



(12)发明专利

(10)授权公告号 CN 107181776 B

(45)授权公告日 2020.04.28

(21)申请号 201610136091.0

(22)申请日 2016.03.10

(65)同一申请的已公布的文献号
申请公布号 CN 107181776 A

(43)申请公布日 2017.09.19

(73)专利权人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 谭卫国 汪芳山

(74)专利代理机构 广州三环专利商标代理有限公司 44202

代理人 郝传鑫 熊永强

(51)Int.Cl.

H04L 29/08(2006.01)

H04L 12/26(2006.01)

(56)对比文件

US 2014143446 A1,2014.05.22,

US 7761425 B1,2010.07.20,

CN 101454781 A,2009.06.10,

CN 104794239 A,2015.07.22,

CN 102915347 A,2013.02.06,

US 8781442 B1,2014.07.15,

审查员 李亢亢

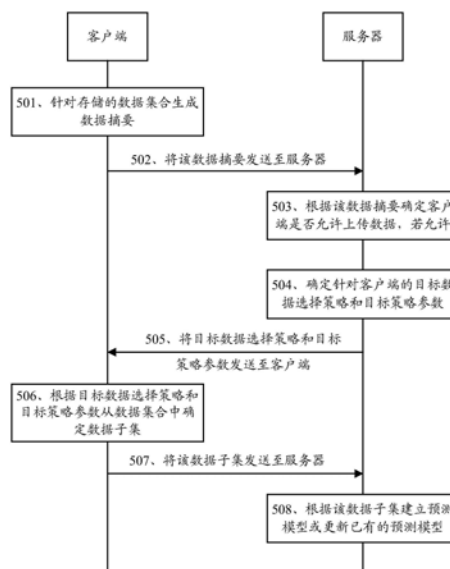
权利要求书7页 说明书25页 附图11页

(54)发明名称

一种数据处理方法及相关设备、系统

(57)摘要

本发明实施例公开了一种数据处理方法及相关设备、系统,其中,该方法包括:客户端针对存储的数据集合生成数据摘要,并将该数据摘要发送至服务器,服务器根据该数据摘要确定针对客户端的反馈信息,并将该反馈信息发送给客户端,当客户端接收到的反馈信息中包括目标数据选择策略和目标策略参数时,客户端可以根据该目标数据选择策略和该目标策略参数从数据集合中确定出数据子集,并将该数据子集发送至服务器,服务器可以根据该数据子集建立预测模型或者更新已有的预测模型。实施本发明实施例,可以控制客户端向服务器传输的数据量,有效降低数据处理过程中的资源消耗;此外,能够精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型。



1. 一种数据处理方法,其特征在于,包括:

客户端针对存储的数据集合生成数据摘要,并将所述数据摘要发送至服务器,所述数据摘要用于触发所述服务器以所述数据摘要为依据确定针对所述客户端的反馈信息;

所述客户端接收所述服务器发送的所述反馈信息;

当所述反馈信息包括目标数据选择策略和目标策略参数时,所述客户端根据所述目标数据选择策略和所述目标策略参数从所述数据集合中确定数据子集,其中,所述目标策略参数用于限定所述数据子集中数据的数目;

所述客户端将所述数据子集发送至所述服务器。

2. 根据权利要求1所述的方法,其特征在于,当所述反馈信息包括用于指示所述客户端禁止上传数据的信息时,所述方法还包括:

所述客户端响应所述反馈信息,禁止向所述服务器发送数据。

3. 根据权利要求1所述的方法,其特征在于,所述方法还包括:

所述客户端计算所述数据摘要与至少一个历史数据摘要的相似度;

所述客户端判断所述计算出的相似度中是否存在相似度超过预设阈值;

当所述计算出的相似度中不存在相似度超过所述预设阈值时,所述客户端执行所述的将所述数据摘要发送至服务器的步骤。

4. 根据权利要求2所述的方法,其特征在于,所述方法还包括:

所述客户端计算所述数据摘要与至少一个历史数据摘要的相似度;

所述客户端判断所述计算出的相似度中是否存在相似度超过预设阈值;

当所述计算出的相似度中不存在相似度超过所述预设阈值时,所述客户端执行所述的将所述数据摘要发送至服务器的步骤。

5. 根据权利要求1所述的方法,其特征在于,当所述数据集合中包括多种类型的数据时,所述客户端针对存储的数据集合生成数据摘要,包括:

所述客户端针对每种类型的数据生成对应的数据摘要子集,并将所述每种类型的数据对应的数据摘要子集成数据摘要。

6. 根据权利要求2所述的方法,其特征在于,当所述数据集合中包括多种类型的数据时,所述客户端针对存储的数据集合生成数据摘要,包括:

所述客户端针对每种类型的数据生成对应的数据摘要子集,并将所述每种类型的数据对应的数据摘要子集成数据摘要。

7. 根据权利要求3所述的方法,其特征在于,当所述数据集合中包括多种类型的数据时,所述客户端针对存储的数据集合生成数据摘要,包括:

所述客户端针对每种类型的数据生成对应的数据摘要子集,并将所述每种类型的数据对应的数据摘要子集成数据摘要。

8. 根据权利要求4所述的方法,其特征在于,当所述数据集合中包括多种类型的数据时,所述客户端针对存储的数据集合生成数据摘要,包括:

所述客户端针对每种类型的数据生成对应的数据摘要子集,并将所述每种类型的数据对应的数据摘要子集成数据摘要。

9. 根据权利要求1-8中任一项所述的方法,其特征在于,所述目标数据选择策略包括基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略中的至少一种。

10. 根据权利要求9所述的方法, 其特征在于, 当所述目标数据选择策略为所述基于置信度的选择策略时, 所述目标策略参数为置信度阈值, 所述客户端根据所述目标数据选择策略和所述目标策略参数从所述数据集合中确定数据子集, 包括:

所述客户端根据所述服务器发送的预测模型, 计算所述数据集合中的数据的置信度;

所述客户端从所述数据集合中提取置信度小于所述置信度阈值的数据, 并组合为数据子集。

11. 根据权利要求9所述的方法, 其特征在于, 当所述目标数据选择策略为所述抽样选择策略时, 所述目标策略参数为抽样比例, 所述客户端根据所述目标数据选择策略和所述目标策略参数从所述数据集合中确定数据子集, 包括:

所述客户端根据所述抽样比例确定所述数据集合中的数据的抽样数目;

所述客户端从所述数据集合中抽取抽样数目条数据, 并组合为数据子集。

12. 根据权利要求9所述的方法, 其特征在于, 当所述目标数据选择策略为所述基于聚类的选择策略时, 所述目标策略参数为聚类的群的数量, 所述客户端根据所述目标数据选择策略和所述目标策略参数从所述数据集合中确定数据子集, 包括:

所述客户端根据所述聚类的群的数量确定所述数据集合中的数据的待上传数目;

所述客户端从所述数据集合中选取待上传数目条数据, 并组合为数据子集。

13. 根据权利要求12所述的方法, 其特征在于, 当所述客户端产生新的数据时, 所述方法还包括:

所述客户端从所述数据子集中获取产生时间与所述新的数据的产生时间的间隔小于预设时间间隔的所有目标数据;

所述客户端计算所述所有目标数据与所述新的数据的相似度, 以获得最大相似度和所述最大相似度对应的目标数据;

所述客户端判断所述最大相似度是否大于预设门限值;

当所述最大相似度大于所述预设门限值时, 所述客户端将所述新的数据替换所述最大相似度对应的目标数据, 以获得第一数据子集;

其中, 所述客户端将所述数据子集发送至所述服务器, 包括:

所述客户端将所述第一数据子集发送至所述服务器。

14. 根据权利要求13所述的方法, 其特征在于, 所述方法还包括:

当所述最大相似度不大于所述预设门限值时, 所述客户端将所述新的数据添加至所述数据子集中, 以获得第二数据子集;

其中, 所述客户端将所述数据子集发送至所述服务器, 包括:

所述客户端将所述第二数据子集发送至所述服务器。

15. 一种数据处理方法, 其特征在于, 包括:

服务器接收若干个客户端发送的数据摘要;

所述服务器根据所述数据摘要从所述若干个客户端中确定允许上传数据的目标客户端;

所述服务器根据所述目标客户端对应的数据摘要确定目标数据选择策略, 以及根据所述目标客户端对应的数据摘要和所述目标数据选择策略确定目标策略参数;

所述服务器将所述目标数据选择策略和所述目标策略参数发送至所述目标客户端, 以

使所述目标客户端根据所述目标数据选择策略和所述目标策略参数确定待传输的数据子集,其中,所述目标策略参数用于限定所述数据子集中数据的数目;

所述服务器接收所述目标客户端发送的所述数据子集,并根据所述数据子集建立预测模型或更新已有的预测模型。

16. 根据权利要求15所述的方法,其特征在于,所述方法还包括:

所述服务器向所述若干个客户端中的除所述目标客户端外的其余客户端发送反馈信息,所述反馈信息用于指示所述若干个客户端中的除所述目标客户端外的其余客户端禁止向所述服务器发送数据。

17. 根据权利要求15所述的方法,其特征在于,所述服务器根据所述数据摘要从所述若干个客户端中确定允许上传数据的目标客户端,包括:

所述服务器根据所述数据摘要将所述若干个客户端划分为多个群;

所述服务器从所述多个群的每一个群中选取至少一个客户端,并将所述每一个群中选取的所述至少一个客户端作为允许上传数据的目标客户端。

18. 根据权利要求16所述的方法,其特征在于,所述服务器根据所述数据摘要从所述若干个客户端中确定允许上传数据的目标客户端,包括:

所述服务器根据所述数据摘要将所述若干个客户端划分为多个群;

所述服务器从所述多个群的每一个群中选取至少一个客户端,并将所述每一个群中选取的所述至少一个客户端作为允许上传数据的目标客户端。

19. 根据权利要求15所述的方法,其特征在于,所述服务器根据所述目标客户端对应的数据摘要和所述目标数据选择策略确定目标策略参数,包括:

所述服务器针对存储的数据集合生成所述服务器对应的数据摘要;

所述服务器计算所述服务器对应的数据摘要与所述目标客户端对应的数据摘要的相似度;

所述服务器按照所述目标数据选择策略对应的转换函数将所述相似度转换为目标策略参数。

20. 根据权利要求16所述的方法,其特征在于,所述服务器根据所述目标客户端对应的数据摘要和所述目标数据选择策略确定目标策略参数,包括:

所述服务器针对存储的数据集合生成所述服务器对应的数据摘要;

所述服务器计算所述服务器对应的数据摘要与所述目标客户端对应的数据摘要的相似度;

所述服务器按照所述目标数据选择策略对应的转换函数将所述相似度转换为目标策略参数。

21. 根据权利要求17所述的方法,其特征在于,所述服务器根据所述目标客户端对应的数据摘要和所述目标数据选择策略确定目标策略参数,包括:

所述服务器针对存储的数据集合生成所述服务器对应的数据摘要;

所述服务器计算所述服务器对应的数据摘要与所述目标客户端对应的数据摘要的相似度;

所述服务器按照所述目标数据选择策略对应的转换函数将所述相似度转换为目标策略参数。

22. 根据权利要求18所述的方法,其特征在于,所述服务器根据所述目标客户端对应的数据摘要和所述目标数据选择策略确定目标策略参数,包括:

所述服务器针对存储的数据集合生成所述服务器对应的数据摘要;

所述服务器计算所述服务器对应的数据摘要与所述目标客户端对应的数据摘要的相似度;

所述服务器按照所述目标数据选择策略对应的转换函数将所述相似度转换为目标策略参数。

23. 根据权利要求15-18中任一项所述的方法,其特征在于,所述服务器根据所述目标客户端对应的数据摘要确定目标数据选择策略,包括:

所述服务器根据预先建立的模型变化幅度、历史数据摘要与数据选择策略三者之间的对应关系,利用所述目标客户端对应的数据摘要计算不同数据选择策略下的模型变化幅度;

所述服务器从所述计算出的不同数据选择策略下的模型变化幅度中获取最大的模型变化幅度所对应的数据选择策略作为目标数据选择策略。

24. 根据权利要求23所述的方法,其特征在于,所述服务器根据所述目标客户端对应的数据摘要和所述目标数据选择策略确定目标策略参数,包括:

所述服务器根据所述目标数据选择策略下模型变化幅度与策略参数的对应关系,利用所述最大的模型变化幅度确定目标策略参数。

25. 根据权利要求15-22中任一项所述的方法,其特征在于,所述目标数据选择策略包括基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略中的至少一种。

26. 根据权利要求23所述的方法,其特征在于,所述目标数据选择策略包括基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略中的至少一种。

27. 根据权利要求24所述的方法,其特征在于,所述目标数据选择策略包括基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略中的至少一种。

28. 一种客户端,其特征在于,包括:

生成单元,用于针对存储的数据集合生成数据摘要;

发送单元,用于将所述数据摘要发送至服务器,所述数据摘要用于触发所述服务器以所述数据摘要为依据确定针对所述客户端的反馈信息;

接收单元,用于接收所述服务器发送的所述反馈信息;

确定单元,用于当所述反馈信息包括目标数据选择策略和目标策略参数时,根据所述目标数据选择策略和所述目标策略参数从所述数据集合中确定数据子集,其中,所述目标策略参数用于限定所述数据子集中数据的数目;

所述发送单元,还用于将所述数据子集发送至所述服务器。

29. 根据权利要求28所述的客户端,其特征在于,当所述反馈信息包括用于指示所述客户端禁止上传数据的信息时,所述客户端还包括:

控制单元,用于响应所述反馈信息,禁止向所述服务器发送数据。

30. 根据权利要求28所述的客户端,其特征在于,所述客户端还包括:

第一计算单元,用于计算所述数据摘要与至少一个历史数据摘要的相似度;

第一判断单元,用于判断所述计算出的相似度中是否存在相似度超过预设阈值,当所

述计算出的相似度中不存在相似度超过所述预设阈值时,触发所述发送单元执行所述的将所述数据摘要发送至服务器的操作。

31. 根据权利要求29所述的客户端,其特征在于,所述客户端还包括:

第一计算单元,用于计算所述数据摘要与至少一个历史数据摘要的相似度;

第一判断单元,用于判断所述计算出的相似度中是否存在相似度超过预设阈值,当所述计算出的相似度中不存在相似度超过所述预设阈值时,触发所述发送单元执行所述的将所述数据摘要发送至服务器的操作。

32. 根据权利要求28所述的客户端,其特征在于,当所述数据集合中包括多种类型的数据时,所述生成单元具体用于针对每种类型的数据生成对应的数据摘要子集,并将所述每种类型的数据对应的数据摘要子集合成数据摘要。

33. 根据权利要求29所述的客户端,其特征在于,当所述数据集合中包括多种类型的数据时,所述生成单元具体用于针对每种类型的数据生成对应的数据摘要子集,并将所述每种类型的数据对应的数据摘要子集合成数据摘要。

34. 根据权利要求30所述的客户端,其特征在于,当所述数据集合中包括多种类型的数据时,所述生成单元具体用于针对每种类型的数据生成对应的数据摘要子集,并将所述每种类型的数据对应的数据摘要子集合成数据摘要。

35. 根据权利要求31所述的客户端,其特征在于,当所述数据集合中包括多种类型的数据时,所述生成单元具体用于针对每种类型的数据生成对应的数据摘要子集,并将所述每种类型的数据对应的数据摘要子集合成数据摘要。

36. 根据权利要求28-35中任一项所述的客户端,其特征在于,所述目标数据选择策略包括基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略中的至少一种。

37. 根据权利要求36所述的客户端,其特征在于,当所述目标数据选择策略为所述基于置信度的选择策略时,所述目标策略参数为置信度阈值,所述确定单元包括:

计算子单元,用于根据所述服务器发送的预测模型,计算所述数据集合中的数据的置信度;

提取子单元,用于从所述数据集合中提取置信度小于所述置信度阈值的数据,并组合为数据子集。

38. 根据权利要求36所述的客户端,其特征在于,当所述目标数据选择策略为所述抽样选择策略时,所述目标策略参数为抽样比例,所述确定单元包括:

第一确定子单元,用于根据所述抽样比例确定所述数据集合中的数据的抽样数目;

抽取子单元,用于从所述数据集合中抽取抽样数目条数据,并组合为数据子集。

39. 根据权利要求36所述的客户端,其特征在于,当所述目标数据选择策略为所述基于聚类的选择策略时,所述目标策略参数为聚类的群的数量,所述确定单元包括:

第二确定子单元,用于根据所述聚类的群的数量确定所述数据集合中的数据的待上传数目;

选取子单元,用于从所述数据集合中选取待上传数目条数据,并组合为数据子集。

40. 根据权利要求39所述的客户端,其特征在于,当所述客户端产生新的数据时,所述客户端还包括:

