐标，以及将 TP 作为 Lift 图的纵坐标。

S211、利用每个阈值点对应的 Lift 坐标绘制 Lift 图。

进一步地，在获取到每个阈值点对应的 Lift 坐标后，将所有阈值点对应的 Lift 坐标绘制 Lift 图。

S212、接收用户的显示指令，根据显示指令将评估指标进行可视化展示。

在获取到评估指标后，用户可以发送显示评估指标的显示指令，在接收到显示指令后，向用户可视化展示计算出的评估指标，使得用户能够直观地判断分类模型的优良情况。

本实施例中，可以在服务器上执行该评估指标获取方法，在计算出评估指标，用户可以向服务器进行发送显示指令，在接收到显示指令，服务器可以将评估指标下发给本地终端，这样本地终端通过显示屏将评估指标进行可视化展示，如向用户展示 ROC 曲线、Lift 图等。

可选地，对于大规模数据，计算直方图时数据量较大，可以在服务器上进行计算，在计算完直方图后，可以将直方图结果下发到本地终端，在本地终端上计算评估指标，这样可以减缓服务器的压力。在计算出评估指标后，用户可以向本地终端发送显示指令，在接收到显示指令后，本地终端通过显示屏将评估指标进行可视化展示，如向用户展示 ROC 曲线、Lift 图等。当用户点击 ROC 曲线上的点时，可以将该点对应的混淆矩阵进行

展示。

可选地，可以在本地终端上执行该评估指标获取方法，在计算出评估指标后，用户可以向本地终端发送显示指令，在接收到显示指令后，在显示屏上进行可视化展示，如向用户展示 ROC 曲线、Lift 图等。当用户点击 ROC 曲线上的点时，可以将该点对应的混淆矩阵进行展示。

为了更好地理解本实施例提供的评估指标获取方法，下面举例进行说明：样本为用户 0~用户 99，样本用户具有如下的特征参数：年龄（age）、工作性质（workclass）、取样量（fnlwgt）学历（education）、教育程度（education_num）、婚姻状况（matrital_status）、职业（occupation）、家庭情况（relationship）、种族（race）、性别（sex）、资本收益（capital_gain）、资本损失（capital_loss）、每周工作时长（hours_per_week）、国籍（native_country）等，将这些用户的特征参数输入到分类模型中进行分类训练，能够获取到一个用于用户收入情况的分类结果。在该例子中用“0”表示为低收入，“1”表示高收入。将高收入作为正样本属性，将低收入作为负样本属性。分类模型的输出数据中包括每个样本原始的样本属性、预测的样本属性以及每个样本被预测成高收入类别的概率，如下表 2 所示。

表 2 为分类模型的输出数据

| 样本序号 | 原始的样本属性 | 预测的样本属性 | 样本被预测成高收入的概率 |
|---|---|---|---|
| 0 | 0 | 0 | 0.059091 |
| 1 | 0 | 0 | 0.205556 |
| 2 | 0 | 0 | 0.009091 |
| 3 | 0 | 0 | 0.026482 |
| 4 | 0 | 0 | 0.283333 |
| 5 | 0 | 0 | 0.1 |
| 6 | 0 | 0 | 0.059091 |
| 7 | 1 | 1 | 0.65235 |
| 8 | 1 | 0 | 0.475366 |
| 9 | 1 | 1 | 0.683333 |
| 10 | 1 | 1 | 0.885714 |
| 11 | 1 | 1 | 0.738889 |
| 12 | 0 | 0 | 0.023377 |

| | | | |
|---|---|---|---|
| 13 | 0 | 0 | 0.109091 |
| 14 | 1 | 1 | 0.631795 |
| 15 | 0 | 0 | 0.026482 |
| 16 | 0 | 0 | 0.009091 |
| 17 | 0 | 0 | 0.009091 |
| 18 | 0 | 0 | 0.026482 |
| 19 | 1 | 1 | 0.655556 |
| 20 | 1 | 1 | 0.975 |
| 21 | 0 | 0 | 0.009091 |
| 22 | 0 | 0 | 0.076482 |
| 23 | 0 | 0 | 0.026482 |
| 24 | 0 | 0 | 0.05 |
| 25 | 1 | 1 | 0.854762 |
| 26 | 0 | 0 | 0.02 |
| 27 | 1 | 1 | 0.744805 |
| 28 | 0 | 0 | 0.067033 |
| 29 | 0 | 0 | 0.250128 |
| 30 | 0 | 0 | 0.171429 |
| 31 | 0 | 0 | 0.009091 |
| 32 | 0 | 0 | 0.15 |
| 33 | 0 | 0 | 0.109091 |
| 34 | 0 | 0 | 0.244805 |
| 35 | 0 | 0 | 0.009091 |
| 36 | 0 | 0 | 0.009091 |
| 37 | 0 | 0 | 0.090768 |
| 38 | 1 | 1 | 0.534805 |
| 39 | 0 | 0 | 0.12735 |
| 40 | 0 | 0 | 0.026482 |
| 41 | 0 | 0 | 0.238889 |
| 42 | 0 | 0 | 0.278571 |
| 43 | 0 | 0 | 0.023377 |
| 44 | 0 | 0 | 0.009091 |

| | | | |
|---|---|---|---|
| 45 | 1 | 1 | 0.671429 |
| 46 | 0 | 0 | 0.051482 |
| 47 | 0 | 0 | 0.033333 |
| 48 | 0 | 0 | 0.24735 |
| 49 | 0 | 0 | 0.148571 |
| 50 | 0 | 0 | 0.166667 |
| 51 | 0 | 0 | 0.049091 |
| 52 | 1 | 1 | 1 |
| 53 | 1 | 1 | 0.821429 |
| 54 | 0 | 0 | 0.410684 |
| 55 | 1 | 1 | 0.869048 |
| 56 | 0 | 0 | 0.026482 |
| 57 | 0 | 0 | 0.046482 |
| 58 | 0 | 0 | 0.115768 |
| 59 | 0 | 0 | 0.076482 |
| 60 | 0 | 0 | 0.209816 |
| 61 | 0 | 0 | 0.049091 |
| 62 | 0 | 0 | 0.051482 |
| 63 | 1 | 1 | 0.975 |
| 64 | 0 | 0 | 0.028571 |
| 65 | 0 | 0 | 0.112033 |
| 66 | 0 | 0 | 0.023377 |
| 67 | 1 | 0 | 0.390768 |
| 68 | 1 | 1 | 0.724603 |
| 69 | 0 | 0 | 0.009091 |
| 70 | 0 | 0 | 0.1 |
| 71 | 0 | 0 | 0.042424 |
| 72 | 1 | 1 | 0.626482 |
| 73 | 0 | 0 | 0.009091 |
| 74 | 0 | 0 | 0 |
| 75 | 0 | 0 | 0.009091 |
| 76 | 0 | 0 | 0.090768 |

| | | | |
|---|---|---|---|
| 77 | 0 | 0 | 0.126482 |
| 78 | 0 | 0 | 0.009091 |
| 79 | 0 | 0 | 0.026482 |
| 80 | 0 | 0 | 0.009091 |
| 81 | 0 | 0 | 0.026482 |
| 82 | 0 | 0 | 0.076482 |
| 83 | 0 | 0 | 0.101482 |
| 84 | 1 | 0 | 0.458462 |
| 85 | 0 | 0 | 0.009091 |
| 86 | 1 | 1 | 0.83489 |
| 87 | 0 | 0 | 0.375366 |
| 88 | 0 | 0 | 0.009091 |
| 89 | 1 | 1 | 0.871429 |
| 90 | 0 | 0 | 0.125128 |
| 91 | 0 | 0 | 0.12 |
| 92 | 0 | 0 | 0.009091 |
| 93 | 0 | 0 | 0.166482 |
| 94 | 1 | 1 | 0.971429 |
| 95 | 0 | 0 | 0.009091 |
| 96 | 1 | 1 | 1 |
| 97 | 1 | 1 | 0.694805 |
| 98 | 0 | 0 | 0 |
| 99 | 0 | 0 | 0.009091 |

