(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106599922 A
(43)申请公布日 2017.04.26

(21)申请号 201611165253.X

(22)申请日 2016.12.16

(71)申请人 中国科学院计算技术研究所
地址 100190 北京市海淀区中关村科学院南路6号

(72)发明人 陈益强 王晋东 沈建飞 胡春雨 王记伟 张宇欣 忽丽莎

(74)专利代理机构 北京泛华伟业知识产权代理有限公司 11280
代理人 王勇 苏晓丽

(51)Int.Cl.
G06K 9/62(2006.01)

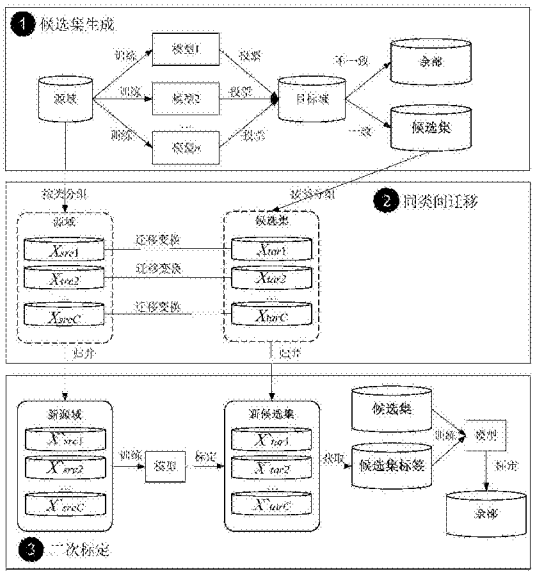
权利要求书1页 说明书6页 附图4页

(54)发明名称

用于大规模数据标定的迁移学习方法及系统

(57)摘要

本发明提供一种迁移学习方法，该方法利用基于已标定的源域数据训练的至少两个分类器对待标定的目标域数据进行初次标定，根据标定结果将目标域数据划分为候选集和余部；在具有相同标定的源域数据组和候选集中目标域数据组之间进行迁移变换，生成新源域和新候选集；基于在新源域上训练的分类器对新候选集中的目标域数据进行标定，并利用新候选集中各数据的标定结果更新对未经变换的候选集中各数据的二次标定；以及基于经更新标定后的候选集训练分类器，并利用该分类器完成对余部中目标数据的标定。该方法缩短了迁移的时间，提高了迁移标定的效率，更适用于大规模数据的标定。



CN 106599922 A

1.一种迁移学习方法,包括:

步骤a)利用基于已标定的源域数据训练的至少两个分类器分别对待标定的目标域数据进行标定,将至少两个分类器的标定结果相同的目标域数据添加到候选集,其余目标域数据构成余部;

步骤b)对于源域数据和候选集的目标域数据,分别将数据按其标定进行分组,将具有相同标定的源域数据组和目标域数据组变换至同一空间使得变换后的源域数据组和目标域数据组满足相同分布,并将变换后得到的各源域数据组和目标域数据组分别归并成新源域和新候选集;

步骤c)基于在新源域上训练的分类器对新候选集中的目标域数据进行标定,并利用新候选集中各数据的标定结果更新对未经变换的候选集中各数据的标定;

步骤d)基于经更新标定后的候选集训练分类器,并利用该分类器完成对余部中目标数据的标定。

2.根据权利要求1所述的方法,步骤a)包括基于已标定的源域数据的相同特征来训练至少两个分类器。

3.根据权利要求1所述的方法,步骤a)包括基于已标定的源域数据的不同特征来训练至少两个分类器。

4.根据权利要求1所述的方法,在步骤b)采用下列方法中的一个来对源域数据组和目标域数据组进行变换:迁移成分分析方法、测地学流式核方法、谱特征对齐方法。

5.根据前述任一权利要求所述的方法,所述分类器选自下列中的一个或多个:支持向量机、随机森林、决策树。

6.一种迁移学习系统,包括:

候选集生成装置,用于利用基于已标定的源域数据训练的至少两个分类器分别对待标定的目标域数据进行标定,将至少两个分类器的标定结果相同的目标域数据添加到候选集,其余目标域数据构成余部;

同类迁移装置,用于对于源域数据和候选集的目标域数据,分别将数据按其标定进行分组,将具有相同标定的源域数据组和目标域数据组变换至同一空间使得变换后的源域数据组和目标域数据组满足相同分布,并将变换后得到的各源域数据组和目标域数据组分别归并成新源域和新候选集;

候选集标定装置,用于基于在新源域上训练的分类器对新候选集中的目标域数据进行标定,并利用新候选集中各数据的标定结果更新对未经变换的候选集中各数据的标定;

余部标定装置,用于基于经更新标定后的候选集训练分类器,并利用该分类器完成对余部中目标数据的标定。

7.根据权利要求6所述的系统,其中所述至少两个分类器为基于已标定的源域数据的相同的特征来训练的。

8.根据权利要求6所述的系统,其中所述至少两个分类器为基于已标定的源域数据的不同的特征来训练的。

9.根据权利要求6所述的系统,所述同类迁移装置采用下列方法中的一个来对源域数据组和目标域数据组进行变换:迁移成分分析方法、测地学流式核方法、谱特征对齐方法。

用于大规模数据标定的迁移学习方法及系统

技术领域

[0001] 本发明涉及机器学习、迁移学习及数据标定，尤其涉及用于不同数据分布下的迁移学习方法。

背景技术

[0002] 不同数据分布下的大规模数据标定是机器学习领域的一个热点问题。随着大数据时代的到来，可穿戴计算领域产生了大量的人群行为、交通模式、生活数据、健康、办公、医疗等各个方面的用户数据。尽管这些数据可以很容易地被获取到，但是它们往往都以无标定的形态出现，即，通常我们只能获取用户的数据特征，却不知道数据特征和具体行为的对应关系。并且，可获取的数据通常也具有不同的性质：或者具有不同的数据特征维度，或者具有不同的特征分布，又或者具有不同的行为类别。机器学习方法是解决数据分类和数据标定问题的常用手段。传统的机器学习方法利用带有标签的样本数据训练相关的分类器模型来实现对测试数据的标签标定，但其假定样本数据与测试数据均属于同一种数据分布。而在大数据环境中，由于这些数据分布的高动态性和高差异性，传统的机器学习方法并不能很好地进行不同数据分布下的数据标定。

[0003] 近年来，迁移学习受到了越来越多的关注，其可以利用已知领域中有标签的训练样本(可称为源域数据)训练分类模型来对目标领域的数据(可称为目标域数据)进行标定，而并不要求源域和目标域数据具有相同的数据分布。迁移学习实际上是通过找寻待标定数据和已知标签数据之间的联系，例如采用核函数的方式将源域和目标域数据映射到同一空间中，在该空间下源域数据和目标域数据拥有相同的分布，从而可以利用该空间表示的有标签的源域样本数据训练分类器来对目标领域进行标定。然而，传统的迁移学习方法计算复杂度高，并不适用于大数据环境下的数据标定。

发明内容

[0004] 因此，本发明的目的在于克服上述现有技术的缺陷，提供一种新的迁移学习方法，实现对不同数据分布下的大规模数据的快速标定。

[0005] 本发明的目的是通过以下技术方案实现的：

[0006] 一方面，本发明提供了一种迁移学习方法，包括：

[0007] 步骤a)利用基于已标定的源域数据训练的至少两个分类器分别对待标定的目标域数据进行标定，将至少两个分类器的标定结果相同的目标域数据添加到候选集，其余目标域数据构成余部；

[0008] 步骤b)对于源域数据和候选集的目标域数据，分别将数据按其标定进行分组，将具有相同标定的源域数据组和目标域数据组变换至同一空间使得变换后的源域数据组和目标域数据组满足相同分布，并将变换后得到的各源域数据组和目标域数据组分别归并成新源域和新候选集；