获取单元,用于从所述数据子集中获取产生时间与所述新的数据的产生时间的间隔小

于预设时间间隔的所有目标数据；

第二计算单元，用于计算所述所有目标数据与所述新的数据的相似度，以获得最大相似度和所述最大相似度对应的目标数据；

第二判断单元，用于判断所述最大相似度是否大于预设门限值；

替换单元，用于当所述第二判断单元判断出所述最大相似度大于所述预设门限值时，将所述新的数据替换所述最大相似度对应的目标数据，以获得第一数据子集；

所述发送单元将所述数据子集发送至所述服务器的方式具体为：

所述发送单元将所述第一数据子集发送至所述服务器。

41. 根据权利要求40所述的客户端，其特征在于，所述客户端还包括：

添加单元，用于当所述第二判断单元判断出所述最大相似度不大于所述预设门限值时，将所述新的数据添加至所述数据子集中，以获得第二数据子集；

所述发送单元将所述数据子集发送至所述服务器的方式具体为：

所述发送单元将所述第二数据子集发送至所述服务器。

42. 一种服务器，其特征在于，包括：

接收单元，用于接收若干个客户端发送的数据摘要；

第一确定单元，用于根据所述数据摘要从所述若干个客户端中确定允许上传数据的目标客户端；

第二确定单元，用于根据所述目标客户端对应的数据摘要确定目标数据选择策略；

第三确定单元，用于根据所述目标客户端对应的数据摘要和所述目标数据选择策略确定目标策略参数；

发送单元，用于将所述目标数据选择策略和所述目标策略参数发送至所述目标客户端，以使所述目标客户端根据所述目标数据选择策略和所述目标策略参数确定待传输的数据子集，其中，所述目标策略参数用于限定所述数据子集中数据的数目；

所述接收单元，还用于接收所述目标客户端发送的所述数据子集，并根据所述数据子集建立预测模型或更新已有的预测模型。

43. 根据权利要求42所述的服务器，其特征在于，

所述发送单元，还用于向所述若干个客户端中的除所述目标客户端外的其余客户端发送反馈信息，所述反馈信息用于指示所述若干个客户端中的除所述目标客户端外的其余客户端禁止向所述服务器发送数据。

44. 根据权利要求42所述的服务器，其特征在于，所述第一确定单元包括：

划分子单元，用于根据所述数据摘要将所述若干个客户端划分为多个群；

选取子单元，用于从所述多个群的每一个群中选取至少一个客户端，并将所述每一个群中选取的所述至少一个客户端作为允许上传数据的目标客户端。

45. 根据权利要求43所述的服务器，其特征在于，所述第一确定单元包括：

划分子单元，用于根据所述数据摘要将所述若干个客户端划分为多个群；

选取子单元，用于从所述多个群的每一个群中选取至少一个客户端，并将所述每一个群中选取的所述至少一个客户端作为允许上传数据的目标客户端。

46. 根据权利要求42-45中任一项所述的服务器，其特征在于，

所述第三确定单元包括：

生成子单元,用于针对存储的数据集合生成所述服务器对应的数据摘要;

第一计算子单元,用于计算所述服务器对应的数据摘要与所述目标客户端对应的数据摘要的相似度;

转换子单元,用于按照所述目标数据选择策略对应的转换函数将所述相似度转换为目标策略参数。

47.根据权利要求42-45中任一项所述的服务器,其特征在于,

所述第二确定单元包括:

第二计算子单元,用于根据预先建立的模型变化幅度、历史数据摘要与数据选择策略三者之间的对应关系,利用所述目标客户端对应的数据摘要计算不同数据选择策略下的模型变化幅度;

获取子单元,用于从所述第二计算子单元计算出的不同数据选择策略下的模型变化幅度中获取最大的模型变化幅度所对应的数据选择策略作为目标数据选择策略。

48.根据权利要求47所述的服务器,其特征在于,所述第三确定单元根据所述目标客户端对应的数据摘要和所述目标数据选择策略确定目标策略参数的方式具体为:所述第三确定单元根据所述目标数据选择策略下模型变化幅度与策略参数的对应关系,利用所述最大的模型变化幅度确定目标策略参数。

49.根据权利要求42-45中任一项所述的服务器,其特征在于,所述目标数据选择策略包括基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略中的至少一种。

50.根据权利要求46所述的服务器,其特征在于,所述目标数据选择策略包括基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略中的至少一种。

51.根据权利要求47所述的服务器,其特征在于,所述目标数据选择策略包括基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略中的至少一种。

52.根据权利要求48所述的服务器,其特征在于,所述目标数据选择策略包括基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略中的至少一种。

53.一种客户端,其特征在于,包括处理器、存储器、总线和通信接口;所述存储器用于存储执行指令,所述处理器与所述存储器通过所述总线连接,当所述客户端运行时,所述处理器执行所述存储器存储的所述执行指令,以使所述客户端执行权利要求1-14中任一项所述的方法。

54.一种服务器,其特征在于,包括处理器、存储器、总线和通信接口;所述存储器用于存储执行指令,所述处理器与所述存储器通过所述总线连接,当所述服务器运行时,所述处理器执行所述存储器存储的所述执行指令,以使所述服务器执行权利要求15-27中任一项所述的方法。

55.一种数据处理系统,其特征在于,包括至少一个如权利要求28-41中任一项所述的客户端以及如权利要求42-52中任一项所述的服务器。

56.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质存储有计算机程序,所述计算机程序被硬件执行时能够实现权利要求1-14中任一项所述的方法。

57.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质存储有计算机程序,所述计算机程序被硬件执行时能够实现权利要求15-27中任一项所述的方法。

一种数据处理方法及相关设备、系统

技术领域

[0001] 本发明涉及数据处理技术领域,尤其涉及一种数据处理方法及相关设备、系统。

背景技术

[0002] 数据挖掘是指从大量的数据中通过算法搜索隐藏于其中信息的过程,通过统计、分析处理、情报检索、机器学习和模式识别等诸多方法来实现上述目标。例如,根据标注好的短信类型(如垃圾短信,正常短信)的短信文本,学习一个预测模型,对于新的短信文本,基于该模型,自动识别该短信属于哪一种类型,就是一个数据挖掘的过程。

[0003] 目前,数据挖掘大多是基于Client-Server架构的场景。例如,智能手环自动识别当前用户行为模式的预测模型(如走路,骑车,游泳等),模型是在服务器中学习并同步给手环的,手环需要不断传送新的数据给服务器,以便服务器不断更新得到更准确的模型。又如,智能交通监控摄像头自动识别车辆违章的预测模型(如闯红灯、逆行、违章变线等),模型是在服务器中学习并同步给摄像头的,摄像头需要不断传送新的数据给服务器,以便服务器不断更新得到更准确的模型。然而,由于客户端需要不断传送新的数据给服务器,服务器需要处理的数据量巨大,对带宽、流量、功耗或计算能力等方面的资源消耗很大,因此,如何降低数据处理过程中的资源消耗,是当前业界亟需解决的问题。

发明内容

[0004] 本发明实施例公开了一种数据处理方法及相关设备、系统,用于解决如何降低数据处理过程中的资源消耗的问题。

[0005] 本发明实施例第一方面公开了一种数据处理方法,包括:

[0006] 客户端针对存储的数据集合生成数据摘要,并将该数据摘要发送至服务器,该数据摘要用于触发服务器以该数据摘要为依据确定针对客户端的反馈信息;

[0007] 客户端接收服务器发送的该反馈信息;

[0008] 当该反馈信息包括目标数据选择策略和目标策略参数时,客户端根据目标数据选择策略和目标策略参数从数据集合中确定数据子集,其中,目标数据选择策略用于指示客户端以何种方式选择数据,目标策略参数用于限定选取出的该数据子集中数据的数目;

[0009] 客户端将该数据子集发送至服务器。这样可以控制客户端向服务器传输的数据量,从而有效降低数据处理过程中的资源消耗;此外,能够精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型。

[0010] 结合本发明实施例第一方面,在本发明实施例第一方面的第一种可能的实施方式中,当该反馈信息包括用于指示客户端禁止上传数据的信息时,该方法还包括:

[0011] 客户端响应该反馈信息,禁止向服务器发送数据。客户端可以根据需求对无需发送的数据进行清除,以释放客户端的存储空间。

[0012] 结合本发明实施例第一方面或第一方面的第一种可能的实施方式,在本发明实施例第一方面的第二种可能的实施方式中,该方法还包括:

- [0013] 客户端计算该数据摘要与至少一个历史数据摘要的相似度；
- [0014] 客户端判断计算出的相似度中是否存在相似度超过预设阈值；
- [0015] 当计算出的相似度中不存在相似度超过预设阈值时，客户端执行上述的将该数据摘要发送至服务器的步骤。这样可以避免多次发送相似的数据摘要到服务器，减少不必要的数据发送。
- [0016] 结合本发明实施例第一方面或第一方面的第一种或第二种可能的实施方式，在本发明实施例第一方面的第三种可能的实施方式中，当数据集合中包括多种类型的数据时，客户端针对存储的数据集合生成数据摘要，包括：
- [0017] 客户端针对每种类型的数据生成对应的数据摘要子集，并将每种类型的数据对应的数据摘要子集成数据摘要。
- [0018] 结合本发明实施例第一方面或第一方面的第一种至第三种中任一种可能的实施方式，在本发明实施例第一方面的第四种可能的实施方式中，目标数据选择策略可以包括但不限于基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略等中的至少一种。其中，抽样选择策略可以包括但不限于随机抽样选择策略、等距抽样选择策略、分层抽样选择策略等中的任一种。
- [0019] 结合本发明实施例第一方面的第四种可能的实施方式，在本发明实施例第一方面的第五种可能的实施方式中，当目标数据选择策略为基于置信度的选择策略时，目标策略参数为置信度阈值，客户端根据目标数据选择策略和目标策略参数从数据集合中确定数据子集，包括：
- [0020] 客户端根据服务器发送的预测模型，计算数据集合中的数据的置信度；
- [0021] 客户端从数据集合中提取置信度小于置信度阈值的数据，并组合为数据子集。
- [0022] 结合本发明实施例第一方面的第四种可能的实施方式，在本发明实施例第一方面的第六种可能的实施方式中，当目标数据选择策略为抽样选择策略时，目标策略参数为抽样比例，客户端根据目标数据选择策略和目标策略参数从数据集合中确定数据子集，包括：
- [0023] 客户端根据所述抽样比例确定数据集合中的数据的抽样数目；
- [0024] 客户端从数据集合中抽取该抽样数目条数据，并组合为数据子集。
- [0025] 结合本发明实施例第一方面的第四种可能的实施方式，在本发明实施例第一方面的第七种可能的实施方式中，当目标数据选择策略为基于聚类的选择策略时，目标策略参数为聚类的群的数量，客户端根据目标数据选择策略和目标策略参数从数据集合中确定数据子集，包括：
- [0026] 客户端根据聚类的群的数量确定数据集合中的数据的待上传数目；
- [0027] 客户端从数据集合中选取该待上传数目条数据，并组合为数据子集。
- [0028] 结合本发明实施例第一方面的第七种可能的实施方式，在本发明实施例第一方面的第八种可能的实施方式中，当客户端产生新的数据时，该方法还包括：
- [0029] 客户端从该数据子集中获取产生时间与新的数据的产生时间的间隔小于预设时间间隔的所有目标数据；
- [0030] 客户端计算所有目标数据与该新的数据的相似度，以获得最大相似度和该最大相似度对应的目标数据；
- [0031] 客户端判断最大相似度是否大于预设门限值；

[0032] 当最大相似度大于预设门限值时,客户端将该新的数据替换该最大相似度对应的目标数据,以获得第一数据子集;

[0033] 其中,客户端将该数据子集发送至服务器,包括:

[0034] 客户端将第一数据子集发送至服务器。

[0035] 结合本发明实施例第一方面的第八种可能的实施方式,在本发明实施例第一方面的第九种可能的实施方式中,该方法还包括:

[0036] 当目标数据不存在或最大相似度不大于预设门限值时,客户端将该新的数据添加至该数据子集中,以获得第二数据子集;

[0037] 其中,客户端将该数据子集发送至服务器,包括:

[0038] 客户端将第二数据子集发送至服务器。

[0039] 本发明实施例第二方面公开了一种数据处理方法,包括:

[0040] 服务器接收若干个客户端发送的数据摘要;

[0041] 服务器根据该数据摘要从上述若干个客户端中确定允许上传数据的目标客户端;

[0042] 服务器根据目标客户端对应的数据摘要确定目标数据选择策略,以及根据目标客户端对应的数据摘要和目标数据选择策略确定目标策略参数;

[0043] 服务器将目标数据选择策略和目标策略参数发送至目标客户端,以使目标客户端根据目标数据选择策略和目标策略参数确定待传输的数据子集,其中,目标数据选择策略用于指示目标客户端以何种方式选择数据,目标策略参数用于限定选取出的该数据子集中数据的数目;

[0044] 服务器接收目标客户端发送的该数据子集,并根据该数据子集建立预测模型或更新已有的预测模型。这样可以控制客户端向服务器传输的数据量,从而有效降低服务器在数据处理过程中的资源消耗;此外,能够精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型。

[0045] 结合本发明实施例第二方面,在本发明实施例第二方面的第一种可能的实施方式中,所述方法还包括:

[0046] 服务器向上述若干个客户端中的除目标客户端外的其余客户端发送反馈信息,该反馈信息用于指示上述若干个客户端中的除目标客户端外的其余客户端禁止向服务器发送数据。客户端可以根据需求对无需发送的数据进行清除,以释放客户端的存储空间。

[0047] 结合本发明实施例第二方面或第二方面的第一种可能的实施方式,在本发明实施例第二方面的第二种可能的实施方式中,服务器根据该数据摘要从上述若干个客户端中确定允许上传数据的目标客户端,包括:

[0048] 服务器根据该数据摘要将上述若干个客户端划分为多个群;

[0049] 服务器从上述多个群的每一个群中选取至少一个客户端,并将每一个群中选取的上述至少一个客户端作为允许上传数据的目标客户端。这样可以极大降低客户端传输数据到服务器的并发量。

[0050] 结合本发明实施例第二方面或第二方面的第一种或第二种可能的实施方式,在本发明实施例第二方面的第三种可能的实施方式中,服务器根据目标客户端对应的数据摘要和目标数据选择策略确定目标策略参数,包括:

[0051] 服务器针对存储的数据集合生成服务器对应的数据摘要;

- [0052] 服务器计算服务器对应的数据摘要与目标客户端对应的数据摘要的相似度；
- [0053] 服务器按照目标数据选择策略对应的转换函数将该相似度转换为目标策略参数。
- [0054] 结合本发明实施例第二方面或第二方面的第一种或第二种可能的实施方式，在本发明实施例第二方面的第四种可能的实施方式中，服务器根据目标客户端对应的数据摘要确定目标数据选择策略，包括：
- [0055] 服务器根据预先建立的模型变化幅度、历史数据摘要与数据选择策略三者之间的对应关系，利用目标客户端对应的数据摘要计算不同数据选择策略下的模型变化幅度；
- [0056] 服务器从计算出的不同数据选择策略下的模型变化幅度中获取最大的模型变化幅度所对应的数据选择策略作为目标数据选择策略。
- [0057] 结合本发明实施例第二方面的第四种可能的实施方式，在本发明实施例第二方面的第五种可能的实施方式中，服务器根据目标客户端对应的数据摘要和目标数据选择策略确定目标策略参数，包括：
- [0058] 服务器根据目标数据选择策略下模型变化幅度与策略参数的对应关系，利用该最大的模型变化幅度确定目标策略参数。
- [0059] 结合本发明实施例第二方面或第二方面的第一种至第五种中任一种可能的实施方式，在本发明实施例第二方面的第六种可能的实施方式中，目标数据选择策略可以包括但不限于基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略等中的至少一种。其中，抽样选择策略可以包括但不限于随机抽样选择策略、等距抽样选择策略、分层抽样选择策略等中的任一种。
- [0060] 本发明实施例第三方面公开了一种客户端，包含用于执行本发明实施例第一方面公开的方法中的任一种可能的实施方式的模块。
- [0061] 本发明实施例第四方面公开了一种服务器，包含用于执行本发明实施例第二方面公开的方法中的任一种可能的实施方式的模块。
- [0062] 本发明实施例第五方面公开了一种计算机可读存储介质，其中存储有可执行的程序代码，该程序代码用于实现本发明实施例第一方面公开的方法中的全部或部分步骤。
- [0063] 本发明实施例第六方面公开了一种计算机可读存储介质，其中存储有可执行的程序代码，该程序代码用于实现本发明实施例第二方面公开的方法中的全部或部分步骤。
- [0064] 本发明实施例第七方面公开了一种客户端，包括处理器，存储器，通信接口和总线，处理器与存储器、通信接口通过总线连接。存储器用于存储指令，处理器用于执行该指令，在控制器的控制下，通信接口用于与服务器通信。当处理器执行存储器存储的指令时，使得处理器执行本发明实施例第一方面公开的方法及其中的任一种可能的实施方式。
- [0065] 本发明实施例第八方面公开了一种服务器，包括处理器，存储器，通信接口和总线，处理器与存储器、通信接口通过总线连接。存储器用于存储指令，处理器用于执行该指令，在控制器的控制下，通信接口用于与客户端通信。当处理器执行存储器存储的指令时，使得处理器执行本发明实施例第二方面公开的方法及其中的任一种可能的实施方式。
- [0066] 本发明实施例第九方面公开了一种数据处理系统，包括至少一个如本发明实施例第三方面公开的客户端以及如本发明实施例第四方面公开的服务器。
- [0067] 本发明实施例中，客户端针对存储的数据集合生成数据摘要，并将该数据摘要发送至服务器，服务器根据该数据摘要确定针对客户端的反馈信息，并将该反馈信息发送给