对分类模型的输出数据进行直方图计算，得到如下表 3 和表 4，表 3 为正样本对应的第一直方图结果，表 4 为负样本对应的第二直方图结果。

表 3 为正样本的第一直方图结果

| 概率区间 | 概率区间内正样本数量 |
|---|---|
| [0, 0.04) | 0 |
| [0.04, 0.08) | 0 |
| [0.08, 0.12) | 0 |
| [0.12, 0.16) | 0 |
| [0.16, 0.2) | 0 |

| | |
|---|---|
| [0.2, 0.24) | 0 |
| [0.24, 0.28) | 0 |
| [0.28, 0.32) | 0 |
| [0.32, 0.36) | 0 |
| [0.36, 0.4) | 1 |
| [0.4, 0.44) | 0 |
| [0.44, 0.48) | 2 |
| [0.48, 0.52) | 0 |
| [0.52, 0.56) | 1 |
| [0.56, 0.6) | 0 |
| [0.6, 0.64) | 2 |
| [0.64, 0.68) | 3 |
| [0.68, 0.72) | 2 |
| [0.72, 0.76) | 3 |
| [0.76, 0.8) | 0 |
| [0.8, 0.84) | 2 |
| [0.84, 0.88) | 3 |
| [0.88, 0.92) | 1 |
| [0.92, 0.96) | 0 |
| [0.96, 1) | 5 |

表 4 为负样本的第二直方图

| 概率区间 | 概率区间内负样本数量 |
|---|---|
| [0, 0.04) | 34 |
| [0.04, 0.08) | 13 |
| [0.08, 0.12) | 10 |
| [0.12, 0.16) | 5 |
| [0.16, 0.2) | 3 |
| [0.2, 0.24) | 3 |
| [0.24, 0.28) | 4 |
| [0.28, 0.32) | 1 |
| [0.32, 0.36) | 0 |

| | |
|---|---|
| [0.36, 0.4) | 1 |
| [0.4, 0.44) | 1 |
| [0.44, 0.48) | 0 |
| [0.48, 0.52) | 0 |
| [0.52, 0.56) | 0 |
| [0.56, 0.6) | 0 |
| [0.6, 0.64) | 0 |
| [0.64, 0.68) | 0 |
| [0.68, 0.72) | 0 |
| [0.72, 0.76) | 0 |
| [0.76, 0.8) | 0 |
| [0.8, 0.84) | 0 |
| [0.84, 0.88) | 0 |
| [0.88, 0.92) | 0 |
| [0.92, 0.96) | 0 |
| [0.96, 1) | 0 |

在获取到第一直方图和第二直方图的结果后，可以获取到概率区间，将每个概率区间的下限制作为阈值点构成阈值集。该示例中阈值集为：0、0.04、0.08、0.12、0.16、0.2、0.24、0.28、0.32、0.36、0.4、0.44、0.48、0.52、0.56、0.6、0.64、0.68、0.72、0.76、0.8、0.84、0.88、0.92、0.96

此处仅以两个阈值点作为示例说明阈值点对应评估参数的计算过程：

当阈值点选择为 0.4 时，根据第一直方图和第二直方图可以获取阈值点为 0.4 时的混淆矩阵：TP=24，FP=1，FN=1，TN=74。

当阈值点选择为 0.6 时，根据第一直方图结果和第二直方图结果可以获取阈值点为 0.6 时的混淆矩阵：TP=21，FP=4，FN=0，TN=75。

对于每个阈值点，根据混淆矩阵可以计算出对应的 ROC 坐标和 Lift 坐标。

ROC 坐标：横坐标 X=FP/(FP+TN)；纵坐标 Y=TP/(TP+FN)。Lift 坐标：横坐标 X=（TP+FN）/样本总量；纵坐标 Y=TP。在获取到所有的阈值点对应的 ROC 坐标和 Lift 坐标后，就可以描点绘制 ROC 曲线以及 Lift 图。图 3 为分类模型的 ROC 曲线，图 3 中 ROC 曲线的纵坐标为击中率 TPR（True Positive Rate），击中率可用于指示出分类模型识别出正样本的灵敏度（Sensitivity）。TPR=TP/(TP+FN)；横坐标为假正率 FPR（False

Positive Rate），其中，FPR=FP/(FP+TN)。其中，假正率可以通过特异率（Spcificity 表示，假正率=1-Spcificity，特异率为负例的覆盖率（True Negative Rate，TNR）TNR=TN /（TN + FP）。

图 4 为分类模型的 Lift 图，图 4 中纵坐标为实际正样本的数量，横坐标为正样本预测比例=（TP+FN）/样本总量。

在获取到每个阈值点对应的 ROC 坐标后，可以绘制出 ROC 曲线后，由相邻阈值点对应的 ROC 坐标与 ROC 曲线可以构成一个曲边梯形，根据相邻的 ROC 坐标能够计算一个曲边梯形的面积。在获取到所有的曲边梯形的面积后，将所有曲边梯形的面积相加得到 ROC 曲线对应的 AUC 值。

下面为计算评估参数的代码：

输入：N，icProb，icTrue，icFalse　#N 为概率区间的个数、icProb 概率区间的下限值、icTrue 概率区间内实际正样本的数量、icFalse 概率区间内实际负样本的数量#

输出：每个阈值点对应的 ROC 坐标，Lift 坐标，混淆矩阵，AUC 值；

计算过程：

1.计算总体正样本数量：totalTrue=∑（icTrue）；总体负样本数量：totalFalse=∑（icFalse）

2.初始化累计正负样本数量 curTrue=0，curFalse=0

3.For　i：　0 to N

a)阈值点 p=icProb[N-1-i]

b)curTrue+=icTrue[N-1-i];curFalse+=icFalse[N-1-i]#对实际正样本被预测成正样本数量进行累积得到 TP，对实际负样本被预测成正样本数量进行累积得到 FN#

c)混淆矩阵坐标：cm.p=p; cm.tp= curTrue, cm.fp= curFalse

cm.fn= totalTrue - curTrue, cm.tn= totalFalse – curFalse

d)ROC 坐标：roc.p=p;

roc.x=curFalse / totalFalse

roc.y= curTrue / totalTrue

e)Lift 坐标：lift.p=plift.x= (curTrue + curFalse) / (totalTrue+ totalFalse)

lift.y= curTrue

4.根据 ROC 坐标计算曲线下方的面积，即 AUC 值。

通过上述实施例可以看出，根据直方图计算结果计算得出的混淆矩阵，然后基于该混淆矩阵就可以方便的计算出其他评估指标，并生成可视化图像，用户可以直观地判断

分类模型的优良。

实施例三

如图 5 所示，其为本发明实施例三的评估指标获取装置的结构示意图。该评估指标获取装置包括：分类训练模块 11、概率统计模块 12 和计算模块 13。

分类训练模块 11，用于将样本输入分类模型进行分类训练，获取分类模型的输出数据。

为了对分类模型进行评估，分类训练模块 11 需要将样本输入分类模型中进行分类训练，在训练完成后，分类训练模块 11 会对每个样本进行分类和概率预测。具体地，分类训练模块 11 在训练完成后为每个样本输出训练后的样本属性，训练后的样本属性可以指示出样本经过分类模型后是正样本还是负样本。