[0009] 步骤c)基于在新源域上训练的分类器对新候选集中的目标域数据进行标定，并利

用新候选集中各数据的标定结果更新对未经变换的候选集中各数据的标定；

[0010] 步骤d)基于经更新标定后的候选集训练分类器，并利用该分类器完成对余部中目标数据的标定。

[0011] 在上述方法中，步骤a)可包括基于已标定的源域数据的相同特征来训练至少两个分类器。

[0012] 在上述方法中，步骤a)可包括基于已标定的源域数据的不同特征来训练至少两个分类器。

[0013] 在上述方法中，在步骤b)可采用下列方法中的一个来对源域数据组和目标域数据组进行变换：迁移成分分析方法、测地学流式核方法、谱特征对齐方法。

[0014] 在上述方法中，所述分类器可以选自下列中的一个或多个：支持向量机、随机森林、决策树。

[0015] 又一方面，本发明提供了一种迁移学习系统，包括：

[0016] 候选集生成装置，用于利用基于已标定的源域数据训练的至少两个分类器分别对待标定的目标域数据进行标定，将至少两个分类器的标定结果相同的目标域数据添加到候选集，其余目标域数据构成余部；

[0017] 同类迁移装置，用于对于源域数据和候选集的目标域数据，分别将数据按其标定进行分组，将具有相同标定的源域数据组和目标域数据组变换至同一空间使得变换后的源域数据组和目标域数据组满足相同分布，并将变换后得到的各源域数据组和目标域数据组分别归并成新源域和新候选集；

[0018] 候选集标定装置，用于基于在新源域上训练的分类器对新候选集中的目标域数据进行标定，并利用新候选集中各数据的标定结果更新对未经变换的候选集中各数据的标定；

[0019] 余部标定装置，用于基于经更新标定后的候选集训练分类器，并利用该分类器完成对余部中目标数据的标定。

[0020] 在上述系统中，所述至少两个分类器可以是基于已标定的源域数据的相同的特征来训练的。

[0021] 在上述系统中，所述至少两个分类器可以是基于已标定的源域数据的不同的特征来训练的。

[0022] 在上述系统中，所述同类迁移装置可以采用下列方法中的一个来对源域数据组和目标域数据组进行变换：迁移成分分析方法、测地学流式核方法、谱特征对齐方法。

[0023] 与现有技术相比，本发明的优点在于：

[0024] 基于源域已有知识对目标域部分数据进行分类，在同类型的源域数据和目标域数据之间进行迁移，而并非直接将所有的源域和目标域进行迁移；这样可以大大缩短迁移的时间，提高了迁移标定的效率，更适用于大规模数据的标定。

附图说明

[0025] 以下参照附图对本发明实施例作进一步说明，其中：

[0026] 图1为根据本发明实施例的迁移学习方法的过程示意图；

[0027] 图2为用于进行跨位置行为识别实验的位置示意图；

[0028] 图3(a)和图3(b)为根据本发明实施例的方法与现有方法识别精度对比示意图；

[0029] 图4(a)和图4(b)为根据本发明实施例的方法与现有方法识别时间效率对比示意图。

具体实施方式

[0030] 为了使本发明的目的，技术方案及优点更加清楚明白，以下结合附图通过具体实施例对本发明进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。