客户端,当客户端接收到的反馈信息中包括目标数据选择策略和目标策略参数时,客户端可以根据该目标数据选择策略和该目标策略参数从数据集合中确定出数据子集,并将该数据子集发送至服务器,服务器可以根据该数据子集建立预测模型或者更新已有的预测模型。可见,实施本发明实施例,在客户端选择数据发送至服务器之前,通过客户端生成数据摘要,将数据摘要发送服务器,由服务器来决定合适于客户端的数据选择策略和策略参数,以使得客户端按照数据选择策略来进行数据的选择,以及根据策略参数来确定需要选择的数据的数目,从而可以控制客户端向服务器传输的数据量,有效降低数据处理过程中的资源消耗;此外,能够精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型。

附图说明

[0068] 为了更清楚地说明本发明实施例中的技术方案,下面将对实施例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发明的一些实施例,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。

[0069] 图1是本发明实施例公开的一种基于C/S结构的网络架构示意图;

[0070] 图2是本发明实施例公开的一种计算机设备的结构示意图;

[0071] 图3是本发明实施例公开的一种数据处理方法的流程示意图;

[0072] 图4是本发明实施例公开的另一种数据处理方法的流程示意图;

[0073] 图5是本发明实施例公开的又一种数据处理方法的流程示意图;

[0074] 图6是本发明实施例公开的一种预测模型更新的示意图;

[0075] 图7是本发明实施例公开的一种客户端的结构示意图;

[0076] 图8是本发明实施例公开的另一种客户端的结构示意图;

[0077] 图9是本发明实施例公开的又一种客户端的结构示意图;

[0078] 图10是本发明实施例公开的又一种客户端的结构示意图;

[0079] 图11是本发明实施例公开的又一种客户端的结构示意图;

[0080] 图12是本发明实施例公开的一种服务器的结构示意图;

[0081] 图13是本发明实施例公开的另一种服务器的结构示意图;

[0082] 图14是本发明实施例公开的又一种服务器的结构示意图;

[0083] 图15是本发明实施例公开的又一种服务器的结构示意图;

[0084] 图16是本发明实施例公开的一种数据处理系统的结构示意图。

具体实施方式

[0085] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0086] 本发明实施例公开了一种数据处理方法及相关设备、系统,可以精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型,并可以有效降低客户端向服务器传输的数据量。以下分别进行详细说明。

[0087] 为了更好的理解本发明实施例,下面先对本发明实施例公开的一种基于C/S结构的网络架构进行描述。请参阅图1,图1是本发明实施例公开的一种基于C/S结构的网络架构示意图。其中,图1所示的网络架构是基于客户端/服务器(Client/Server,C/S)结构的,可以用于进行数据处理,如数据挖掘。在图1所示的网络架构中,可以包括至少一个客户端(Client)和服务器(Server),其中,客户端可以包括智能手机、平板电脑、掌上电脑、个人数字助理(Personal Digital Assistant,PDA)、移动互联网设备(Mobile Internet Device,MID)、智能穿戴设备(如智能手表、智能手环等)、智能监控摄像头等各类设备,本发明实施例不作限定。服务器可以同时支持多个客户端,如客户端1、客户端2、……、客户端n,其中,n为正整数。服务器可以是数据库服务器,主要用于处理数据查询或数据操纵的请求,可以为用户提供查询、更新、事务管理、缓存、多用户存取控制等服务。

[0088] 在图1所示的网络架构中,客户端与服务器之间可以预先建立通信连接,具体地,客户端与服务器之间可以通过标准的协议,如HTTP(HyperText Transfer Protocol,超文本传输协议)协议等来传输信息;客户端与服务器之间也可以通过专门的协议,如传感器专用的协议等来传输信息,本发明实施例不作限定。在现有的数据挖掘中,客户端可以不断地向服务器发送数据,以使得服务器可以利用接收到的数据来建立预测模型,或者来更新服务器中已有的预测模型。然而,由于客户端需要不断传送新的数据给服务器,服务器需要处理的数据量巨大,对带宽、流量、功耗或计算能力等方面的资源消耗很大。当客户端中没有产生新的数据、或者产生的新的数据价值不大的情况下,客户端没有必要向服务器发送数据。此外,当服务器连接的客户端数量较多时,服务器接收并处理多个客户端的数据会造成负载压力过大(如某一时刻处理数据的数量过大)。基于现有技术的缺陷,本发明实施例中,在客户端向服务器发送数据之前,通过客户端生成本地数据摘要,并将该数据摘要发送给服务器,由服务器根据该数据摘要来决定客户端是否需要发送数据,从而可以有效降低服务器的负载压力。此外,服务器还可以根据该数据摘要来决定客户端用于选择数据的数据选择策略以及用于限定选择数据的数量的策略参数,从而可以控制客户端向服务器传输的数据量,有效降低数据处理过程中的资源消耗;此外,可以精确、便捷地选取具有代表性的数据来发送至服务器,以便获得更准确的预测模型。

[0089] 基于图1所示的网络架构,本发明实施例公开了一种客户端和服务器。其中,客户端和服务器均可以通过计算机设备的形式实现。请参阅图2,图2是本发明实施例公开的一种计算机设备的结构示意图。如图2所示,该计算机设备可以包括:处理器202、存储器204、通信接口206和总线208。其中,处理器202、存储器204和通信接口206通过总线208实现彼此之间的通信连接。本领域技术人员可以理解,图2所示的计算机设备的结构并不构成对本发明的限定,它既可以是总线形结构,也可以是星型结构,还可以包括比图2所示的结构更多或更少的部件,或者组合某些部件,或者不同的部件布置。

[0090] 处理器202可以采用通用的中央处理器(Central Processing Unit,CPU),微处理器,应用专用集成电路(Application Specific Integrated Circuit,ASIC),或者一个或多个集成电路,用于执行相关程序,以实现本发明实施例所提供的技术方案。

[0091] 存储器204可以是只读存储器(Read Only Memory,ROM),静态存储设备,动态存储设备或者随机存取存储器(Random Access Memory,RAM)。存储器204可以存储操作系统2041和其他应用程序2042。在通过软件或者固件来实现本发明实施例提供的技术方案时,

用于实现本发明实施例提供的技术方案的程序代码保存在存储器204中,并由处理器202来执行。

[0092] 通信接口206使用例如但不限于收发器一类的收发装置,来实现与其他设备或通信网络之间的通信。

[0093] 总线208可包括一通路,在各个部件(例如处理器202、存储器204、通信接口206)之间传送信息。

[0094] 在一个实施例中,当图2所示的计算机设备为客户端时,处理器202可以调用存储器204中存储的程序代码,用于执行以下操作:

[0095] 针对存储器204中存储的数据集合生成数据摘要,并控制通信接口206将该数据摘要发送至服务器,该数据摘要用于触发服务器以该数据摘要为依据确定针对客户端的反馈信息;

[0096] 控制通信接口206接收服务器发送的反馈信息;

[0097] 当该反馈信息包括目标数据选择策略和目标策略参数时,根据目标数据选择策略和目标策略参数从数据集合中确定数据子集,其中,目标策略参数用于限定该数据子集中数据的数目;

[0098] 控制通信接口206将该数据子集发送至服务器。

[0099] 通过实施本实施例,客户端在选择数据发送至服务器之前,通过生成数据摘要,将数据摘要发送服务器,由服务器来决定合适于客户端的数据选择策略和策略参数,以使得客户端按照数据选择策略来进行数据的选择,以及根据策略参数来确定需要选择的数据的数目,从而可以控制客户端向服务器传输的数据量,有效降低数据处理过程中的资源消耗;此外,能够精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型。

[0100] 在一个实施例中,当图2所示的计算机设备为服务器时,处理器202可以调用存储器204中存储的程序代码,用于执行以下操作:

[0101] 控制通信接口206接收若干个客户端发送的数据摘要;

[0102] 根据该数据摘要从上述若干个客户端中确定允许上传数据的目标客户端;

[0103] 根据目标客户端对应的数据摘要确定目标数据选择策略,以及根据目标客户端对应的数据摘要和目标数据选择策略确定目标策略参数;

[0104] 控制通信接口206将目标数据选择策略和目标策略参数发送至目标客户端,以使目标客户端根据目标数据选择策略和目标策略参数确定待传输的数据子集,其中,目标策略参数用于限定数据子集中数据的数目;

[0105] 控制通信接口206接收目标客户端发送的数据子集,并根据该数据子集建立预测模型或更新已有的预测模型。

[0106] 通过实施本实施例,服务器在接收客户端发送的数据之前,先接收客户端生成的数据摘要,并根据该数据摘要来决定是否允许上传数据以及客户端的数据选择策略和策略参数,以使得客户端按照数据选择策略来进行数据的选择,以及根据策略参数来确定需要选择的数据的数目,从而可以控制客户端向服务器传输的数据量,从而有效降低服务器在数据处理过程中的资源消耗;此外,能够精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型。

[0107] 基于图1所示的网络架构,本发明实施例公开了一种数据处理方法。请参阅图3,图

3是本发明实施例公开的一种数据处理方法的流程示意图。如图3所示,该数据处理方法可以包括以下步骤:

[0108] 301、客户端针对存储的数据集合生成数据摘要。

[0109] 302、客户端将该数据摘要发送至服务器。

[0110] 本发明实施例中,客户端中可以预先存储有多项数据,这些数据可以构成一个数据集合。客户端可以针对该数据集合生成数据摘要,具体地,客户端可以是针对该数据集合中的全部或者部分数据来生成数据摘要,其中,可能的实施方式可以包括但不限于以下几种情况:

[0111] (1) 客户端针对该数据集合中的所有数据生成数据摘要;

[0112] (2) 客户端针对该数据集合中排除已发送给服务器的那部分数据生成数据摘要;

[0113] (3) 客户端针对该数据集合中给定时间段内产生的所有数据生成数据摘要,如最近10分钟内产生的数据;

[0114] (4) 客户端针对该数据集合中指定的若干条数据生成数据摘要,如最近产生的1000条数据;

[0115] (5) 客户端针对该数据集合中指定一次或若干次行为所产生的数据生成数据摘要,如最近3次运动产生的数据。

[0116] 本发明实施例中,数据摘要可以是指数据的统计特征和描述信息,即用于描述数据的特征。数据摘要可以包括但不限于统计量(如数据量、最大值、最小值、均值、方差等),数据的分布(如数据在不同区间上的数据比例),有标注的数据比例(如数据量为6,有标注的数据有3条,则有标注的数据比例为50%),异常的数据比例(如可以通过区间来判断,落在该区间外的为异常数据),置信度的最大值、最小值、平均值、标准方差、分布等等。

[0117] 作为一种可选的实施方式,当该数据集合中包括多种类型的数据时,步骤301客户端针对存储的数据集合生成数据摘要的具体实施方式可以包括以下步骤:

[0118] 30) 客户端针对每种类型的数据生成对应的数据摘要子集,并将每种类型的数据对应的数据摘要子集成数据摘要。

[0119] 下面以运动数据来举例说明,例如运动数据包括速度、心率、海拔等不同类型的数
据,如下表1所示:

[0120] 表1

[0121]

时间	10:15:01	10:15:02	10:15:03	10:15:04	10:15:05	10:15:06
速度 (公里/小时)	5	7	9	10	6	20
心率 (次/分钟)	80	95	110	140	130	90
海拔 (米)	20	10	30	15	22	18
标注	无标注	无标注	跑步	跑步	步行	无标注

[0122] 对于每种类型的数据(速度、心率、海拔)可以分别计算对应的数据摘要子集,下面以速度为例来说明如何计算,对于其他类型的数据,计算方法都是相同的。

[0123] 速度的数据摘要为:

[0124] 统计量:数据量=6,最小值=5,最大值=20,均值=9.5,标准方差=5.468等。

[0125] 数据的分布:数据的分布是数据在不同区间上的数据比例。例如把数据分为4个区间:[0,6),[6,8),[8,10),[10,∞),每个区间的数据量为1、2、1、2,对应的数据比例为0.17、0.33、0.17、0.33,即为该数据的分布。

[0126] 有标注的数据比例:表1中有标注的数据为3条,无标注的数据为3条,则有标注的数据比例为50%。

[0127] 异常的数据比例:检测数据异常的方法有多种,一般可以通过(均值-3*标准方差,均值+3*标准方差)区间来判断,落在该区间之外的为异常数据。本例中该区间为(-6.904,25.904),没有落在该区间内的数据,因此异常的数据比例为0。

[0128] 按照上述方法可以得出速度、心率、海拔的数据摘要,如下表2所示,表2中示出的仅是部分内容。其中,置信度的最大值,最小值,平均值,方差,分布等统计量:基于客户端本地预测模型(即为服务器反馈给客户端的预测模型),对于每条数据(即由某时刻下的每种类型的数据构成,如某时刻下的速度、心率和海拔三个数据构成一条数据),可以计算出置信度。基于由每条数据的置信度构成的置信度序列,可以计算出最大值,最小值,平均值,方差,分布等置信度的统计量。表2中置信度的平均值即为置信度序列的平均值。置信度的计算方法可以有多种,举例来说,预测模型的输入是手环运动数据,输出是用户当前状态(如走路、骑车、游泳)的概率值,每条数据预测出一组概率值。若 X_1, X_2, X_3 分别表示一条数据里速度、心率、海拔的取值,基于预测模型计算每条数据的概率的公式为:走路的概率 $P_1 = 1 / (1 + \exp(c_{11} * X_1 + c_{21} * X_2 + c_{31} * X_3))$,骑车的概率 $P_2 = 1 / (1 + \exp(c_{12} * X_1 + c_{22} * X_2 + c_{32} * X_3))$,游泳的概率 $P_3 = 1 / (1 + \exp(c_{13} * X_1 + c_{23} * X_2 + c_{33} * X_3))$,其中, $c_{11}, c_{21}, c_{31}, c_{12}, c_{22}, c_{32}, c_{13}, c_{23}, c_{33}$ 等预测模型的参数是已知。在已知一条数据(X_1, X_2, X_3)的情况下,将该条数据代入上述公式,可以计算出用户走路,骑车,游泳的概率值。假设 $P_1 = 0.7, P_2 = 0.1, P_3 = 0.2$ 。基于上述概率值,使用置信度计算公式,可以计算出对该条数据预测结果的置信度。例如,置信度计算公式为:置信度=最大的概率-次大的概率,则对于上面的例子,计算出该条

数据的置信度为 $0.7-0.2=0.5$ 。

[0129] 表2

[0130]	速度的平 均值	速度的 方差	心率 90-120 的 比例	海拔的最 大值	置信度的平 均值	有标注的数 据比例	...
	9.5	5.468	0.5	30	0.2	0.5	

[0131] 其中,数据摘要是一个向量,向量的每一个元素(如速度的方差)称之为一个数据摘要特征。

[0132] 本发明实施例中,客户端与服务器之间可以建立通信连接,两者之间可以通过标准的协议(如HTTP协议)连接,也可以通过专门的协议(如传感器专有的协议)连接,以使客户端将生成的数据摘要发送至服务器。其中,该数据摘要用于触发服务器以该数据摘要为依据确定针对客户端的反馈信息。一个服务器可以同时与多个客户端进行交互,使得一个服务器可以在同时段内接收多个客户端发送的数据摘要,服务器可以对每个客户端发送来的数据摘要进行统计分析,以决定哪些客户端允许上传数据,哪些客户端不允许上传数据,并通过反馈信息告知客户端。

[0133] 本发明实施例中,触发客户端生成和发送数据摘要的条件可以包括但不限于以下几种情况中的一种或多种:

[0134] 1、定时:例如客户端每隔特定时间(如5分钟)计算1次数据摘要并发送至服务器。

[0135] 2、数据量触发:例如客户端的数据量累积到指定值(如1000条)后,计算数据摘要并发送至服务器。

[0136] 3、存储触发:例如客户端数据存储空间(如磁盘或内存)达到设定上限后,计算数据摘要并发送至服务器。

[0137] 4、事件触发:例如客户端在某些事件(例如手环用户完成一次运动)产生后,计算数据摘要并发送至服务器。

[0138] 5、服务器触发:服务器主动向客户端发送请求,要求客户端计算数据摘要并发送至服务器。

[0139] 作为一种可选的实施方式,在执行步骤301之后,执行步骤302之前,图3所描述的方法还可以包括以下步骤:

[0140] 31a) 客户端计算该数据摘要与至少一个历史数据摘要的相似度;

[0141] 31b) 客户端判断计算出的相似度中是否存在相似度超过预设阈值,若是,则不将该数据摘要发送给服务器;若否,则触发执行步骤302将该数据摘要发送至服务器。

[0142] 在该实施方式中,当客户端生成数据摘要后,可以将该数据摘要与一个或多个历史数据摘要进行相似度计算,分别计算出该数据摘要与这些历史数据摘要的相似度。其中,客户端可以在本地存储前一次或多次的数据摘要,即历史数据摘要。因为数据摘要是一个向量,两个数据摘要的相似度可以用向量的余弦相似度公式来计算。

[0143] 在该实施方式中,客户端可以将计算出的相似度与预设阈值进行比较,当有相似度大于预设阈值时,可以说明新生成的数据摘要与某个历史数据摘要的相似度高,此时不

需要将该数据摘要发送至服务器,从而可以避免多次发送相似的数据摘要到服务器。当计算出的相似度都小于预设阈值时,可以说明新生成的数据摘要与历史数据摘要的相似度不大,即有较大区别,则可以进一步执行步骤302将新生成的数据摘要发送至服务器,从而更新服务器中的预测模型。

[0144] 303、客户端接收服务器发送的反馈信息。

[0145] 304、当该反馈信息包括目标数据选择策略和目标策略参数时,客户端根据目标数据选择策略和目标策略参数从数据集合中确定数据子集。

[0146] 本发明实施例中,当客户端接收到的反馈信息中包括目标数据选择策略和目标策略参数时,可以表明服务器允许客户端向其传输数据,此时客户端可以根据目标数据选择策略和目标策略参数从数据集合中确定出待传输的数据子集。客户端可以是数据集合中的所有数据中选取数据子集,也可以是从数据集合中用于生成数据摘要的那部分数据中选取数据子集,本发明实施例不作限定。目标数据选择策略用于指示客户端以何种方式从数据集合中选择数据,可以包括但不限于基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略等中的至少一种。其中,抽样选择策略可以包括但不限于随机抽样选择策略、等距抽样选择策略以及分层抽样选择策略等中的任一种。目标策略参数用于指示客户端从数据集合中选取多少条数据,即用于限定数据子集中数据的数目。

[0147] 作为一种可选的实施方式,当目标数据选择策略为基于置信度的选择策略时,目标策略参数为置信度阈值,步骤304客户端根据目标数据选择策略和目标策略参数从数据集合中确定数据子集的具体实施方式可以包括以下步骤:

[0148] 32a) 客户端根据服务器发送的预测模型,计算数据集合中的数据的置信度;

[0149] 32b) 客户端从数据集合中提取置信度小于置信度阈值的数据,并组合为数据子集。

[0150] 在该实施方式中,对于方差较大,但置信度普遍较高的数据,可以采用基于置信度的方法。服务器可以将训练得到的预测模型反馈给客户端,客户端可以利用该预测模型计算出数据集合中各条数据的置信度,并从中提取出置信度小于置信度阈值的数据,例如置信度阈值为0.5,可以从数据集合中选取所有置信度小于0.5的数据。选取的数据具有较高的代表性,且选取的数据的量也会比较少,从而可以降低客户端向服务器传输的数据量。

[0151] 作为一种可选的实施方式,当目标数据选择策略为抽样选择策略时,目标策略参数为抽样比例,步骤304客户端根据目标数据选择策略和目标策略参数从数据集合中确定数据子集的具体实施方式可以包括以下步骤:

[0152] 33a) 客户端根据抽样比例确定数据集合中的数据的抽样数目;

[0153] 33b) 客户端从数据集合中抽取抽样数目条数据,并组合为数据子集。

[0154] 在该实施方式中,对于方差较小的数据,说明数据的波动幅度较小,可以采用抽样的方法,计算量小,且可以选择得到代表性较好的数据。选取的数据的量由抽样比例决定,例如,当抽样比例为0.1,表示从100条数据中抽取10条数据。

[0155] 在该实施方式中,抽样选择策略可以包括但不限于随机抽样选择策略、等距抽样选择策略以及分层抽样选择策略等中的任一种。例如有100条数据,需要抽取10条数据,随机抽样选择策略可以从这100条数据中随机抽取10条数据组成数据子集;等距抽样选择策略,即间隔若干条数据抽取1条数据,可以抽取数据中的第1、11、21、31、41、51、61、71、81、

91条数据,组合为数据子集;分层抽样选择策略将这100条数据按某种特征或某种规则划分为不同的层,然后从不同的层中独立、随机或等距地抽取代表数据。

[0156] 作为一种可选的实施方式,当目标数据选择策略为基于聚类的选择策略时,目标策略参数为聚类的群的数量,步骤304客户端根据目标数据选择策略和目标策略参数从数据集合中确定数据子集的具体实施方式可以包括以下步骤:

[0157] 34a) 客户端根据聚类的群的数量确定数据集合中的数据的待上传数目;

[0158] 34b) 客户端从数据集合中选取上述待上传数目条数据,并组合为数据子集。

[0159] 在该实施方式中,对于方差较大,置信度普遍较低的数据,可以采用基于聚类的方法,例如,聚类的群的数量为10,则将数据集合中的数据划分为10个群,从每个群中选取1条代表数据,最终将得到的10条代表数据构成数据子集。这10条代表数据可以是随机选取的,也可以是基于一定的规则进行选取的,该实施方式不作限定。

[0160] 需要说明的是,上述三种数据选择策略可以单独使用,也可以结合使用,本发明实施例不作限定。

[0161] 305、客户端将该数据子集发送至服务器。

[0162] 本发明实施例中,客户端向服务器发送该数据子集后,服务器可以利用该数据子集来建立新的预测模型;或者利用该数据子集来更新已有的预测模型,以便获得更准确的预测模型。客户端将数据子集发送至服务器后,根据需求或基于存储空间考虑,可以对已发送的数据进行清除处理。

[0163] 作为一种可选的实施方式,图3所描述的方法还可以包括以下步骤:

[0164] 35) 当该反馈信息包括用于指示客户端禁止上传数据的信息时,客户端响应该反馈信息,禁止向服务器发送数据。

[0165] 在该实施方式中,当该反馈信息包括用于指示客户端禁止上传数据的信息时,可以表明服务器不允许客户端向其传输数据,此时客户端本次操作将不向服务器发送数据。通过实施本实施方式,可以减少不必要的数据传输,且能够有效降低服务器的负载压力。

[0166] 本发明实施例中,当客户端在向服务器发送数据子集前又产生新的数据时,此时可以对该数据子集进行更新,并将更新后的数据子集发送给服务器,以使服务器获得更具代表性的数据。

[0167] 作为一种可选的实施方式,基于聚类的数据选择策略,当客户端产生新的数据时,图3所描述的方法还可以包括以下步骤:

[0168] 36) 客户端从该数据子集中获取产生时间与新的数据的产生时间的间隔小于预设时间间隔的所有目标数据;

[0169] 37) 客户端计算所有目标数据与该新的数据的相似度,以获得最大相似度和最大相似度对应的目标数据;

[0170] 38) 客户端判断该最大相似度是否大于预设门限值;

[0171] 39) 当该最大相似度大于预设门限值时,客户端将该新的数据替换该最大相似度对应的目标数据,以获得更新后的第一数据子集;

[0172] 相应地,步骤305客户端将该数据子集发送至服务器的具体实施方式可以为:

[0173] 客户端将更新后的第一数据子集发送至服务器。

[0174] 作为另一种可选的实施方式,图3所描述的方法还可以包括以下步骤:

[0175] 40) 当该最大相似度不大于预设门限值时,客户端将该新的数据添加至该数据子集中,以获得更新后的第二数据子集;

[0176] 相应地,步骤305客户端将该数据子集发送至服务器的具体实施方式可以为:

[0177] 客户端将更新后的第二数据子集发送至服务器。

[0178] 在该实施方式中,对于客户端中产生的新的数据,可以从先前确定的数据子集中选取产生时间与该新的数据的产生时间的间隔在预设时间间隔内的所有数据,这里考虑的是客户端的数据具有时效性,两条相似的数据,在不同的时间段出现,可能具有不同的意义,不能相互替代;当在较相近的时间出现时,可以相互替代。

[0179] 在该实施方式中,当没有产生时间与新的数据的产生时间的间隔小于预设时间间隔的数据,即目标数据为空集时,可以认为该新的数据与数据子集中的代表数据都相隔太久,此时可以将该新的数据作为一条新的代表数据,将其直接扩充到该数据子集中。如果目标数据不为空,可以计算该新的数据与每条目标数据的相似度,从而找出相似度最大的那条目标数据。其中,计算两条数据相似度的方法与具体的数据来源相关,例如:手环的运动数据,每条运动数据是多维的,可以看作是一个向量,因此可以利用余弦相似度来计算两条运动数据的相似度;摄像头的视频监控数据,其相似度是摄像头采集的图片帧的相似度。当最大相似度大于预设门限值时,则说明该新的数据替代掉数据子集中最大相似度对应的目标数据,从而得到第一数据子集;反之,两者不可替代,将该新的数据直接添加至数据子集中,以得到第二数据子集。

[0180] 举例来说,数据子集中包含数据1,产生时间为10:34:00;数据2,产生时间为10:35:06;数据3,产生时间为10:35:56;数据4,产生时间为10:36:15。客户端中有新产生的数据A,产生时间为10:36:45,预设时间间隔为2分钟,则获得的所有目标数据为数据2、数据3和数据4。分别计算出这三条数据与数据A的相似度为0.12、0.75、0.50,因此得出最大相似度为0.75,对应的是数据3。当0.75大于给定的预设门限值时,将数据A替换数据3;当0.75小于预设门限值时,数据A与数据3不可替换,直接将数据A添加进数据子集中。

[0181] 在图3所描述的方法中,客户端针对存储的数据集合生成数据摘要,并将该数据摘要发送至服务器,服务器根据该数据摘要确定针对客户端的反馈信息,并将该反馈信息发送给客户端,当客户端接收到的反馈信息中包括目标数据选择策略和目标策略参数时,客户端可以根据该目标数据选择策略和该目标策略参数从数据集合中确定出数据子集,并将该数据子集发送至服务器,服务器可以根据该数据子集建立预测模型或者更新已有的预测模型。通过实施图3所描述的方法,客户端在选择数据发送至服务器之前,通过生成数据摘要,将数据摘要发送服务器,由服务器来决定合适于客户端的数据选择策略和策略参数,以使得客户端按照数据选择策略来进行数据的选择,以及根据策略参数来确定需要选择的数据的数目,从而可以控制客户端向服务器传输的数据量,有效降低数据处理过程中的资源消耗;此外,能够精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型。

[0182] 基于图1所示的网络架构,本发明实施例公开了另一种数据处理方法。请参阅图4,图4是本发明实施例公开的另一种数据处理方法的流程示意图。如图4所示,该数据处理方法可以包括以下步骤:

[0183] 401、服务器接收若干个客户端发送的数据摘要。

[0184] 本发明实施例中,一个服务器可以同时与一个或多个客户端建立连接,以使得一

个服务器可以在同一时段内分别接收多个客户端发送的数据摘要。

[0185] 本发明实施例中,服务器响应客户端的请求至少有两种方式:

[0186] 1) 同步响应:即当客户端发送请求后,服务器立即响应。在这里客户端发送数据摘要至服务器后,服务器立即开始对该数据摘要进行分析处理。

[0187] 2) 异步响应:在这里服务器接收到客户端的数据摘要后,先缓存到服务器,并在满足触发条件后才开始对数据摘要进行分析处理。

[0188] 其中,异步响应的触发条件包括但不限于以下几种情况:

[0189] 1、服务器每隔一段时间(例如1分钟)定时触发对数据摘要的分析处理。

[0190] 2、发送数据摘要的客户端数量超过一定数量(例如100个)后,服务器开始对数据摘要的分析处理。

[0191] 3、客户端的数据摘要数量中,累积的数据量超过一定规模后,服务器开始进行响应。每个客户端的数据摘要包括了数据量的信息,服务器对多个客户端的数据量进行求和汇总,当总的的数据量超过一定规模(如100万条数据)后,开始对数据摘要进行分析处理。

[0192] 402、服务器根据该数据摘要从上述若干个客户端中确定允许上传数据的目标客户端。

[0193] 本发明实施例中,由于服务器可以连接多个客户端,当多个客户端均向服务器发送数据时,会使得服务器的负载压力过大,因此,可以从多个客户端中有目的的选取一部分客户端允许其上传数据,而阻止一部分客户端上传数据。允许上传数据的目标客户端可以是一个或多个。

[0194] 作为一种可选的实施方式,步骤402服务器根据该数据摘要从上述若干个客户端中确定允许上传数据的目标客户端的具体实施方式可以包括以下步骤:

[0195] 41) 服务器根据上述若干个客户端发送的数据摘要将上述若干个客户端划分为多个群;

[0196] 42) 服务器从上述多个群的每一个群中选取至少一个客户端,并将每一个群中选取的上述至少一个客户端作为允许上传数据的目标客户端。

[0197] 在该实施方式中,是通过聚类的方法,将数据摘要相似的客户端聚成一个群(也成为簇,cluster),聚类的输出是多个群,每个群包括多个客户端。聚类之后,从每个群中选至少一个客户端,其中,至少一个客户端不包括所有的客户端,则选中的客户端需要传输数据到服务器,该群中未选中的其它客户端则不用传输数据。实施该实施方式,可以极大地降低客户端传输数据到服务器的并发量。

[0198] 下面举例描述这个过程。例如有M个客户端,相应的数据摘要如下表3:

[0199] 表3

[0200]

Client ID	速度的 平均值	速度的 异常数 据比例	速度的 方差	心 率 90-120的 比例	海拔的 最大值	...	有标注 的数据 比例
1	6.7	0.3	2	0.6	0	...	0.9
2	4.8	0.5	6	0.3	1	...	0.9
...	...	...	...	...	...	...	...
M	5.5	0	8	0.2	7	...	0.5

[0201] 可以以k-means的聚类方法对上表中的M个客户端进行聚类处理。当选取的聚类中心个数为3时,假设得到的聚类结果如表4所示:

[0202] 表4

[0203]

Client ID	群ID
1	2
2	1
3	1
4	3
...	...
M	2

[0204] 对每个群中选择一个或几个代表的客户端。可以使用随机选择的方法,也可以使用基于规则的方法,例如选择距离该群的聚类中心最近的客户端。继续上面的例子,输出如下表5所示的结果:

[0205] 表5

[0206]

群ID	代表Client ID
1	2
2	100
3	4

[0207] 对于群1需要发送数据给服务器的代表客户端为群1中的客户端2,群1中的其它客户端则不需要发送数据给服务器。对于群2需要发送数据给服务器的代表客户端为群2中的客户端100,群2中的其它客户端则不需要发送数据给服务器。对于群3需要发送数据给服务器的代表客户端为群3中的客户端4,群3中的其它客户端则不需要发送数据给服务器。

[0208] 403、服务器根据目标客户端对应的数据摘要确定目标数据选择策略,以及根据目标客户端对应的数据摘要和目标数据选择策略确定目标策略参数。

[0209] 本发明实施例中,目标数据选择策略可以包括但不限于基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略等中的至少一种。其中,抽样选择策略可以包括但

不限于随机抽样选择策略、等距抽样选择策略以及分层抽样选择策略等中的任一种。

[0210] 本发明实施例中,服务器可以根据目标客户端对应的数据摘要中的特征信息来决定目标数据选择策略。例如,可以是基于经验规则来决定,当数据摘要中数据的方差较小时,则可以选用抽样选择策略作为目标客户端的目标数据选择策略;当数据摘要中数据的方差较大,且置信度普遍较高时,则可以选用基于置信度的选择策略作为目标客户端的目标数据选择策略;当数据摘要中数据的方差较大,但置信度普遍较低时,可以选用基于聚类的选择策略作为目标客户端的目标数据选择策略。

[0211] 作为一种可选的实施方式,步骤403中的服务器根据目标客户端对应的数据摘要和目标数据选择策略确定目标策略参数的具体实施方式可以包括以下步骤:

[0212] 43) 服务器针对存储的数据集合生成服务器对应的数据摘要;

[0213] 44) 服务器计算服务器对应的数据摘要与目标客户端对应的数据摘要的相似度;

[0214] 45) 服务器按照目标数据选择策略对应的转换函数将该相似度转换为目标策略参数。

[0215] 在该实施方式中,服务器中的数据集合是指服务器中当前存储的数据组合而成的,生成服务器对应的数据摘要与客户端生成数据摘要的方法相同。服务器的数据量更大,但是因为数据摘要的计算大多是可以增量(Increment)计算的,所以不会存在性能的问题,即在前一次计算出的数据摘要的基础上,只需利用新增的数据来更新前一次的数据摘要,而无需再使用所有的数据。例如,计算平均值时,只需利用前一次的平均值和新增的数据即可以计算出新的平均值,而无需再将所有的数据求和再计算平均值。计算服务器对应的数据摘要与每一个目标客户端对应的数据摘要的相似度,可以使用多种相似度计算方法,常见的有余弦相似度法、Pearson相似度法等。