进一步地，分类训练模块 11 在训练完成后还会为每个样本进行概率预测，用户可以根据实际需要选择输出每个样本经过分类模型预测成正样本的概率，或者选择输出每个样本经过分类模型预测成负样本的概率。其中，样本经过分类模型被预测成正样本的概率和被预测成负样本的概率的和为 1。

其中，输入的每个样本都有一个原始的样本属性。本实施例中，样本属性包括正样本属性和负样本属性。原始的样本属性表示样本实际是正样本还是负样本。

概率统计模块 12，用于对输出数据进行概率分布统计获取概率统计结果。

其中，概率统计结果包括概率区间以及每个概率区间内实际正样本数量和实际负样本数量。

在获取到输出数据后，由于分类训练模块 11 会对每个样本进行概率预测，这样输出数据中每个样本会有一个预测概率，本实施例中，分类训练模块 11 输出的每个样本的概率为每个样本被分类模型预测成正样本的预测概率。

进一步地，概率统计模块 12 根据预测概率对输出数据进行概率分布统计，获取概率统计结果。概率统计模块 12 在进行概率统计时首先需要划分概率区间，然后在每个概率区间内基于输出数据中每个样本原始的样本属性统计实际正样本数量和实际负样本数量，得到正样本和负样本的概率分布图，基于正样本的概率分布图获取每个概率区间内实际正样本数量，基于负样本的概率分布图获取每个概率区间内实际负样本数量。

优选地，概率统计模块 12 基于直方图算法对输出数据进行概率分布的统计，获取正样本的直方图和负样本的直方图，基于正样本的直方图和负样本的直方图能够获取到上述概率统计结果。

计算模块 13，用于根据阈值集和概率统计结果计算分类模型的评估指标。

在获取到概率统计结果后，需要获取阈值集，其中阈值集中包括多个阈值点，然后基于每个阈值点和概率统计结果中每个概率区间内实际正样本的第一数据和实际负样本的第二数据，获取每个阈值点对应的评估参数，利用所有阈值点对应的评估参数生成分类模型的评估指标。

本实施例中，在概率统计结果后，计算模块 13 可以将概率统计结果中的概率区间的端点值作为阈值点构成阈值集。例如，可以利用每个概率区间的下限值作为阈值点构成阈值集。或者将部分概率区间的下限值作为阈值点构成阈值集。在概率统计的过程中，对概率区间进行划分，本实施例中概率区间的端点可作为分界点，直接将概率区间的端点值作为阈值点，不需要进行阈值点的重新设定，进而提高了评估指标的计算效率。

可选地，计算模块 13 可以接收用户输入的利用概率区间端值点作为阈值点阈值集。例如，用户可以将每个概率区间的下限值作为阈值点构成阈值集，或者用户选取部分概率区间的下限值作为阈值点构成阈值集本实施例中，用户根据反馈的概率统计结果，可以初步对分类模型的效果有一定的了解，从而能够选取合适的阈值点构成阈值集，用户交互较好，而且对分类模型的评估更加准确。

进一步地，计算模块 13 根据阈值集中的阈值点和概率统计结果计算评估指标。其中，评估指标包括混淆矩阵、ROC 曲线、AUC 值和 Lift 图。

其中，混淆矩阵中包括：TP、FP、TN 和 FN。

在获取到阈值点之后，计算模块 13 将阈值点作为分界点，对于正样本的概率分布来说，大于阈值点的所有概率区间内实际正样本被分类模型预测成正样本，对实际正样本被分类模型预测成正样本的数量进行累积，将累积的实际正样本被分类模型预测成正样本的数量作为混淆矩阵的 TP。而小于阈值点的所有概率区间内实际正样本被分类模型预测成负样本，对实际正样本被分类模型预测成负样本的数量进行累计，将累计后的实际正样本被分类模型预测成负样本的数量作为混淆矩阵的 FP。

对于负样本的概率分布来说，大于阈值点的所有概率区间内实际负样本被分类模型预测成正样本，对实际负样本被分类模型预测成正样本的数量进行累积，将累积的实际负样本被分类模型预测成正样本的数量作为混淆矩阵的 FN。而小于阈值点的所有概率区间内实际负样本被分类模型预测成负样本，对实际负样本被分类模型预测成负样本的数量进行累计，将累计后的实际负样本被分类模型预测成负样本的数量作为混淆矩阵的 TN。

在获取到阈值点对应的混淆矩阵后，计算模块 13 可以利用混淆矩阵中的 TP、FP、TN 和 FN，计算得到其他评估指标的该阈值点对应的评估参数，当所有阈值点对应的评估参数计算完成后，利用每个阈值点对应的评估参数生成评估指标。例如，根据一个阈值点对应的混淆矩阵可以计算出在该阈值点处 ROC 曲线的坐标，将坐标作为该阈值点 ROC 曲线的评估参数。当所有阈值点对应的评估参数计算完成后，利用每个阈值点对应的 ROC 曲线的坐标绘制 ROC 曲线。

本实施例提供的评估指标获取装置，通过对分类模型的输出数据进行概率统计，基于得到的概率统计结果对评估指标进行计算，解决了在评估指标的计算过程中多次扫描输出数据的问题，尤其在输出数据为大规模数据时可以提高评估指标的计算效率。

实施例四

如图 6 所示，其为本发明实施例四的评估指标获取装置的结构示意图。该评估指标获取装置包括：分类训练模块 21、概率统计模块 22、计算模块 23 和可视化模块 24。

分类训练模块 21，用于将样本输入分类模型进行分类训练，获取分类模型的输出数据。

进一步地，概率统计模块 22，具体用于直方图计算单元 221，用于基于直方图算法对输出数据进行概率区间划分，统计每个概率区间内实际正样本数量和实际负样本数量。

其中，输出数据包括：每个样本原始的样本属性以及每个样本被分类模型预测成正样本的预测概率；其中，样本属性包括正样本属性和负样本属性。

进一步地，概率统计模块 22 一种可选的结构方式包括：扫描单元 221、直方图生成单元 222、步长调整单元 223 和统计单元 224。

扫描单元 221，用于扫描输出数据。

直方图生成单元 222，用于根据每个样本被预测成正样本的预测概率和输出数据中每个样本原始的样本属性生成正样本对应的第一直方图和负样本对应的第二直方图；其中，第一直方图的横轴是预测概率，第一直方图的纵轴是实际正样本数量；第二直方图的横轴是预测概率，第二直方图的纵轴是实际负样本数量。