[0031] 在迁移学习中，通常将已有标签的数据称为源域，待标定的数据称为目标域。图1给出了根据本发明一个实施例的用于不同分布下大规模数据标定的分层迁移学习方法的过程示意图。如图1所示。该方法主要包括三个步骤：候选集生成、同类间迁移以及二次标定。在候选集生成阶段，利用有标签的源域数据学习几个独立的分类器，利用大多数投票机制，利用这些分类器分别对待标定的目标域数据进行标定，将各分类器的标定结果都相同(即投票结果一致)的那些目标域数据称之为候选集，其余目标域数据称为余部，同时，保留这些分类器给候选集中数据标定的标签；在同类间迁移阶段，对于源域数据和候选集数据，分别根据它们的标签进行分组，将相同标定的数据分成一个组，然后对来自同一个类别的源域数据和候选集数据，迁移变换至同一空间使得变换后的源域数据和目标域数据满足相同分布，这样对每个类别均生成新的源域数据和候选集数据，将它们归并生成新源域和新候选集；在二次标定阶段，在新源域数据上训练一个分类器，利用该分类器来标定新候选集数据，即给新候选集中的数据添加相关的标签，接着，用此新候选集中所有标签和旧的候选集数据训练一个分类器，用来标定余部数据。至此，可以实现对全部目标域数据的标定。下面将分别对上述步骤进行详细介绍。

[0032] 1、候选集生成。

[0033] 首先在有标定(即已经带有标签)的源域数据上，训练多个不同的分类器。在训练好分类器之后，分别用每个分类器来对待标定的目标域数据进行标定。对于各个分类器的标定结果，采用大多数投票的准则，大多数分类器取得一致的那些目标域数据组成候选集，用于后续的迁移；而其余的目标域数据统称为余部，等待下一步的标定。例如，对于每个目标域数据，如果全部或大部分的分类器为其标定的标签都相同，则将该目标域数据加入候选集，同时记录该目标域数据对应的标签，否则将其加入余部中。这里为候选集中目标域数据初步标定的标签实际上并不很准确，因此可以将其称为伪标签。

[0034] 在本发明的实施例中不对所采用的分类器模型及其数量进行限制，可以根据实际需求和系统资源状况采用各种数量和类型的分类器模型，例如支持向量机、随机森林、决策树等常用的分类器。并且在训练时可以用不同特征、不同种类的分类器来训练源域数据。另外，在基于各个分类器的标定结果对目标域数据划分候选集和余部时，除了多数投票机制之外，也可以采用如权重投票、打分投票等投票机制。

[0035] 2、同类间迁移。

[0036] 在本发明的实施例中，同类间迁移是指相同类别的源域数据和候选集中目标域数据之间的迁移变换，通过迁移变换将不同分布下的源域和目标域数据变换到同一空间，使得在该空间中的源域和目标域满足相同的数据分布。这样，对每个类别均会生成新的源域

数据和候选集数据，将它们归并生成新源域和新候选集。

[0037] 更具体地，对于源域数据和候选集数据，可以根据各自标签进行分组，将具有相同标签(即相同标定，属于相同的类型)的数据分成一个组，然后在来自同一个类别的源域数据和候选集数据之间进行迁移变换。例如，可以通过迁移成分分析(Transfer Component Analysis，TCA)方法将不同分布下的源域和目标域数据变换到相同的重构希尔伯特空间(reproducing kernel Hilbert space，RKHS)，在此空间中最小化两个域的距离并最大限度地保留它们各自的内部特征，从而使得新空间中的源域和目标域满足相同的数据分布。假定源域和目标域一共有C个类别，则将它们分成对应的C组。对于每个组的源域数据和目标域数据，利用TCA进行迁移变换。以X_{src}和X_{tar}来分别表示源域和目标域中的数据，X_{src_j}和$X_{tar_j}(j=1,2,\cdots,C)$分别表示分组后的源域和候选集中的数据，则有

[0038] $$X_{src}=\begin{bmatrix}X_{src_1}\\X_{src_2}\\\cdots\\X_{src_C}\end{bmatrix},X_{tar}=\begin{bmatrix}X_{tar_1}\\X_{tar_2}\\\cdots\\X_{tar_C}\\X_{residual}\end{bmatrix}$$

[0039] 其中$X_{candidates}$表示候选集，$\{X_{tar_1},X_{tar_2},\cdots,X_{tar_C}\}=X_{candidates}$；而$X_{residual}$表示余部。用$X'_{src_j}$和$X'_{tar_j}$分别表示经过TCA后的源域和候选集数据，则有：