[0216] 在该实施方式中,在计算得到服务器对应的数据摘要与各个目标客户端对应的数据摘要的相似度后,基于各个目标客户端的目标数据选择策略对应的转换函数(转换函数为基于历史的经验),确定各个目标客户端的目标策略参数,例如:对于随机抽样的抽样比例参数,相似度为0.3,转换函数为:抽样比例=相似度/2,则抽样比例为 $0.3/2=0.15$,从而确定了目标策略参数。

[0217] 作为一种可选的实施方式,服务器也可以基于数据挖掘建模来确定出目标客户端的目标数据选择策略和目标策略参数。步骤403中的服务器根据目标客户端对应的数据摘要确定目标数据选择策略的具体实施方式可以包括以下步骤:

[0218] 46) 服务器根据预先建立的模型变化幅度、历史数据摘要与数据选择策略三者之间的对应关系,利用目标客户端对应的数据摘要计算不同数据选择策略下的模型变化幅度;

[0219] 47) 服务器从计算出的不同数据选择策略下的模型变化幅度中获取最大的模型变化幅度所对应的数据选择策略作为目标数据选择策略。

[0220] 相应地,步骤403中的服务器根据目标客户端对应的数据摘要和目标数据选择策略确定目标策略参数的具体实施方式可以包括以下步骤:

[0221] 48) 服务器根据目标数据选择策略下模型变化幅度与策略参数的对应关系,利用该最大的模型变化幅度确定目标策略参数。

[0222] 在该实施方式中,基于历史数据(如客户端上传的历史数据摘要、服务器根据不同

的历史数据摘要得出的数据选择策略和策略参数等等)中,不同的数据摘要使用不同的数据选择策略以及策略参数所带来的模型变化幅度,构建一个训练数据集。基于该训练数据集,对于新数据的数据摘要,预测在服务器的负载(如服务器在某一时刻能够处理数据量的大小)低于上限的前提下,可能带来模型变化幅度最大的数据选择策略。这里的模型变化幅度是指预测模型的变化幅度,即一个预测模型在不同数据摘要下产生的变化程度。预测模型可以是逻辑回归、支持向量机、决策树等模型。

[0223] 在该实施方式中,可以分为训练和预测两个过程,其中,训练的目的在于建立模型变化幅度、历史数据摘要(即客户端前一次或多次上传的数据摘要)与数据选择策略三者之间的对应关系;预测的目的在于对新数据的数据摘要,基于训练得到的对应关系在不同数据选择策略下预计出相应的模型变化幅度,并将最大的模型变化幅度对应的数据选择策略作为目标客户端的目标数据选择策略。服务器可以是每接收到新数据的数据摘要即训练上述三者的对应关系,也可以是每隔特定时间训练一次上述三者的对应关系,还可以是当接收到的新数据的数据摘要的个数达到指定值则训练一次上述三者的对应关系,该实施方式对此不作限定。

[0224] 下面举例描述这个过程。训练的过程首先基于历史数据,构建训练数据集。训练数据集包含数据摘要、数据选择策略、策略参数、模型变化幅度四类字段。如下表6所示(注:表6中的数据只是随机给出的):

[0225] 表6

[0226]	速度的 平均值	速度的 方差	心 率 90-120 的比例	海拔的最 大值	心率的方 差	数据选择 策略ID	策 略 参 数	模 型 变 化幅度
	5	4.7	...	12	10.5	随机	0.1	0.3
	3	1.9	...	90	12	基于置信 度	0.5	1.2
	...	...	...	...		基于聚类	0.3	3.5
	10	7.4	...	52	33	基于置信 度	0.2	0.8

[0227] 模型变化幅度的计算方法为:模型变化幅度=100*(1-新模型参数向量与旧模型参数向量的相似度)。例如对于支持向量机算法,模型参数向量为支持向量这条直线的方程,例如 $3*x+2*y$ 的参数向量为(3,2)。新、旧模型参数向量的相似度可以用余弦相似度来计算,假设旧模型的参数向量为(1.5,2.4),新模型的参数向量为(1.7,2.0),则新旧模型的余弦相似度为 $(1.5*1.7+2.4*2.0)/\sqrt{((1.5*1.5+2.4*2.4)*(1.7*1.7+2.0*2.0))}=0.9894$,模型变化幅度为 $100*(1-0.9894)=1.06$ 。

[0228] 在构建好训练数据集后,使用数据挖掘中的回归模型(例如线性回归,多项式回归,指数回归等,回归模型是现有技术,其原理不详细介绍),可以学习到模型变化幅度与数据摘要和数据选择策略的关系。示例如下:

[0229] 对于抽样选择策略:模型变化幅度=速度的平均值+心率90-120的比例;

[0230] 对于基于置信度的选择策略:模型变化幅度= $0.5 \times$ 速度的方差+海拔的最大值;

[0231] 对于基于聚类的选择策略:模型变化幅度= $3 \times$ 速度的平均值 $-0.3 \times$ 心率的方差。

[0232] 预测的过程,基于新数据的数据摘要,预测对不同的数据选择策略,预计的模型变化幅度。例如新数据的数据摘要如表7所示:

[0233] 表7

[0234]	速度的平均值	速度的方差	心率90-120的比例	海拔的最大值	心率的方差
[0235]	4	3	0.3	6	2

[0236] 则几种选择策略的预计模型变化幅度分别为:对于抽样选择策略,模型变化幅度= $4+0.3=4.3$;对于基于置信度的选择策略:模型变化幅度= $0.5 \times 3+6=7.4$;对于基于聚类的选择策略:模型变化幅度= $3 \times 4-0.3 \times 2=11.4$ 。由于基于聚类的选择策略模型变化幅度最大,因此选择基于聚类的数据选择策略。可以说明,利用聚类的方法选取的代表数据能够使预测模型有较大变化,对于预测模型的更新起重要作用,对于抽样和基于置信度的方法选取的代表数据对预测模型的影响较小。因此可以首选对预测模型影响最大的数据选择策略。

[0237] 对基于聚类的数据选择策略,假定模型变化幅度Y与策略参数X(数据传输条数)的关系为:如果 $X>300$, $Y=0.5$;如果 $X \leq 300$, $Y=X/2$ 。如果服务器负载为40000,有100个客户端,分配给每个客户端的最高负载(数据传输条数)为400,则由于 $X=300$ 后模型变化幅度不再提升,因此最优的策略参数为300。其中,不同数据选择策略下模型变化幅度与策略参数的对应关系可以通过分析训练数据集中的数据得出。

[0238] 404、服务器将目标数据选择策略和目标策略参数发送至目标客户端,以使目标客户端根据目标数据选择策略和目标策略参数确定待传输的数据子集。

[0239] 本发明实施例中,目标策略参数用于限定该数据子集中数据的数目。

[0240] 405、服务器接收目标客户端发送的该数据子集,并根据该数据子集建立预测模型或更新已有的预测模型。

[0241] 本发明实施例中,图4所描述的方法还可以包括以下步骤:

[0242] 49) 服务器向上述若干个客户端中的除目标客户端外的其余客户端发送反馈信息,该反馈信息用于指示上述若干个客户端中的除目标客户端外的其余客户端禁止向服务器发送数据。其中,接收到该反馈信息的客户端可以丢弃相应的数据。

[0243] 通过实施图4所描述的方法,服务器在接收客户端发送的数据之前,先接收客户端生成的数据摘要,并根据该数据摘要来决定是否允许上传数据以及客户端的数据选择策略和策略参数,以使得客户端按照数据选择策略来进行数据的选择,以及根据策略参数来确定需要选择的数据的数目,从而可以控制客户端向服务器传输的数据量,从而有效降低服

务器在数据处理过程中的资源消耗;此外,能够精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型。

[0244] 基于图1所示的网络架构,本发明实施例公开了又一种数据处理方法。请参阅图5,图5是本发明实施例公开的又一种数据处理方法的流程示意图。其中,图5中仅示出的是一个客户端与服务器进行交互的过程,而每个客户端与服务器的交互过程均可以参考图5。如图5所示,该数据处理方法可以包括以下步骤:

[0245] 501、客户端针对存储的数据集合生成数据摘要。

[0246] 502、客户端将该数据摘要发送至服务器。

[0247] 503、服务器根据该数据摘要确定客户端是否允许上传数据,若允许,则执行步骤504。

[0248] 本发明实施例中,服务器可以采用聚类的方法来判断该客户端是否允许上传数据,具体地,服务器可以将该客户端发送的数据摘要与同时段内服务器接收到其他客户端的数据摘要进行相似度计算,将该客户端与该客户端的数据摘要相似的客户端聚成一个群,再从这个群中随机或基于经验规则的方法选取出一个或多个客户端作为允许向服务器上传数据的客户端,且该群中的其余客户端则为不允许向服务器上传数据的客户端。当该客户端为选取出的允许向服务器上传数据的客户端时,该客户端可以向服务器传输数据;当该客户端为不允许向服务器上传数据的客户端,则服务器可以向该客户端发送反馈信息告知该客户端无需上传数据,并结束本次操作。

[0249] 504、服务器确定针对客户端的目标数据选择策略和目标策略参数。

[0250] 本发明实施例中,当客户端允许上传数据时,服务器可以根据该客户端的数据摘要来确定出该客户端的目标数据选择策略和目标策略参数。可以是基于历史经验规则的方法来确定,也可以是基于数据挖掘建模的方法来确定,本发明实施例不作限定。

[0251] 本发明实施例中,目标数据选择策略可以包括但不限于基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略等中的至少一种。其中,抽样选择策略可以包括但不限于随机抽样选择策略、等距抽样选择策略、分层抽样选择策略等中的任一种。目标策略参数用于限定客户端进行数据选择时选取的数据的数量。

[0252] 505、服务器将目标数据选择策略和目标策略参数发送至客户端。

[0253] 506、客户端根据目标选择策略和目标策略参数从数据集合中确定数据子集。

[0254] 507、客户端将该数据子集发送至服务器。

[0255] 508、服务器根据该数据子集建立预测模型或更新已有的预测模型。

[0256] 举例来说,当服务器接收到客户端发送的数据子集后,可以对服务器中已有的预测模型进行更新。如图6所示,图6示出的是一种预测模型更新的示意图,这里涉及的是分类模型,将数据分为两个类,其中,加号表示一个分类的数据点,减号表示另一个分类的数据点,可以利用支持向量机的方法通过学习一条支持向量 L ,把这两个分类的数据区分开来。其中, $L(old)$ 为旧模型的分类型向量(实线所示),当客户端中有的新数据产生,上传新的数据子集时,这里两个类中分别新增了两个数据点,可以利用新的数据子集来更新旧模型,使得模型的分类型向量由 $L(old)$ 更新为 $L(new)$ (虚线所示),以对模型进行微调,使得模型更加准确。

[0257] 本发明实施例中,通过实施图5所描述的方法,在客户端选择数据发送至服务器之

前,通过客户端生成数据摘要,将数据摘要发送服务器,由服务器来决定合适于客户端的数据选择策略和策略参数,以使得客户端按照数据选择策略来进行数据的选择,以及根据策略参数来确定需要选择的数据的数目,从而可以控制客户端向服务器传输的数据量,从而有效降低服务器在数据处理过程中的资源消耗;此外,能够精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型。

[0258] 基于图1所示的网络架构,本发明实施例公开了另一种客户端。请参阅图7,图7是本发明实施例公开的一种客户端的结构示意图,可以用于执行本发明实施例公开的数据处理方法。如图7所示,该客户端可以包括:

[0259] 生成单元701,用于针对存储的数据集合生成数据摘要。

[0260] 本发明实施例中,数据集合为客户端当前存储的一项或多项数据组合而成的。具体地,生成单元701可以是针对该数据集合中的全部或者部分数据来生成数据摘要。

[0261] 本发明实施例中,数据摘要用于描述数据的特征,可以包括但不限于统计量,数据的分布,有标注的数据比例,异常的数据比例,置信度的最大值、最小值、平均值、标准方差、分布等等。数据摘要是一个向量,向量的每一个元素可以称之为一个数据摘要特征。

[0262] 作为一种可选的实施方式,当该数据集合中包括多种类型的数据时,生成单元701具体可以用于针对每种类型的数据生成对应的数据摘要子集,并将每种类型的数据对应的数据摘要子集成数据摘要。

[0263] 发送单元702,用于将该数据摘要发送至服务器。

[0264] 本发明实施例中,客户端与服务器之间可以建立通信连接,两者之间可以通过标准的协议(如HTTP协议)连接,也可以通过专门的协议(如传感器专有的协议)连接,以使发送单元702将生成单元701生成的数据摘要发送至服务器。其中,该数据摘要用于触发服务器以该数据摘要为依据确定针对客户端的反馈信息。

[0265] 接收单元703,用于接收服务器发送的反馈信息。

[0266] 确定单元704,用于当该反馈信息包括目标数据选择策略和目标策略参数时,根据目标数据选择策略和目标策略参数从数据集合中确定数据子集。

[0267] 本发明实施例中,目标数据选择策略用于指示客户端以何种方式从数据集合中选择数据,可以包括但不限于基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略等中的至少一种。其中,抽样选择策略可以包括但不限于随机抽样选择策略、等距抽样选择策略、分层抽样选择策略等中的任一种。目标策略参数用于指示客户端从数据集合中选取多少条数据,即用于限定数据子集中数据的数目。

[0268] 发送单元702,还用于将该数据子集发送至服务器。

[0269] 请一并参阅图8,图8是本发明实施例公开的另一种客户端的结构示意图,可以用于执行本发明实施例公开的数据处理方法。其中,图8所示的客户端是在图7所示的客户端的基础上进一步优化得到的。与图7所示的客户端相比,图8所示的客户端还可以包括:

[0270] 控制单元705,用于当该反馈信息包括用于指示客户端禁止上传数据的信息时,响应该反馈信息禁止向服务器发送数据。

[0271] 作为一种可选的实施方式,图8所示的客户端还可以包括:

[0272] 第一计算单元706,用于在生成单元701生成数据摘要之后,计算该数据摘要与至少一个历史数据摘要的相似度;

[0273] 第一判断单元707,用于判断计算出的相似度中是否存在相似度超过预设阈值,当计算出的相似度中不存在相似度超过预设阈值时,触发发送单元702执行上述的将该数据摘要发送至服务器的操作。

[0274] 作为一种可选的实施方式,当目标数据选择策略为基于置信度的选择策略时,目标策略参数为置信度阈值。请一并参阅图9,图9是本发明实施例公开的又一种客户端的结构示意图,可以用于执行本发明实施例公开的数据处理方法。其中,图9所示的客户端是在图8所示的客户端的基础上进一步优化得到的。与图8所示的客户端相比,图9所示的客户端中确定单元704可以包括:

[0275] 计算子单元7041,用于根据服务器发送的预测模型,计算数据集合中的数据的置信度;

[0276] 提取子单元7042,用于从数据集合中提取置信度小于置信度阈值的数据,并组合为数据子集。

[0277] 作为一种可选的实施方式,当目标数据选择策略为抽样选择策略时,目标策略参数为抽样比例。请一并参阅图10,图10是本发明实施例公开的又一种客户端的结构示意图,可以用于执行本发明实施例公开的数据处理方法。其中,图10所示的客户端是在图8所示的客户端的基础上进一步优化得到的。与图8所示的客户端相比,图10所示的客户端中确定单元704可以包括:

[0278] 第一确定子单元7043,用于根据抽样比例确定数据集合中的数据的抽样数目;

[0279] 抽取子单元7044,用于从数据集合中抽取该抽样数目条数据,并组合为数据子集。

[0280] 作为一种可选的实施方式,当目标数据选择策略为基于聚类的选择策略时,目标策略参数为聚类的群的数量。请一并参阅图11,图11是本发明实施例公开的又一种客户端的结构示意图,可以用于执行本发明实施例公开的数据处理方法。其中,图11所示的客户端是在图8所示的客户端的基础上进一步优化得到的。与图8所示的客户端相比,图11所示的客户端中确定单元704可以包括:

[0281] 第二确定子单元7045,用于根据聚类的群的数量确定数据集合中的数据的待上传数目;

[0282] 选取子单元7046,用于从数据集合中选取该待上传数目条数据,并组合为数据子集。

[0283] 作为一种可选的实施方式,当客户端产生新的数据时,图11所示的客户端还可以包括:

[0284] 获取单元708,用于从该数据子集中获取产生时间与该新的数据的产生时间的间隔小于预设时间间隔的所有目标数据;

[0285] 第二计算单元709,用于计算所有目标数据与该新的数据的相似度,以获得最大相似度和该最大相似度对应的目标数据;

[0286] 第二判断单元710,用于判断该最大相似度是否大于预设门限值;