步长调整单元 223，用于调整横轴步长使第一直方图和第二直方图的概率区间一致，以获取概率统计结果中的概率区间。

统计单元 224，用于统计第一直方图中每个概率区间内实际正样本的数量，以及统计第二直方图中每个概率区间内实际负样本的数量。

本实施例中，计算模块 23 一种可选的结构方式包括：阈值集获取单元 231、混淆矩

阵生成单元 232 和评估指标生成单元 233。

阈值集获取单元 231，用于将每个概率区间的端点值作为阈值点构成阈值集。

进一步地，阈值集获取单元 231，还用于接收用户输入的根据概率区间的端点值构成的阈值集。

混淆矩阵生成单元 232，用于按照由大到小的顺序获取阈值集中每个阈值点对应的混淆矩阵，其中，混淆矩阵包括 TP、FP、TN、FN。

评估指标生成单元 233，用于将每个阈值点对应的混淆矩阵作为分类模块的评估指标。

在获取到每个阈值点的混淆矩阵后，根据混淆矩阵可以获取到分类模型其他的评估指标，例如 ROC 曲线、ROC 曲线下面积 AUC 值以及 Lift 图。

进一步地，混淆矩阵生成单元 232，具体用于对于第一直方图，按照阈值点的大小顺序逐次对大于阈值点的所有概率区间内实际正样本数量进行累积得到 TP，以及对小于阈值点的所有概率区间内实际正样本数量进行累积得到 FN，以及对于第二直方图，按照阈值点的大小顺序逐次对大于阈值点的所有概率区间内负样本数量进行累积得到 FP，以及对小于阈值点的所有概率区间内负样本数量进行累积得到 TN。

评估指标生成单元 233，具体用于将每个阈值对应的混淆矩阵作为评估指标。

评估指标生成单元 233，具体用于针对每一个阈值点，将 FP 与实际负样本总量的比值作为 ROC 的横坐标，以及将 TP 与实际正样本总量的比值作为 ROC 的纵坐标，以及利用所有阈值点对应的 ROC 坐标绘制分类模型的评估指标 ROC 曲线。

评估指标生成单元 233，具体用于获取每个由相邻阈值点对应的 ROC 坐标与所述 ROC 曲线构成的曲边梯形的面积，将所有曲边梯形的面积相加得到所述 ROC 曲线的 AUC 值。

评估指标生成单元 233，具体用于针对每一个阈值点将 TP 和 FP 的和值与样本总量的比值作为 Lift 图的横坐标，以及将 TP 作为 Lift 图的纵坐标以及利用所有阈值点对应的 Lift 坐标绘制分类模型的评估指标 Lift 图。

可视化模块 24，用于接收用户的显示指令，根据显示指令将评估指标进行可视化展示。

本实施例中，评估指标获取装置可以设置在服务器上执行该评估指标获取方法，在计算出评估指标，用户可以向该装置中的可视化模块 24 发送显示指令，在接收到显示指令，可视化模块 24 可以将评估指标下发给本地终端，这样本地终端通过显示屏将评估指

标进行可视化展示，如向用户展示 ROC 曲线、Lift 图等。当用户点击 ROC 曲线上的点时，可以将该点对应的混淆矩阵进行展示。

可选地，对于大规模数据，评估指标获取装置中分类训练模块 21 和概率统计模块 22 可以设置在服务器上，而将计算模块 23 和可视化模块 24 设置在本地终端上，以减少服务器的压力，且便于与用户的交互。在服务器上对样本数据进行分类训练以及直方图计算，在计算完直方图后，概率统计模块 22 可以将直方图结果下发到本地终端的计算模块 23 中，计算模块 23 在本地终端上计算评估指标，这样可以减缓服务器的压力。在计算出评估指标后，用户可以向可视化模块 24 发送显示指令，在接收到显示指令后，可视化模块 24 通过显示屏将评估指标进行可视化展示，如向用户展示 ROC 曲线、Lift 图等。当用户点击 ROC 曲线上的点时，可以将该点对应的混淆矩阵进行展示。

可选地，评估指标获取装置可以设置在本地终端上执行该评估指标获取方法，在计算出评估指标后，用户可以向可视化模块 24 发送显示指令，在接收到显示指令后，可视化模块 24 在显示屏上进行可视化展示，如向用户展示 ROC 曲线、Lift 图等。当用户点击 ROC 曲线上的点时，可以将该点对应的混淆矩阵进行展示。

本实施例提供的评估指标获取装置，对分类模型的输出数据进行概率统计，基于得到包括概率区间以及每个概率区间内实际正样本数量和实际负样本数量的概率统计结果对评估指标进行计算，解决了在评估指标的计算过程中多次扫描输出数据的问题，尤其在输出数据为大规模数据时可以提高评估指标的计算效率。进一步地，在获取到评估指标后，能够将评估指标可视化展示，使用户能够直观地判断分类模型的优良情况。

本领域普通技术人员可以理解：实现上述各方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成。前述的程序可以存储于一计算机可读取存储介质中。该程序在执行时，执行包括上述各方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

最后应说明的是：以上各实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述各实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分或者全部技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的范围。

# 权 利 要 求 书

1、一种评估指标获取方法，其特征在于，包括：

将样本输入分类模型进行分类训练，获取分类模型的输出数据；

对所述输出数据进行概率分布统计获取概率统计结果；其中，所述概率统计结果包括概率区间以及每个概率区间内实际正样本数量和实际负样本数量；

根据阈值集和所述概率统计结果计算所述分类模型的评估指标。

2、根据权利要求 1 所述的评估指标获取方法，其特征在于，所述对所述输出数据进行概率分布统计获取概率统计结果，包括：

基于直方图算法对所述输出数据进行概率区间划分，统计每个概率区间内所述实际正样本数量和所述实际负样本数量。

3、根据权利要求 2 所述的评估指标获取方法，其特征在于，所述输出数据包括：每个样本原始的样本属性以及每个样本被所述分类模型预测成正样本的预测概率；其中，样本属性包括正样本属性和负样本属性。

4、根据权利要求 3 所述的评估指标获取方法，其特征在于，所述基于直方图算法对所述输出数据进行概率区间划分，统计每个概率区间内所述实际正样本数量和所述实际负样本数量，包括：

扫描所述输出数据；

根据每个样本被预测成正样本的预测概率和所述输出数据中每个样本原始的样本属性生成正样本对应的第一直方图和负样本对应的第二直方图；其中，所述第一直方图的横轴是预测概率，所述第一直方图的纵轴是实际正样本数量；所述第二直方图的横轴是预测概率，所述第二直方图的纵轴是实际负样本数量；