[0040] $$[X'_{src_j},X'_{tar_j}]=TCA(X_{src_j},X_{tar_j}),j=1,2,\cdots,C$$

[0041] 最后，将每个类别对应生成新的源域数据和候选集数据进行归并，得到新的具有相同分布的源域数据和候选集数据，以X'_{src}和X'_{tar}进行表示：

[0042] $$X'_{src}=\begin{bmatrix}X'_{src_1}\\X'_{src_2}\\\cdots\\X'_{src_C}\end{bmatrix},X'_{tar}=\begin{bmatrix}X'_{tar_1}\\X'_{tar_2}\\\cdots\\X'_{tar_C}\end{bmatrix}$$

[0043] 应指出，在其他示例中，也可以利用除了TCA之外的其它的迁移变换方法，例如测地学流式核方法(Geodesic flow kernel，GFK)、谱特征对齐(spectral feature alignment，SFA)方法等，将原来处于不同分布的源域和目标域数据变换到一个新空间中，以使得在新空间中的源域和目标域满足相同的数据分布。并且在同类间迁移阶段，可以逐个类别顺序地迁移，也可以利用并行算法实现各类型同时并行迁移。

[0044] 3、二次标定

[0045] 经过上述同类迁移之后，原来的源域数据和原候选集数据被变换到同一个空间，在新空间中以新的形式表达的源域和目标域数据可以称为新的源域数据和新候选集。在该新源域数据上训练一个分类器，利用训练好的分类器对新候选集中的数据进行标定，即识别这些数据属于源域中的哪种类型或哪个标签，并根据识别结果来给新候选集中数据标注相应的标签，这时生成的标签要比之前生成的伪标签更准确。然后，利用新候选集中各数据对应的标签来更新最初生成的原候选集中相应数据的伪标签，从而完成对原候选集中数据的第二次标定。这是因为新候选集只是原候选集的一些形式变换，样本的顺序并没有改变，因此，原来的候选集中的数据实际上也获得了相应的标签。

[0046] 接着，再利用经更新标签之后的原候选集数据训练一个分类器，利用该训练好的分类器完成对余部中数据的标定。至此，完成了全部目标域数据的标定。

[0047] 这里，在新源域上训练的分类器和在更新标签之后的候选集上训练的分类器时均可以根据实际的需要进行选择，可以使用相同或不同的分类器模型。

[0048] 与传统的迁移学习直接在所有的源域数据和目标域数据上进行迁移变换相比，根据本发明实施例的方法首先基于源域知识对目标域进行初步分类，选取部分候选集，然后通过在每个类别的源域数据和候选集之间进行迁移变换来得到具有相同分布的源域和候选集数据，这会降低迁移变换的计算复杂度，大大缩短迁移的时间。

[0049] 为了说明根据本发明实施例的迁移学习方法(下文简称为分层迁移学习方法)的时间优势，在这里将其与传统利用TCA直接对源域和目标域数据进行迁移学习方法(下文简称为TCA方法)所需的时间对比。用p表示经过多数据投票后，候选集占总目标域的比例，S_i，T_i($i=1,2,\cdots,C$)分别表示源域和目标域中每个类的样本个数。TCA方法的时间复杂度为$O(m(n_1+n_2)^2)$，而分层迁移学习方法的时间复杂度为$O\left(m\sum_{i=1}^{C}(S_i+T_i)^2\right)$。用如下公式比较它们的时间复杂度：

[0050]
$$ratio=\frac{O\left(m\sum_{i=1}^{C}(S_i+T_i)^2\right)}{O(m(n_1+n_2)^2)}$$
$$=\frac{O\left(p^2\sum_{i=1}^{C}(S_i+T_i)^2\right)}{O\left(\left(p\sum_{i=1}^{C}S_i+\sum_{i=1}^{C}T_i\right)^2\right)}$$
$$\leqslant\frac{O\left(\sum_{i=1}^{C}(S_i+T_i)^2\right)}{O\left(\left(\sum_{i=1}^{C}S_i+\sum_{i=1}^{C}T_i\right)^2\right)}$$