[0287] 替换单元711,用于当第二判断单元710判断出该最大相似度大于预设门限值时,将该新的数据替换该最大相似度对应的目标数据,以获得第一数据子集;

[0288] 相应地,发送单元702将该数据子集发送至服务器的具体实施方式可以为:

[0289] 发送单元702将第一数据子集发送至服务器。

[0290] 作为一种可选的实施方式,图11所示的客户端还可以包括:

[0291] 添加单元712,用于当第二判断单元710判断出该最大相似度不大于预设门限值时,将该新的数据添加至该数据子集中,以获得第二数据子集;

[0292] 相应地,发送单元702将该数据子集发送至服务器的具体实施方式可以为:

[0293] 发送单元702将第二数据子集发送至服务器。

[0294] 本发明实施例中,通过实施图7至图11所示的客户端,在选择数据发送至服务器之前,通过生成数据摘要,将数据摘要发送服务器,由服务器来决定合适于客户端的数据选择策略和策略参数,以使得客户端按照数据选择策略来进行数据的选择,以及根据策略参数来确定需要选择的数据的数目,从而可以控制客户端向服务器传输的数据量,从而有效降低数据处理过程中的资源消耗;此外,能够精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型。

[0295] 基于图1所示的网络架构,本发明实施例公开了另一种服务器。请参阅图12,图12是本发明实施例公开的一种服务器的结构示意图,可以用于执行本发明实施例公开的数据处理方法。如图12所示,该服务器可以包括:

[0296] 接收单元1201,用于接收若干个客户端发送的数据摘要。

[0297] 本发明实施例中,一个服务器可以同时与一个或多个客户端建立连接,以使得接收单元1201可以在同一时段内分别接收多个客户端发送的数据摘要。

[0298] 第一确定单元1202,用于根据该数据摘要从上述若干个客户端中确定允许上传数据的目标客户端。

[0299] 本发明实施例中,允许上传数据的目标客户端可以是一个或多个。

[0300] 第二确定单元1203,用于根据目标客户端对应的数据摘要确定目标数据选择策略。

[0301] 本发明实施例中,目标数据选择策略可以包括但不限于基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略等中的至少一种。其中,抽样选择策略可以包括但不限于随机抽样选择策略、等距抽样选择策略、分层抽样选择策略等中的任一种。

[0302] 第三确定单元1204,用于根据目标客户端对应的数据摘要和目标数据选择策略确定目标策略参数。

[0303] 发送单元1205,用于将目标数据选择策略和目标策略参数发送至目标客户端,以使目标客户端根据目标数据选择策略和目标策略参数确定待传输的数据子集。

[0304] 本发明实施例中,目标策略参数用于限定该数据子集中数据的数目。

[0305] 接收单元1201,还用于接收目标客户端发送的数据子集,并根据该数据子集建立预测模型或更新已有的预测模型。

[0306] 本发明实施例中,发送单元1205,还用于向上述若干个客户端中的除目标客户端外的其余客户端发送反馈信息,该反馈信息用于指示上述若干个客户端中的除目标客户端外的其余客户端禁止向服务器发送数据。

[0307] 请一并参阅图13,图13是本发明实施例公开的另一种服务器的结构示意图,可以用于执行本发明实施例公开的数据处理方法。其中,图13所示的服务器是在图12所示的服务器的基础上进一步优化得到的。与图12所示的服务器相比,图13所示的服务器中的第一确定单元1202可以包括:

[0308] 划分子单元1202a,用于根据上述若干个客户端发送的数据摘要将上述若干个客户端划分为多个群;

[0309] 选取子单元1202b,用于从上述多个群的每一个群中选取至少一个客户端,并将每一个群中选取的上述至少一个客户端作为允许上传数据的目标客户端。

[0310] 作为一种可选的实施方式,请一并参阅图14,图14是本发明实施例公开的又一种服务器的结构示意图,可以用于执行本发明实施例公开的数据处理方法。其中,图14所示的服务器是在图13所示的服务器的基础上进一步优化得到的。与图13所示的服务器相比,图14所示的服务器中第三确定单元1204可以包括:

[0311] 生成子单元1204a,用于针对存储的数据集合生成服务器对应的数据摘要;

[0312] 第一计算子单元1204b,用于计算服务器对应的数据摘要与目标客户端对应的数据摘要的相似度;

[0313] 转换子单元1204c,用于按照目标数据选择策略对应的转换函数将该相似度转换为目标策略参数。

[0314] 作为一种可选的实施方式,请一并参阅图15,图15是本发明实施例公开的又一种服务器的结构示意图,可以用于执行本发明实施例公开的数据处理方法。其中,图15所示的服务器是在图13所示的服务器的基础上进一步优化得到的。与图13所示的服务器相比,图15所示的服务器中第二确定单元1203可以包括:

[0315] 第二计算子单元1203a,用于根据预先建立的模型变化幅度、历史数据摘要与数据选择策略三者之间的对应关系,利用目标客户端对应的数据摘要计算不同数据选择策略下的模型变化幅度;

[0316] 获取子单元1203b,用于从第二计算子单元1203a计算出的不同数据选择策略下的模型变化幅度中获取最大的模型变化幅度所对应的数据选择策略作为目标数据选择策略。

[0317] 相应地,第三确定单元1204根据目标客户端对应的数据摘要和目标数据选择策略确定目标策略参数的具体实施方式可以为:

[0318] 第三确定单元1204根据目标数据选择策略下模型变化幅度与策略参数的对应关系,利用该最大的模型变化幅度确定目标策略参数。

[0319] 本发明实施例中,通过实施图12至图15所示的服务器,在接收客户端发送的数据之前,先接收客户端生成的数据摘要,并根据该数据摘要来决定是否允许上传数据以及客户端的数据选择策略和策略参数,以使得客户端按照数据选择策略来进行数据的选择,以及根据策略参数来确定需要选择的数据的数目,从而可以控制客户端向服务器传输的数据量,从而有效降低服务器在数据处理过程中的资源消耗;此外,能够精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型。

[0320] 基于图1所示的网络架构,本发明实施例公开了一种数据处理系统。请参阅图16,图16是本发明实施例公开的一种数据处理系统的结构示意图。如图16所示,该数据处理系统可以包括:服务器1601和至少一个客户端1602。其中,服务器1601可以与客户端1602建立通信连接,客户端1602可以包括但不限于智能手机、平板电脑、PDA、智能手表、智能手环、智能监控摄像头等等设备。其中:

[0321] 客户端1602,用于针对存储的数据集合生成数据摘要,并将该数据摘要发送至服务器1601;

[0322] 服务器1601,用于以该数据摘要为依据确定客户端1602是否允许上传数据;

[0323] 服务器1601,还用于当确定出客户端1602允许上传数据后,根据客户端1602对应的数据摘要确定目标数据选择策略,以及根据客户端1602对应的数据摘要和目标数据选择策略确定目标策略参数;

[0324] 服务器1601,还用于将目标数据选择策略和目标策略参数发送至客户端1602;

[0325] 客户端1602,还用于接收目标数据选择策略和目标策略参数,并根据目标数据选择策略和目标策略参数确定待传输的数据子集,其中,目标策略参数用于限定该数据子集中数据的数目;

[0326] 客户端1602,还用于将该数据子集发送至服务器1601;

[0327] 服务器1601,还用于接收该数据子集,并根据该数据子集建立预测模型或更新已有的预测模型。

[0328] 本发明实施例中,服务器1601,还用于当确定出客户端1602不允许上传数据时,向客户端1602发送反馈信息,该反馈信息用于指示客户端1602禁止向服务器1601发送数据。

[0329] 本发明实施例中,目标数据选择策略可以包括但不限于基于置信度的选择策略、抽样选择策略以及基于聚类的选择策略中的至少一种。其中,抽样选择策略可以包括但不限于随机抽样选择策略、等距抽样选择策略、分层抽样选择策略等中的任一种。

[0330] 其中,本发明实施例涉及的服务器1601和客户端1602的其他具体功能可以参考前述实施例中的内容,在此不再赘述。

[0331] 本发明实施例中,通过实施图16所示的数据处理系统,在客户端选择数据发送至服务器之前,通过客户端生成数据摘要,将数据摘要发送服务器,由服务器来决定合适于客户端的数据选择策略和策略参数,以使得客户端按照数据选择策略来进行数据的选择,以及根据策略参数来确定需要选择的数据的数目,从而可以控制客户端向服务器传输的数据量,从而有效降低服务器在数据处理过程中的资源消耗;此外,能够精确、便捷地选择出具有代表性的数据,以便获得更准确的预测模型。

[0332] 需要说明的是,对于前述的各个方法实施例,为了简单描述,故将其都表述为一系列的动作组合,但是本领域技术人员应该知悉,本发明并不受所描述的动作顺序的限制,因为依据本申请,某一些步骤可以采用其他顺序或者同时进行。其次,本领域技术人员也应该知悉,说明书中所描述的实施例均属于优选实施例,所涉及的动作和模块并不一定是本申请所必须的。

[0333] 在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中沒有详细描述的部分,可以参见其他实施例的相关描述。

[0334] 本发明实施例的方法中的步骤可以根据实际需要进行顺序调整、合并和删减。

[0335] 本发明实施例客户端和服务端中的单元或子单元可以根据实际需要进行合并、划分和删减。

[0336] 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通程序来指令相关的硬件来完成,该程序可以存储于一计算机可读存储介质中,存储介质包括只读存储器(Read-Only Memory,ROM)、随机存储器(Random Access Memory, RAM)、可编程只读存储器(Programmable Read-only Memory,PROM)、可擦除可编程只读存储器(Erasable Programmable Read Only Memory,EPR0M)、一次可编程只读存储器(One-

time Programmable Read-Only Memory, OTPROM)、电子抹除式可复写只读存储器 (Electrically-Erasable Programmable Read-Only Memory, EEPROM)、只读光盘 (Compact Disc Read-Only Memory, CD-ROM) 或其他光盘存储器、磁盘存储器、磁带存储器、或者能够用于携带或存储数据的计算机可读的任何其他介质。

[0337] 以上对本发明实施例公开的一种数据处理方法及相关设备、系统进行了详细介绍, 本文中应用了具体个例对本发明的原理及实施方式进行了阐述, 以上实施例的说明只是用于帮助理解本发明的方法及其核心思想; 同时, 对于本领域的一般技术人员, 依据本发明的思想, 在具体实施方式及应用范围上均会有改变之处, 综上所述, 本说明书内容不应理解为对本发明的限制。