调整横轴步长使所述第一直方图和所述第二直方图的概率区间一致，以获取所述概率统计结果中的所述概率区间；

统计所述第一直方图中每个概率区间内所述实际正样本的数量；

统计所述第二直方图中每个概率区间内所述实际负样本的数量。

5、根据权利要求 4 所述的评估指标获取方法，其特征在于，所述根据阈值集和所述概率统计结果计算所述分类模型的评估指标，包括：

将每个概率区间的端点值作为阈值点构成所述阈值集；

按照由大到小的顺序获取所述阈值集中每个阈值点对应的混淆矩阵，其中，所述混淆矩阵包括实际为正样本被预测为正样本的数量 TP、实际为正样本被预测为负样本的

数量 FP、实际为负样本被预测为负样本的数量 TN、实际为负样本被预测为正样本的数量 FN；

将每个阈值点对应的混淆矩阵作为评估指标。

6、根据权利要求 4 所述的评估指标获取方法，其特征在于，所述根据阈值集和所述概率统计结果计算所述分类模型的评估指标，包括：

接收用户输入的根据概率区间的端点值构成的所述阈值集；

按照由大到小的顺序获取所述阈值集中每个阈值点对应的混淆矩阵，其中，所述混淆矩阵包括：TP、FP、TN 和 FN；

将每个阈值点对应的混淆矩阵作为所述评估指标。

7、根据权利要求 5 或 6 所述的评估指标获取方法，其特征在于，所述按照由大到小的顺序获取所述阈值集中每个阈值点对应的混淆矩阵，包括：

对于所述第一直方图，按照阈值点的大小顺序逐次对大于阈值点的所有概率区间内实际正样本数量进行累积得到所述 TP，以及对小于阈值点的所有概率区间内实际正样本数量进行累积得到所述 FN；

对于所述第二直方图，按照阈值点的大小顺序逐次对大于阈值点的所有概率区间内负样本数量进行累积得到所述 FP，以及对小于阈值点的所有概率区间内负样本数量进行累积得到所述 TN。

8、根据权利要求 7 所述的评估指标获取方法，其特征在于，所述按照由大到小的顺序获取所述阈值集中每个阈值点对应的混淆矩阵之后，还包括：

针对每个阈值点，将所述 FP 与实际负样本总量的比值作为所述 ROC 的横坐标；

将所述 TP 与实际正样本总量的比值作为所述 ROC 的纵坐标；

利用所有阈值点对应的 ROC 坐标绘制所述分类模型的评估指标 ROC 曲线。

9、根据权利要求 8 所述的评估指标获取方法，其特征在于，所述利用所有阈值点对应的 ROC 坐标绘制所述分类模型的评估指标 ROC 曲线之后，还包括：

获取每个由相邻阈值点对应的 ROC 坐标与所述 ROC 曲线构成的曲边梯形的面积；

将所有曲边梯形的面积相加得到所述 ROC 曲线对应的 AUC 值。

10、根据权利要求 7 所述的评估指标获取方法，其特征在于，所述按照由大到小的顺序获取所述阈值集中每个阈值点对应的混淆矩阵之后，还包括：

针对每个阈值点，将所述 TP 和所述 FP 的和值与样本总量的比值作为 Lift 图的横坐标；

将所述 TP 作为 Lift 图的纵坐标；

利用所有阈值点对应的 Lift 坐标绘制所述分类模型的评估指标 Lift 图。

11、一种评估指标获取装置，其特征在于，包括：

分类训练模块，用于将样本输入分类模型进行分类训练，获取分类模型的输出数据；

概率统计模块，用于对所述输出数据进行概率分布统计获取概率统计结果；其中，所述概率统计结果包括概率区间以及每个概率区间内实际正样本数量和实际负样本数量；

计算模块，用于根据阈值集和所述概率统计结果计算所述分类模型的评估指标。

12、根据权利要求 11 所述的评估指标获取装置，其特征在于，所述概率统计模块，具体用于基于直方图算法对所述输出数据进行概率区间划分，统计每个概率区间内所述实际正样本数量和所述实际负样本数量。

13、根据权利要求 12 所述的评估指标获取装置，其特征在于，所述输出数据包括：每个样本原始的样本属性以及每个样本被所述分类模型预测成正样本的预测概率；其中，样本属性包括正样本属性和负样本属性。

14、根据权利要求 13 所述的评估指标获取装置，其特征在于，所述概率统计模块，包括：

扫描单元，用于扫描所述输出数据；

直方图生成单元，用于根据每个样本被预测成正样本的预测概率和所述输出数据中每个样本原始的样本属性生成正样本对应的第一直方图和负样本对应的第二直方图；其中，所述第一直方图的横轴是预测概率，所述第一直方图的纵轴是实际正样本数量；所述第二直方图的横轴是预测概率，所述第二直方图的纵轴是实际负样本数量；

步长调整单元，用于调整横轴步长使所述第一直方图和所述第二直方图的概率区间一致，以获取所述概率统计结果中的所述概率区间；

统计单元，用于统计所述第一直方图中每个概率区间内所述实际正样本的数量，以及统计所述第二直方图中每个概率区间内所述实际负样本的数量。

15、根据权利要求 14 所述的评估指标获取装置，其特征在于，所述计算模块，包括：

阈值集获取单元，用于将每个概率区间的端点值作为阈值点生成所述阈值集；

混淆矩阵生成单元，用于按照由大到小的顺序获取所述阈值集中每个阈值点对应的混淆矩阵，其中，所述混淆矩阵包括实际为正样本被预测为正样本的数量 TP、实际为

正样本被预测为负样本的数量 FP、实际为负样本被预测为负样本的数量 TN、实际为负样本被预测为正样本的数量 FN；

评估指标生成单元，用于将每个阈值点对应的混淆矩阵作为所述评估指标。

16、根据权利要求 15 所述的评估指标获取装置，其特征在于，所述阈值集获取单元，还用于接收用户输入的根据概率区间的端点值构成所述阈值集。

17、根据权利要求 16 所述的评估指标获取装置，其特征在于，所述混淆矩阵生成单元，具体用于对于所述第一直方图，按照阈值点的大小顺序逐次对大于阈值点的所有概率区间内实际正样本数量进行累积得到所述 TP，以及对小于阈值点的所有概率区间内实际正样本数量进行累积得到所述 FN，以及对于所述第二直方图，按照阈值点的大小顺序逐次对大于阈值点的所有概率区间内负样本数量进行累积得到所述 FP，以及对小于阈值点的所有概率区间内负样本数量进行累积得到所述 TN。

18、根据权利要求 17 所述的评估指标获取装置，其特征在于，所述评估指标生成单元，具体用于针对每个阈值点，将所述 FP 与实际负样本总量的比值作为所述 ROC 的横坐标，以及将所述 TP 与实际正样本总量的比值作为所述 ROC 的纵坐标，以及利用所有阈值点对应的 ROC 坐标绘制所述分类模型的评估指标 ROC 曲线。

19、根据权利要求 18 所述的评估指标获取装置，其特征在于，所述评估指标生成单元，还具体用于获取每个由相邻阈值点对应的 ROC 坐标与所述 ROC 曲线构成的曲边梯形的面积，将所有曲边梯形的面积相加得到所述 ROC 曲线的 AUC 值。

20、根据权利要求 19 所述的评估指标获取装置，其特征在于，所述评估指标生成单元，具体用于针对每个阈值点，将所述 TP 和所述 FP 的和值与样本总量的比值作为 Lift 图的横坐标，以及将所述 TP 作为 Lift 图的纵坐标以及利用所有阈值点对应的 Lift 坐标绘制所述分类模型的评估指标 Lift 图。

21、根据权利要求 20 所述的评估指标获取装置，其特征在于，所述分类训练模块和所述概率统计模块设置于服务器端，所述计算模块设置于本地终端。

22、根据权利要求 21 所述的评估指标获取装置，其特征在于，还包括：可视化模块，用于接收用户的显示指令，根据显示指令将所述分类模型的评估指标进行可视化展示；

其中，所述可视化模块设置于所述本地终端。

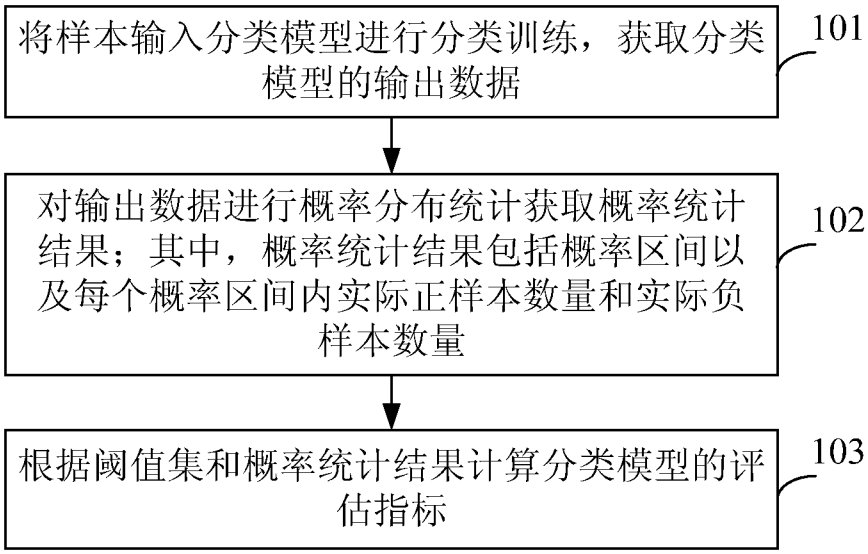

将样本输入分类模型进行分类训练，获取分类模型的输出数据
101
对输出数据进行概率分布统计获取概率统计结果；其中，概率统计结果包括概率区间以及每个概率区间内实际正样本数量和实际负样本数量
102
根据阈值集和概率统计结果计算分类模型的评估指标
103

图 1

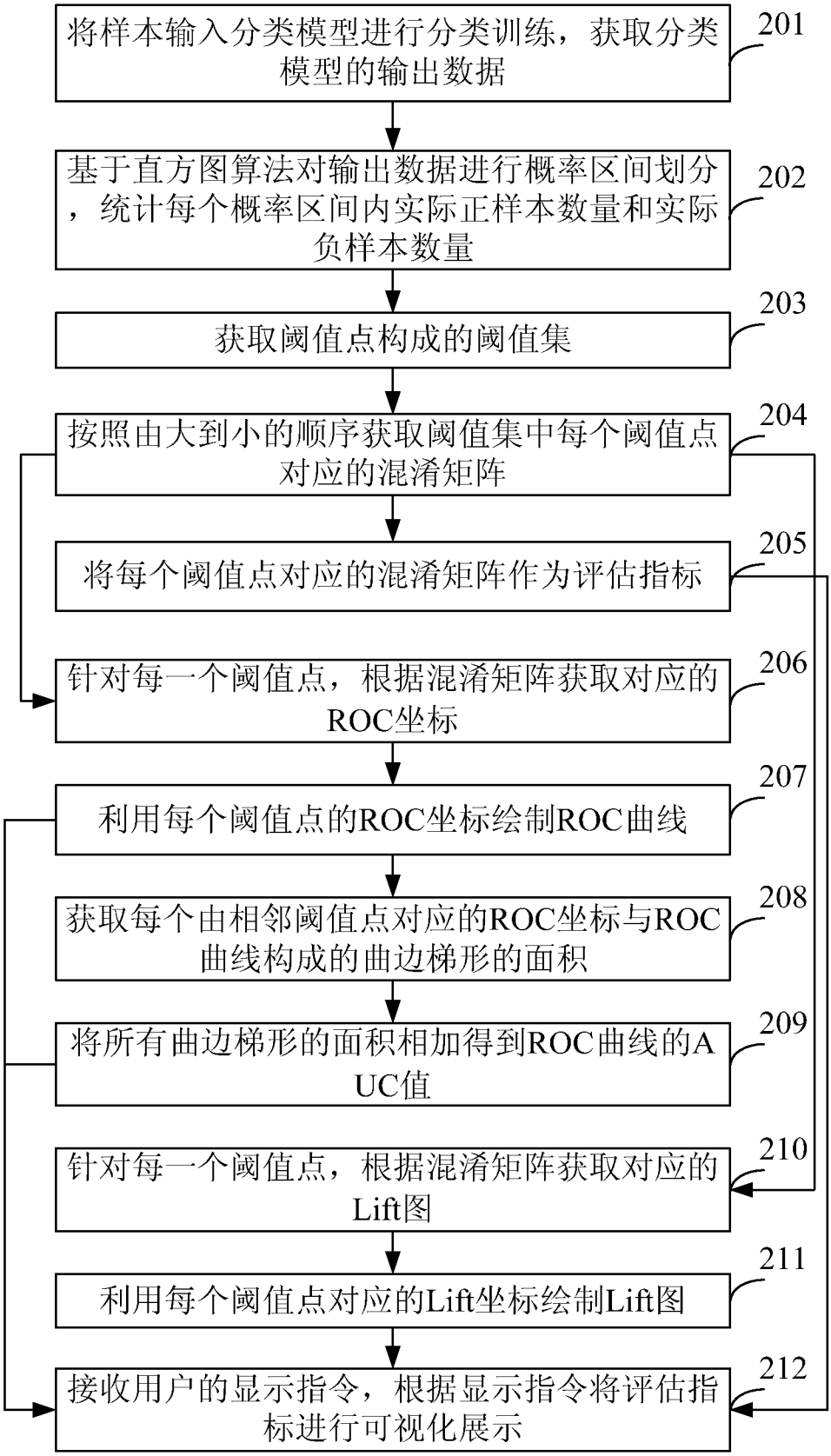

将样本输入分类模型进行分类训练，获取分类模型的输出数据
201
基于直方图算法对输出数据进行概率区间划分，统计每个概率区间内实际正样本数量和实际负样本数量
202
获取阈值点构成的阈值集
203
按照由大到小的顺序获取阈值集中每个阈值点对应的混淆矩阵
204
将每个阈值点对应的混淆矩阵作为评估指标
205
针对每一个阈值点，根据混淆矩阵获取对应的ROC坐标
206
利用每个阈值点的ROC坐标绘制ROC曲线
207
获取每个由相邻阈值点对应的ROC坐标与ROC曲线构成的曲边梯形的面积
208
将所有曲边梯形的面积相加得到ROC曲线的AUC值
209
针对每一个阈值点，根据混淆矩阵获取对应的Lift图
210
利用每个阈值点对应的Lift坐标绘制Lift图
211
接收用户的显示指令，根据显示指令将评估指标进行可视化展示
212