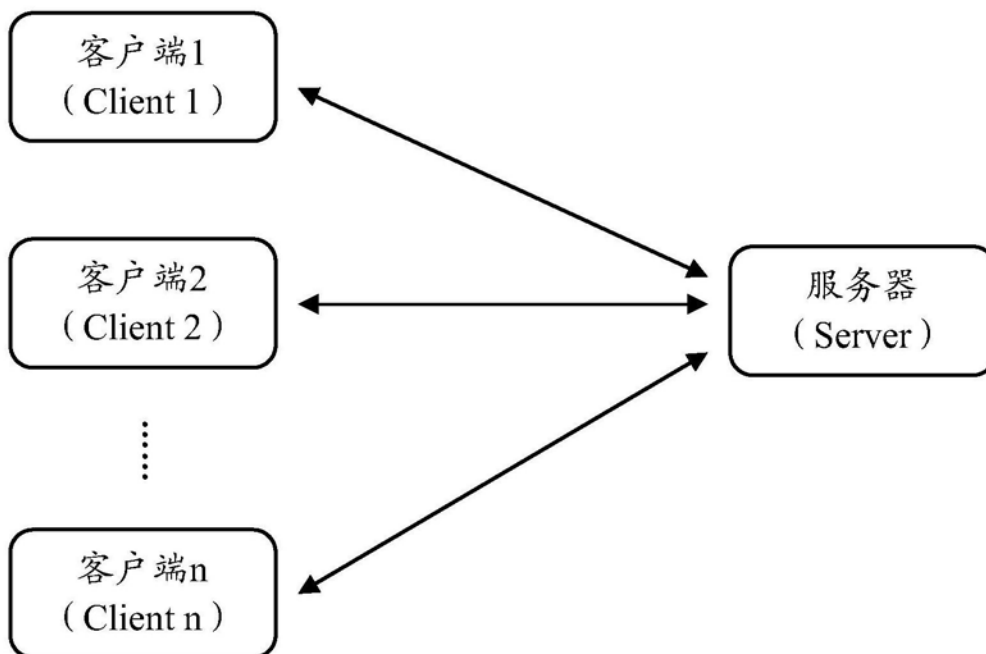


图1

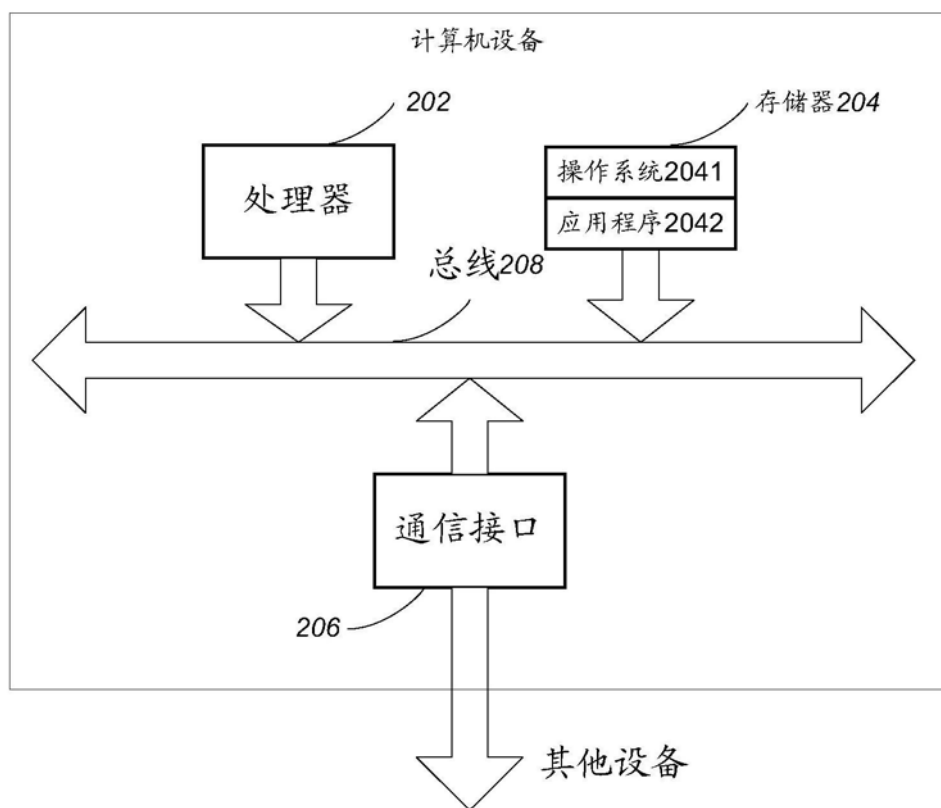


图2

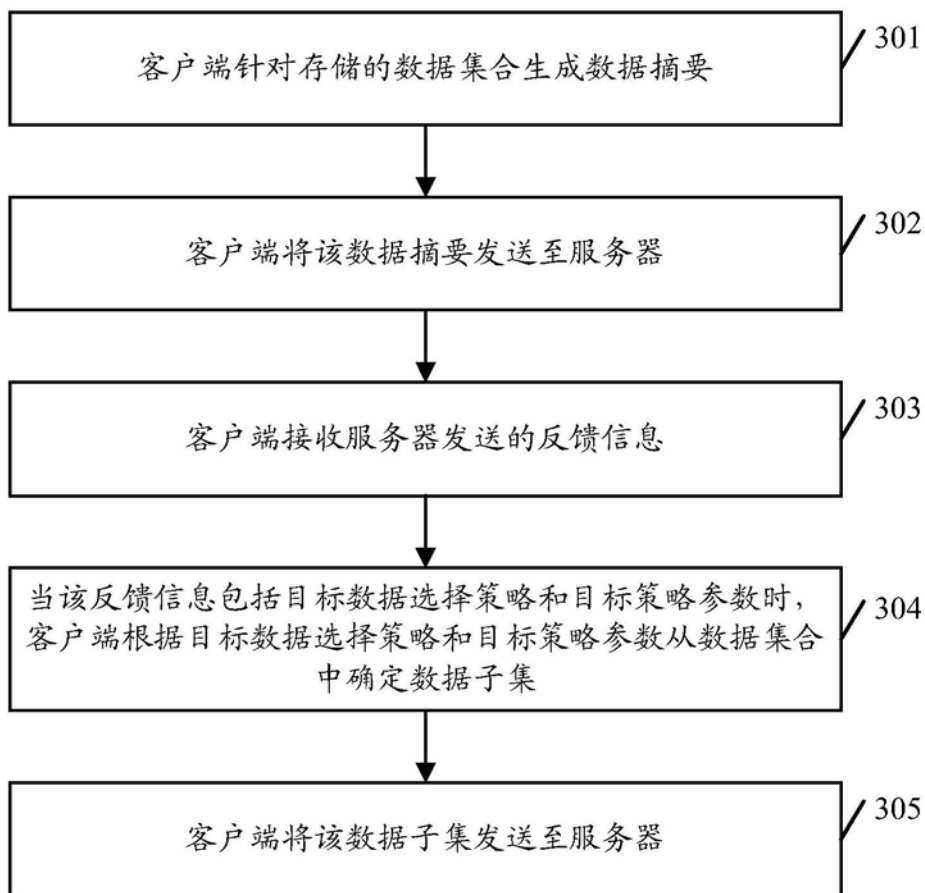


图3

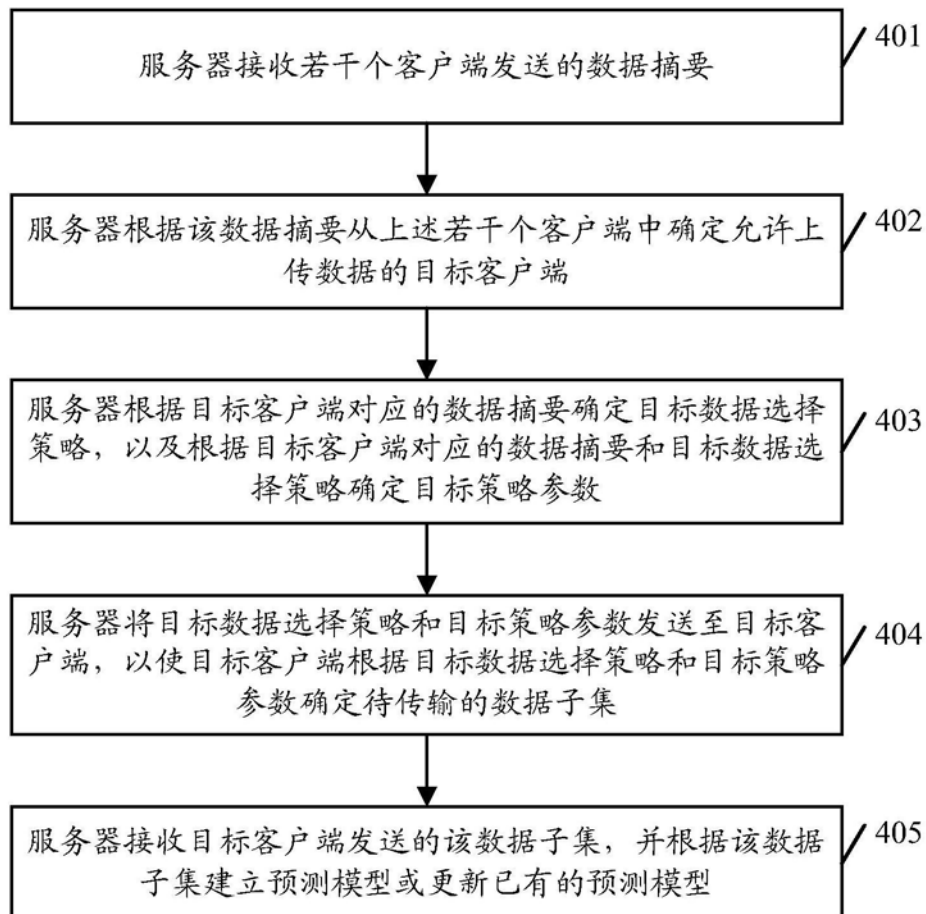


图4

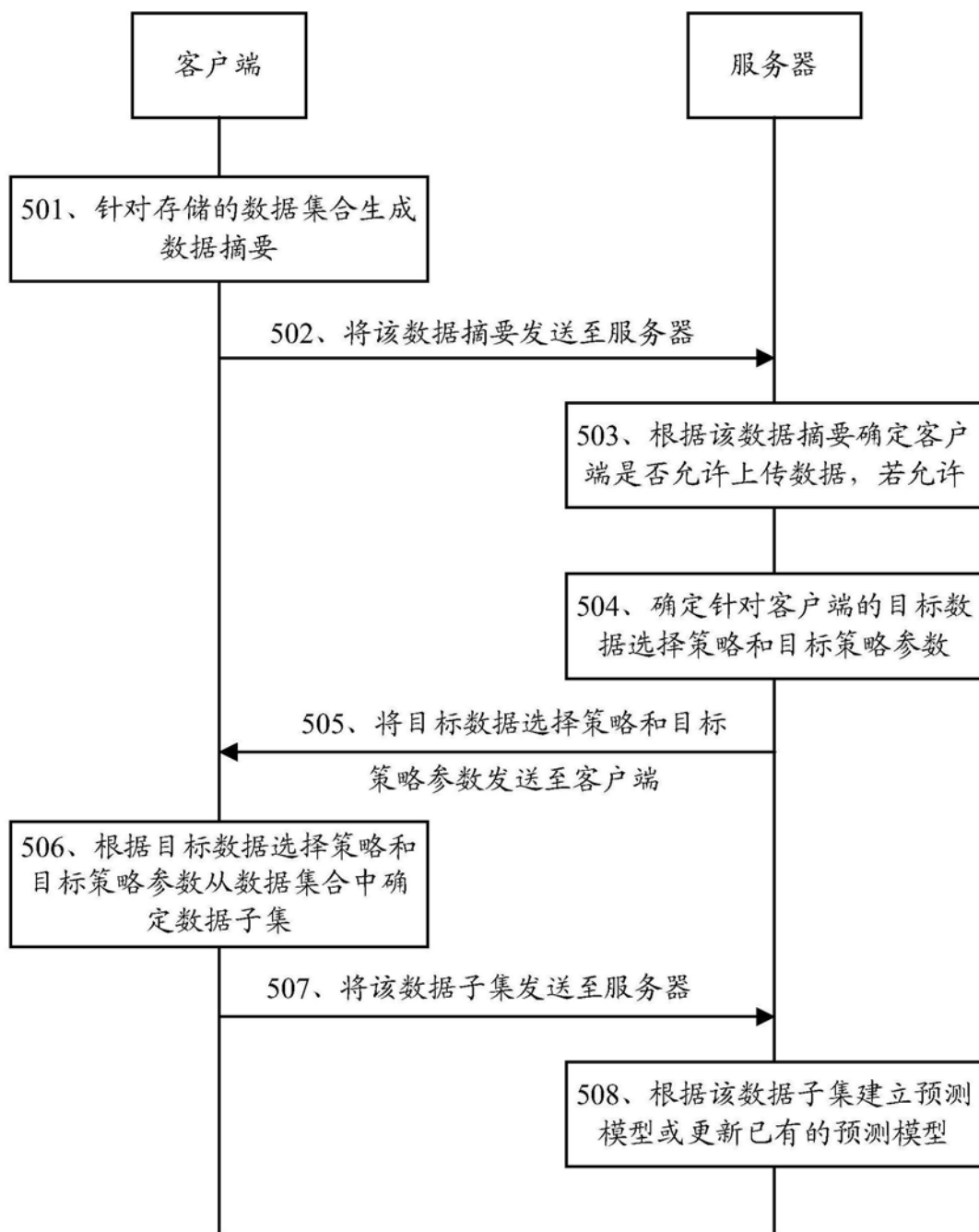


图5

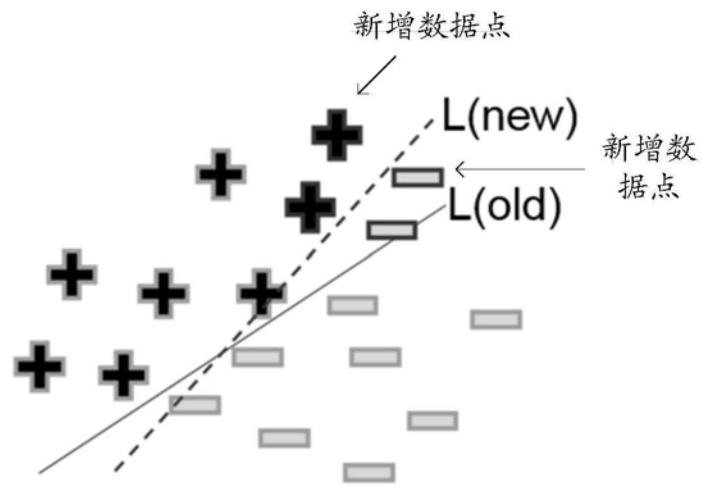


图6

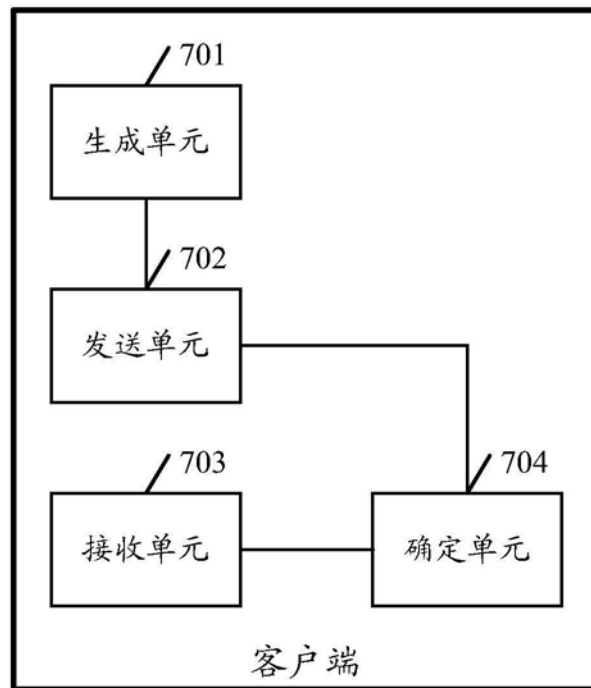


图7

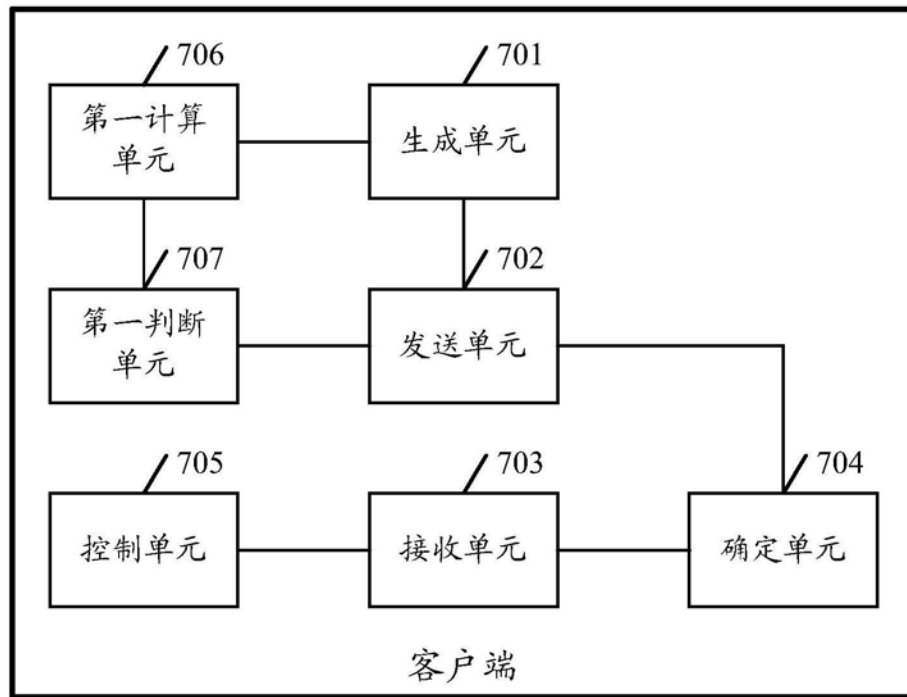


图8