图 2

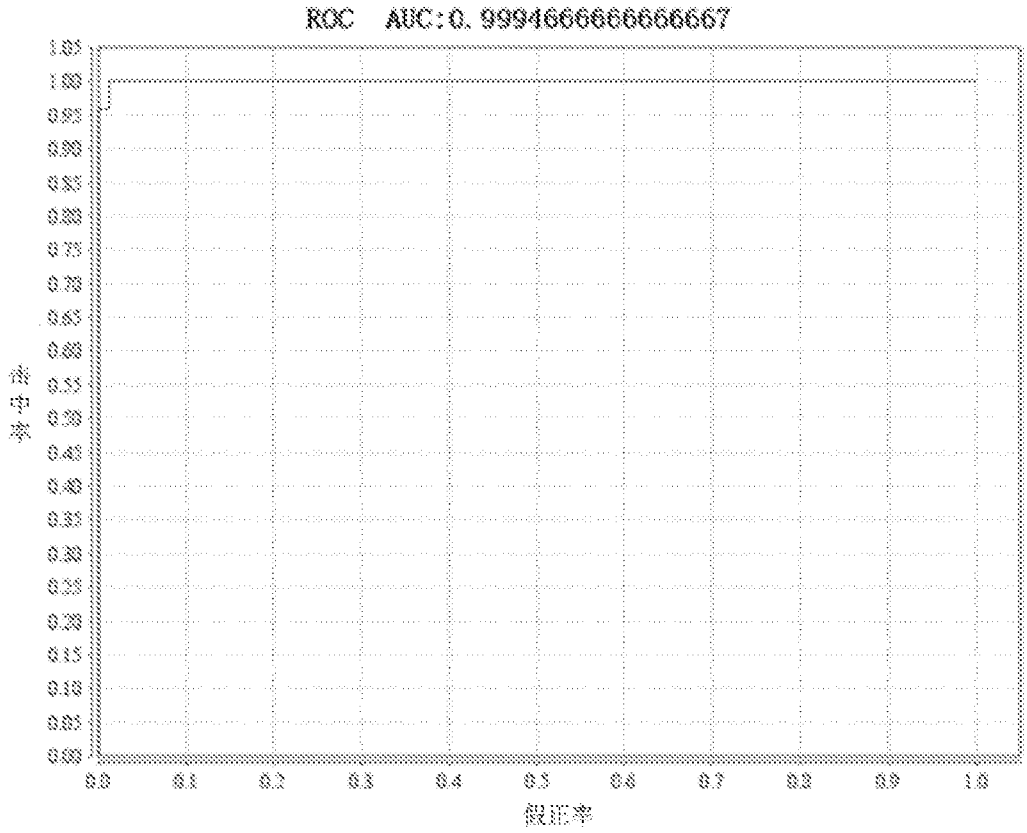

ROC AUC:0.9994666666666667
真正率
假正率
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

图 3

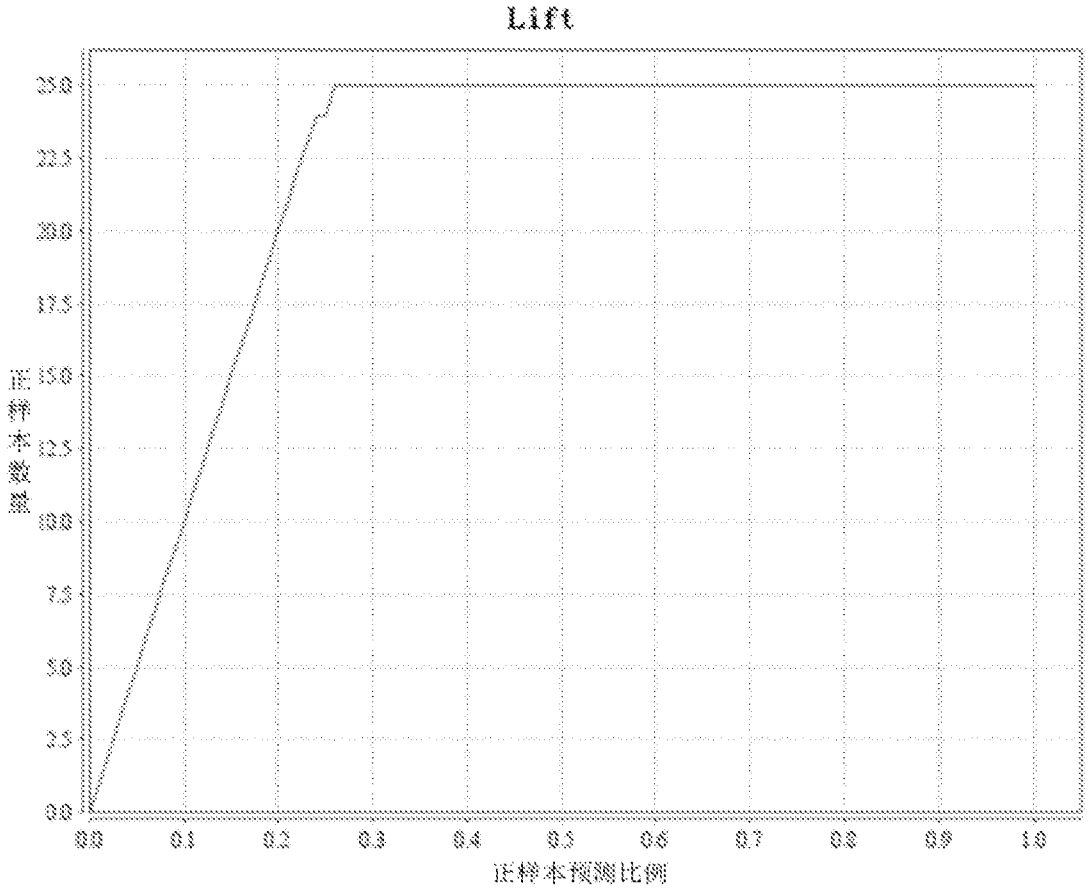

Lift
正样本数累
正样本预测比例
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

图 4


分类训练模块
11
概率统计模块
12
计算模块
13

图 5

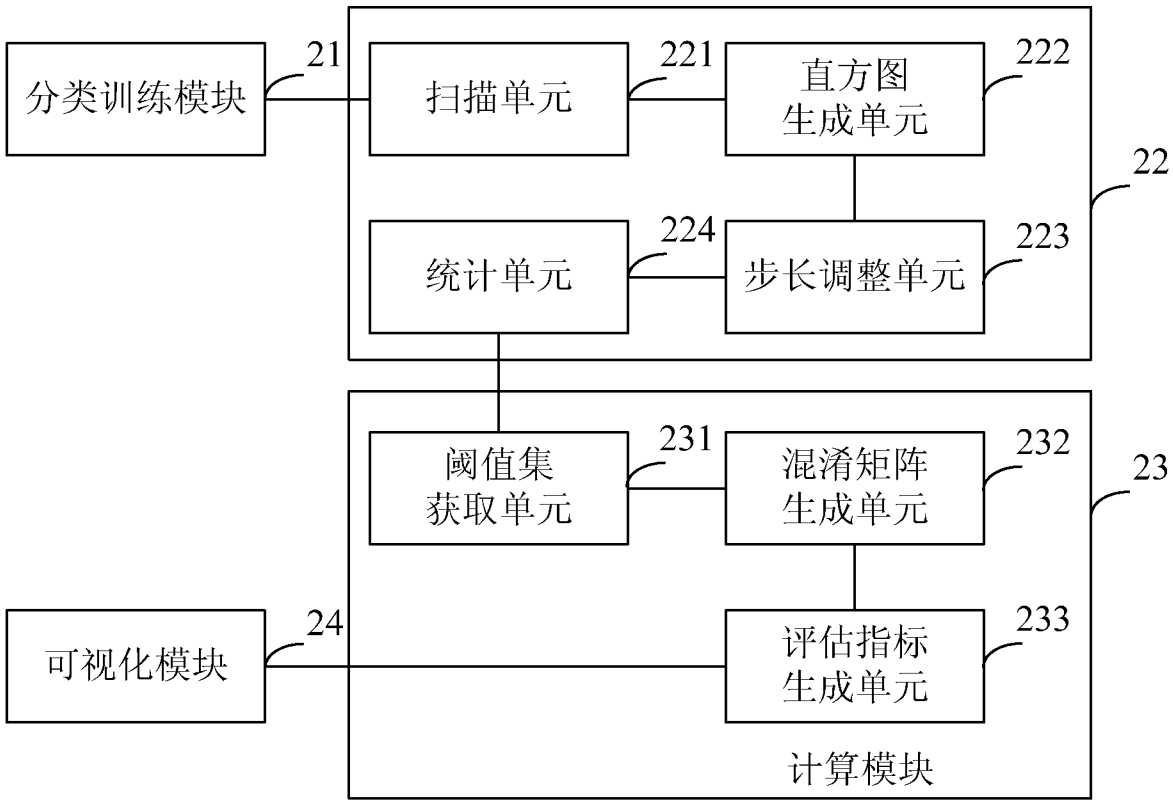

分类训练模块
21
扫描单元
221
直方图
生成单元
222
22
统计单元
224
步长调整单元
223
阈值集
获取单元
231
混淆矩阵
生成单元
232
23
可视化模块
24
评估指标
生成单元
233
计算模块

图 6

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/072405

**A. CLASSIFICATION OF SUBJECT MATTER**

G06F 17/30 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC: classify, model, sample, negative, positive, probability, statistic, zone, threshold, estimate, evaluate, curve, area, scan, arithmetic, confusion, matrix

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | CN 104504583 A (GUANGZHOU VIP.COM NETWORK TECHNOLOGY CO., LTD.) 08 April 2015 (08.04.2015) description, paragraphs [0051], [0052], [0055]-[0057], and [0060]-[0090] | 1-22 |
| A | CN 102663723 A (WUHAN UNIVERSITY) 12 September 2012 (12.09.2012) the whole document | 1-22 |
| A | CN 104361224 A (SHENZHEN INSTITUTE OF INFORMATION TECHNOLOGY) 18 February 2015 (18.02.2015) the whole document | 1-22 |
| A | CN 105069470 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 18 November 2015 (18.11.2015) the whole document | 1-22 |
| A | US 2011137841 A1 (FUJITSU LIMITED) 09 June 2011 (09.06.2011) the whole document | 1-22 |

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&"document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 28 March 2017 | 04 May 2017 |
| Name and mailing address of the ISA<br>State Intellectual Property Office of the P. R. China<br>No. 6, Xitucheng Road, Jimenqiao<br>Haidian District, Beijing 100088, China<br>Facsimile No. (86-10) 62019451 | Authorized officer<br>JIANG, Lingling<br>Telephone No. (86-10) 61648096 |

# INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/072405

| Patent Documents referred in the Report | Publication Date | Patent Family | Publication Date |
|---|---|---|---|
| CN 104504583 A | 08 April 2015 | None | |
| CN 102663723 A | 12 September 2012 | None | |
| CN 104361224 A | 18 February 2015 | None | |
| CN 105069470 A | 18 November 2015 | None | |
| US 2011137841 A1 | 09 June 2011 | WO 2010016109 A1 | 11 February 2010 |
| | | EP 2323080 A1 | 18 May 2011 |
| | | JP 5141767 B2 | 13 February 2013 |
| | | JPW 02010016109 A1 | 12 January 2012 |

# 国际检索报告

国际申请号 PCT/CN2017/072405

A. 主题的分类

G06F 17/30(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNPAT, CNKI, WPI, EPODOC: 分类, 模型, 正样本, 负样本, 概率, 统计, 区间, 阈值, 评估, 评价, 混淆矩阵, 曲线, 面积, 算法, 扫描, classify, model, sample, negative, positive, probability, statistic, zone, threshold, estimate, evaluate, curve, area, scan, arithmetic, cunfusion, matrix

C. 相关文件

| 类 型* | 引用文件，必要时，指明相关段落 | 相关的权利要求 |
|---|---|---|
| X | CN 104504583 A (广州唯品会网络技术有限公司) 2015年 4月 8日 (2015-04-08)<br>说明书第51-52、55-57、60-90段 | 1-22 |
| A | CN 102663723 A (武汉大学) 2012年 9月 12日 (2012-09-12)<br>全文 | 1-22 |
| A | CN 104361224 A (深圳信息职业技术学院) 2015年 2月 18日 (2015-02-18)<br>全文 | 1-22 |
| A | CN 105069470 A (腾讯科技深圳有限公司) 2015年 11月 18日 (2015-11-18)<br>全文 | 1-22 |
| A | US 2011137841 A1 (FUJITSU LIMITED) 2011年 6月 9日 (2011-06-09)<br>全文 | 1-22 |

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

| 国际检索实际完成的日期 | 国际检索报告邮寄日期 |
|---|---|
| 2017年 3月 28日 | 2017年 5月 4日 |
| ISA/CN的名称和邮寄地址<br>中华人民共和国国家知识产权局(ISA/CN)<br>中国北京市海淀区蓟门桥西土城路6号 100088<br>传真号 (86-10)62019451 | 受权官员<br>姜玲玲<br>电话号码 (86-10)61648096 |

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/072405

| 检索报告引用的专利文件 | | | 公布日<br>(年/月/日) | 同族专利 | | | 公布日<br>(年/月/日) |
|---|---|---|---|---|---|---|---|
| CN | 104504583 | A | 2015年 4月 8日 | 无 | | | |
| CN | 102663723 | A | 2012年 9月 12日 | 无 | | | |
| CN | 104361224 | A | 2015年 2月 18日 | 无 | | | |
| CN | 105069470 | A | 2015年 11月 18日 | 无 | | | |
| US | 2011137841 | A1 | 2011年 6月 9日 | WO | 2010016109 | A1 | 2010年 2月 11日 |
| | | | | EP | 2323080 | A1 | 2011年 5月 18日 |
| | | | | JP | 5141767 | B2 | 2013年 2月 13日 |
| | | | | JPW | 02010016109 | A1 | 2012年 1月 12日 |