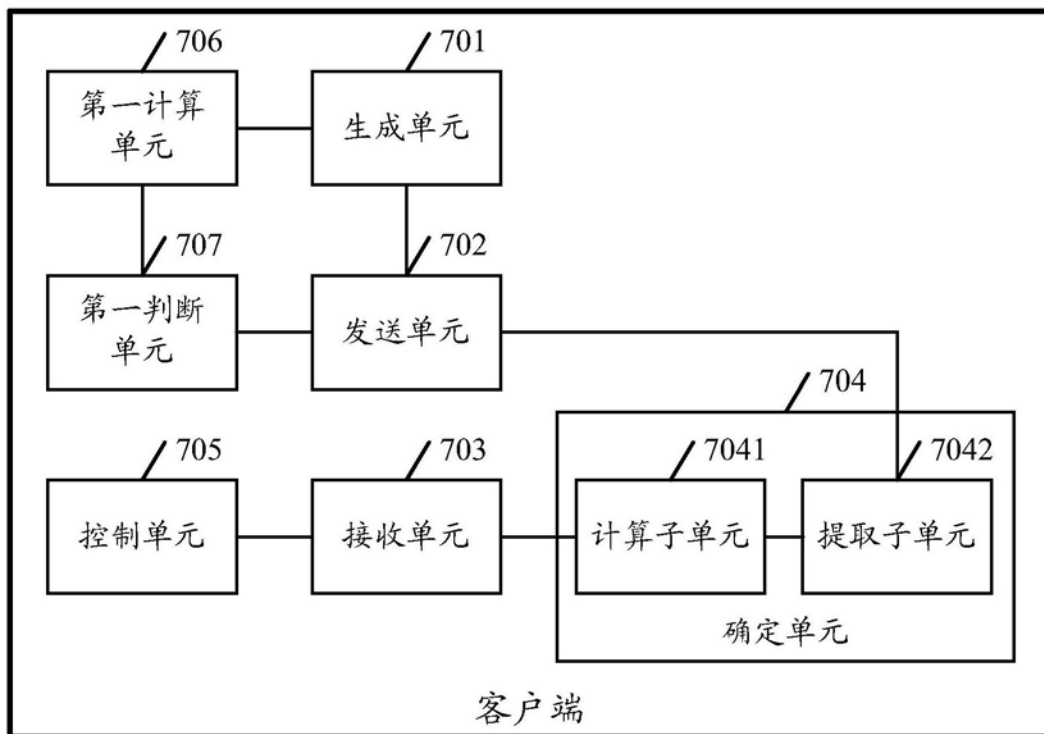


图9

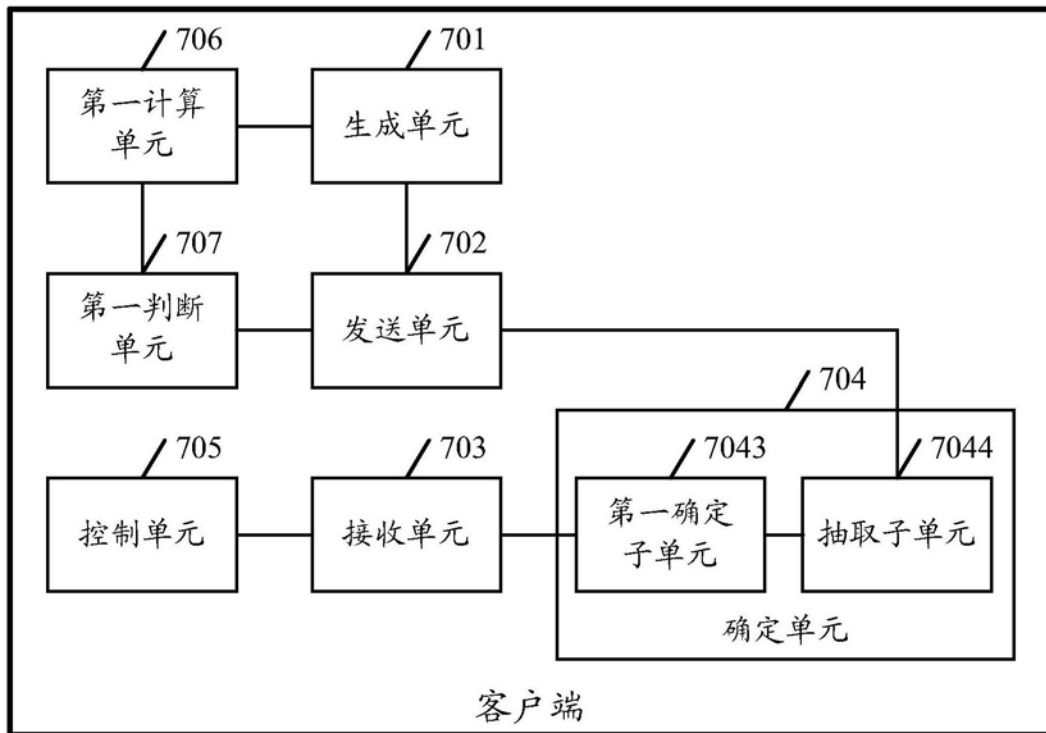


图10

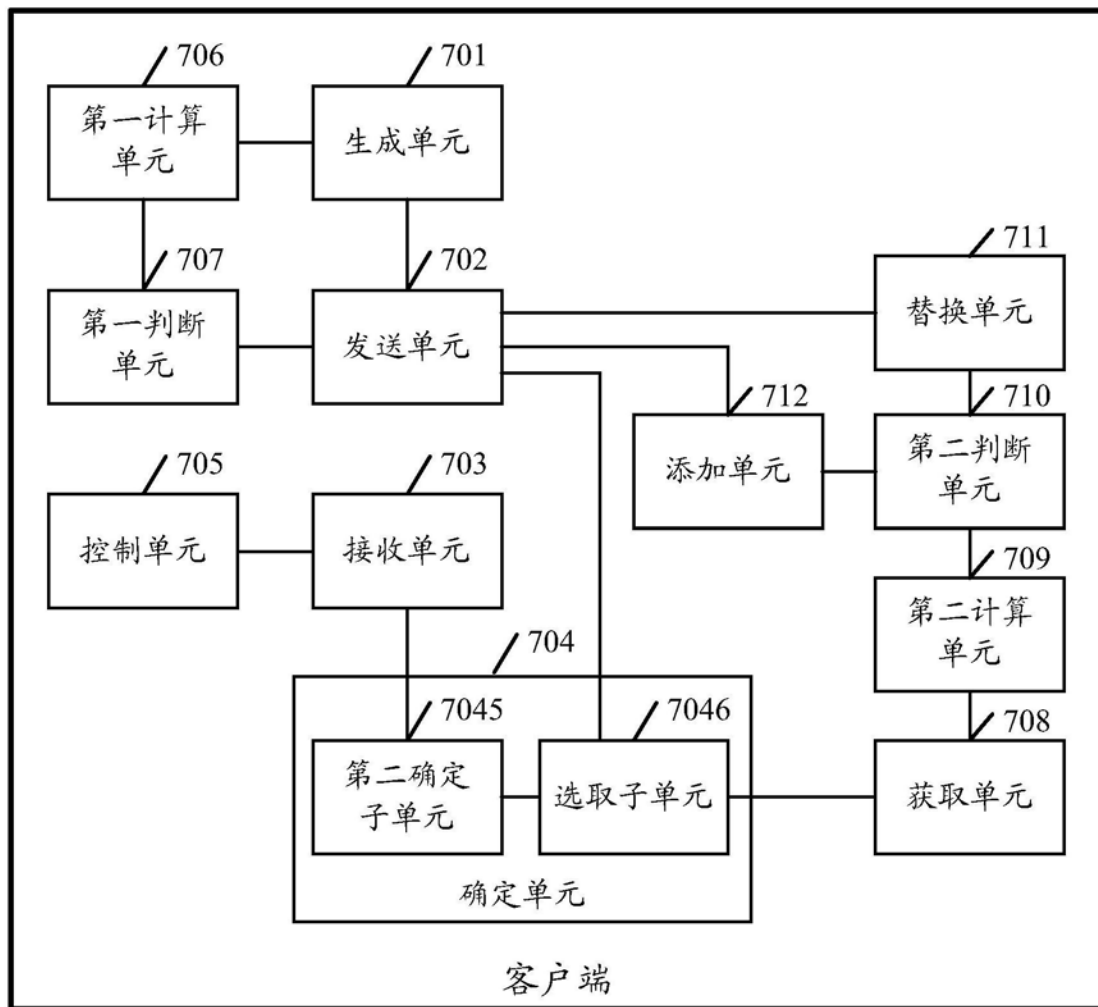


图11

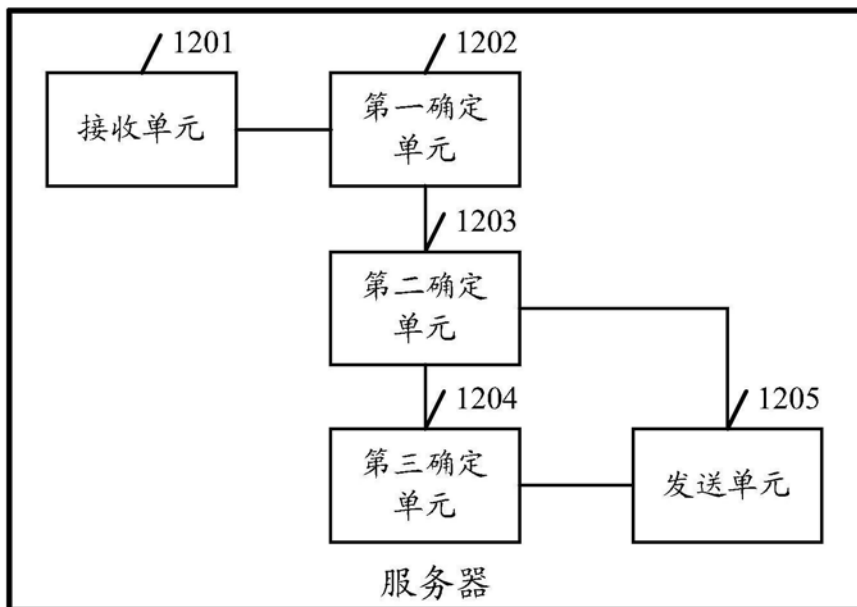


图12

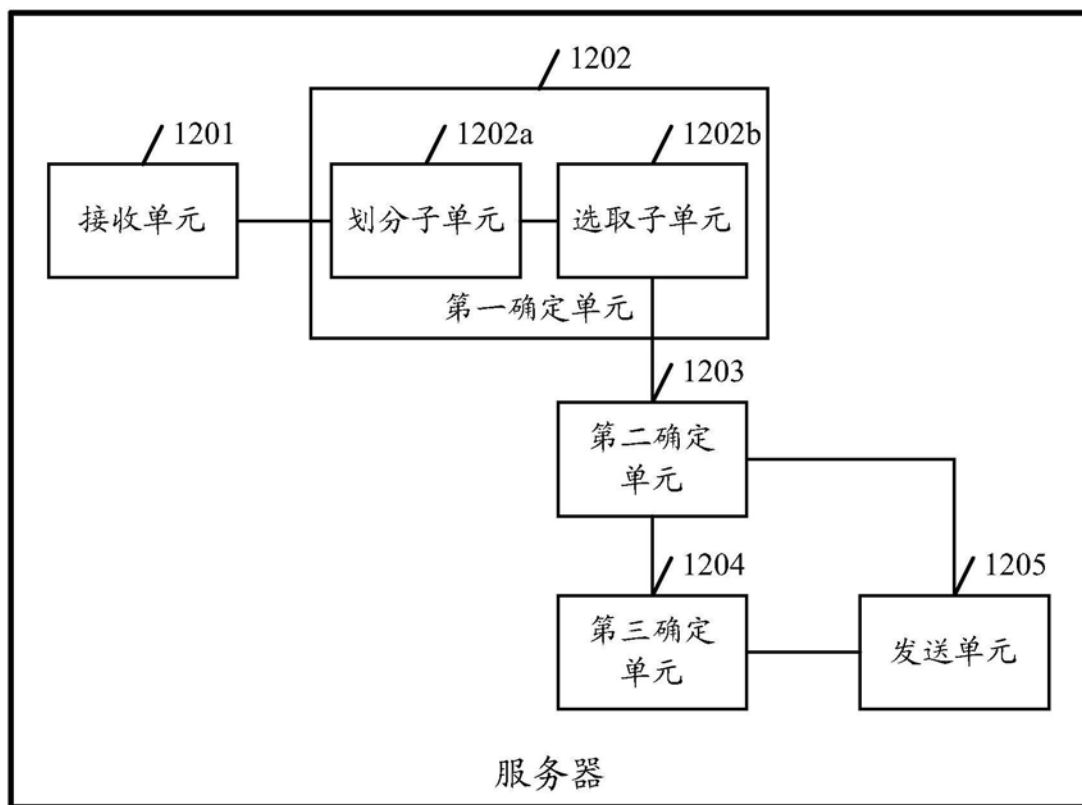


图13

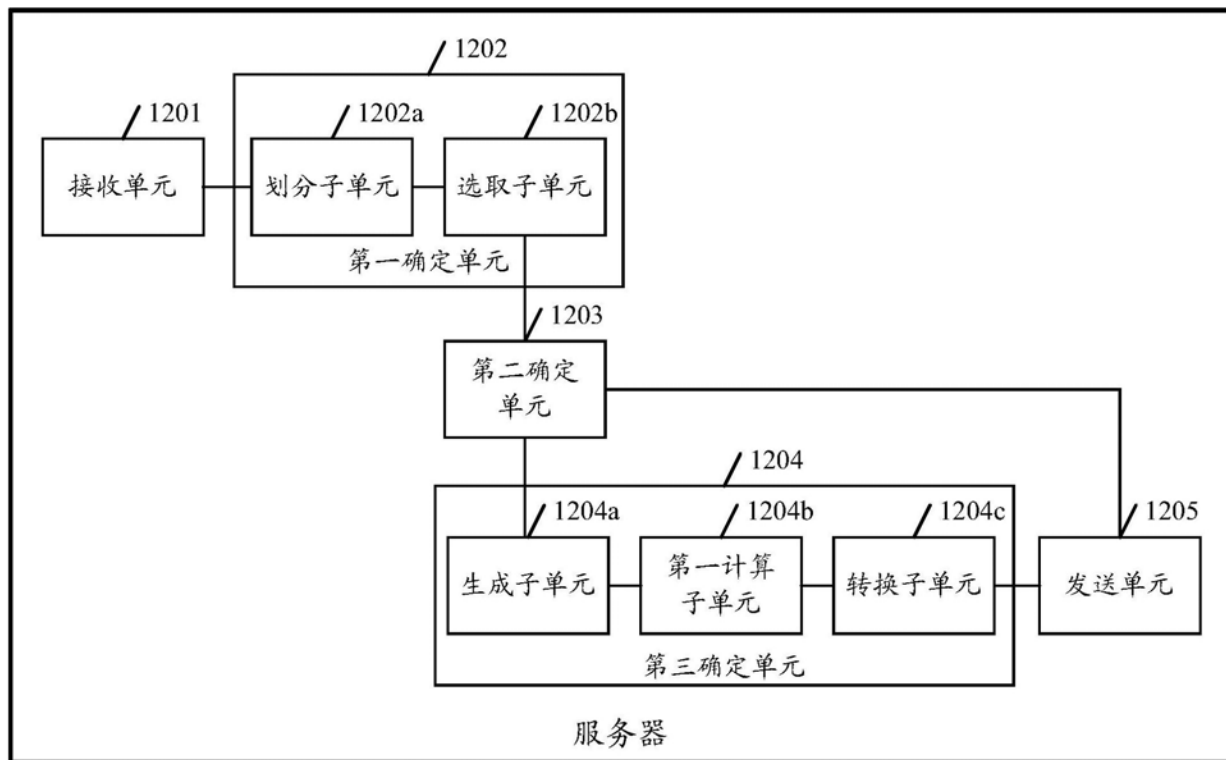


图14

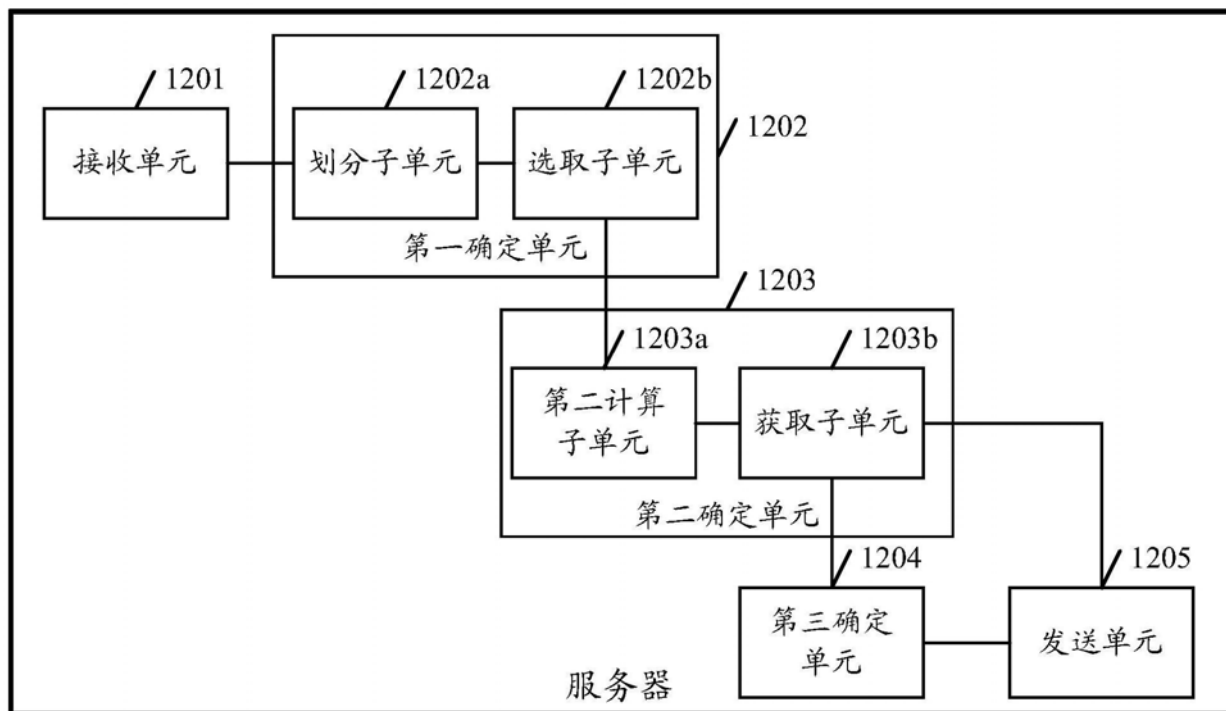


图15

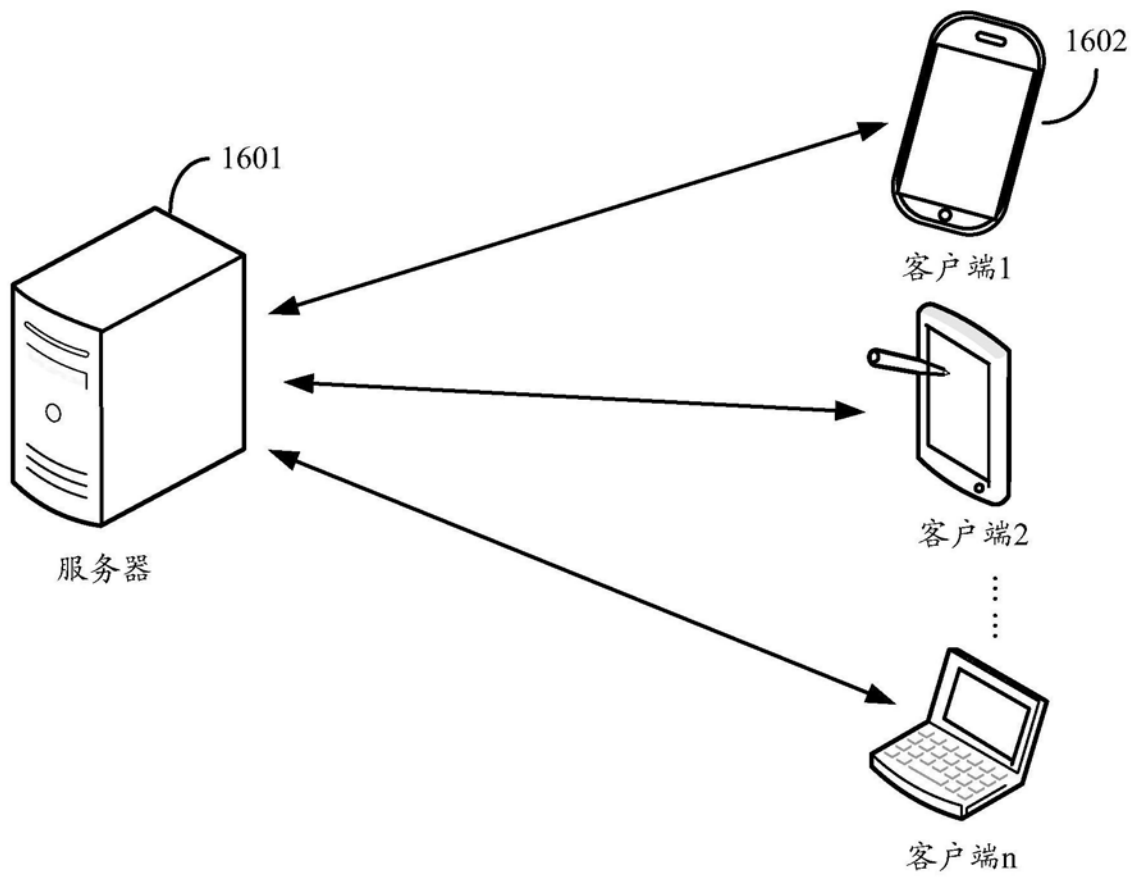


图16