[0051] 从该公式可以看出，分层迁移学习方法与传统的TCA方法的时间复杂度之比(即ratio)永远小于1，可见本发明的分层迁移学习方法较传统的TCA方法更为高效。而且从该公式还可以看出，这个比值ratio与p无关，这说明在最初选取候选集时进行多数投票时的精确度以及候选集中数据量的多少对该分层迁移学习方法的效率本身没有直接的影响。特别地，从上述时间复杂度对比公式可以看出，当源域和目标域中的每个类样本个数近似相等时，可以进一步化简为ratio≤1/C，表示根据本发明的分层迁移学习方法的时间复杂度只是传统TCA方法的1/C，这无疑表明根据本发明的分层迁移学习方法可以以大大缩短迁移的时间，提高了迁移标定的效率，更适用于不同分布下大规模数据的标定。

[0052] 为了进一步验证根据本发明的实施例的分层迁移学习方法的有效性，发明人还在行为识别领域一个公开的数据集上进行了实验。行为识别是通过收集加速度、陀螺仪、无线信号等一些信号来对人体走路、跑步等相应的行为进行识别与预测的研究领域，是可穿戴计算的重要组成部分。所采用的数据集来自加州大学尔湾分校，数据集地址为http://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities，其中包含8个人做19个类别行为的加速度、陀螺仪和磁力计数据。这三种传感器分别放置于每个人身体的5个部位(躯干、左臂、右臂、左腿、右腿)，见图2所示。图2给出了用于在此数据集上进行跨位置行

为识别的位置示意。跨位置为识别是指当身体的一个部位有对应的行为数据和标记时，如何利用此部位的有标记数据来识别同一个人身体的另一个位置的行为。跨位置行为识别在可穿戴计算中属于重要研究问题之一。因为可穿戴设备的位置不可能永远处于固定状态，由此导致了识别模型必须是动态变化的。在下面实验中，任意取一位置，假设其是有标记数据(源域)，然后，针对余下的4个位置(目标域)，分别由这个有标记的位置的数据，对剩下位置进行标记。评价跨位置行为识别实验的标准是识别的精度。也就是由源域数据对目标域数据进行标记后，其标记与原有的目标域标记进行对比，正确的标记所占的比例越大，则说明识别精度越高，表示模型越好。

[0053] 为了便于对比分析，在实验中选用现有的非迁移学习中的主成分分析(principal component analysis，PCA)方法和现有的迁移学习中的迁移成分分析(transfer component analysis，TCA)方法与根据本发明实施例的分层迁移学习方法进行对比。图3(a)示出了参与实验的8个实验对象都统一用右臂的有标记数据来标记余下4个身体部分时的整体精度。图3(b)示出了每个实验对象在由右臂标记左臂时的识别精度。从图3(a)和3(b)中可以明显地看出，根据本发明实施例的分层迁移学习方法在识别精度上要优于现有的主成分分析方法和迁移成分分析方法。

[0054] 如上文对时间复杂度的理论分析部分指出的，根据本发明实施例的分层迁移方法与现有迁移成分分析方法相比，在时间上也有优势。发明人在实验中也验证了这一点，如图4所示。图4(a)示出的是用右臂标记左臂时，在每个实验对象所用时间；图4(b)示出了采用根据本发明实施例的分层迁移方法和现有迁移成分分析方法完成所有位置的标记所用的平均时间。从图4(a)和图4(b)可以明显地看出，根据本发明实施例的分层迁移学习方法比现有的迁移成分分析方法在时间效率上要高出很多。

[0055] 虽然本发明已经通过优选实施例进行了描述，然而本发明并非局限于这里所描述的实施例，在不脱离本发明范围的情况下还包括所做出的各种改变以及变化。

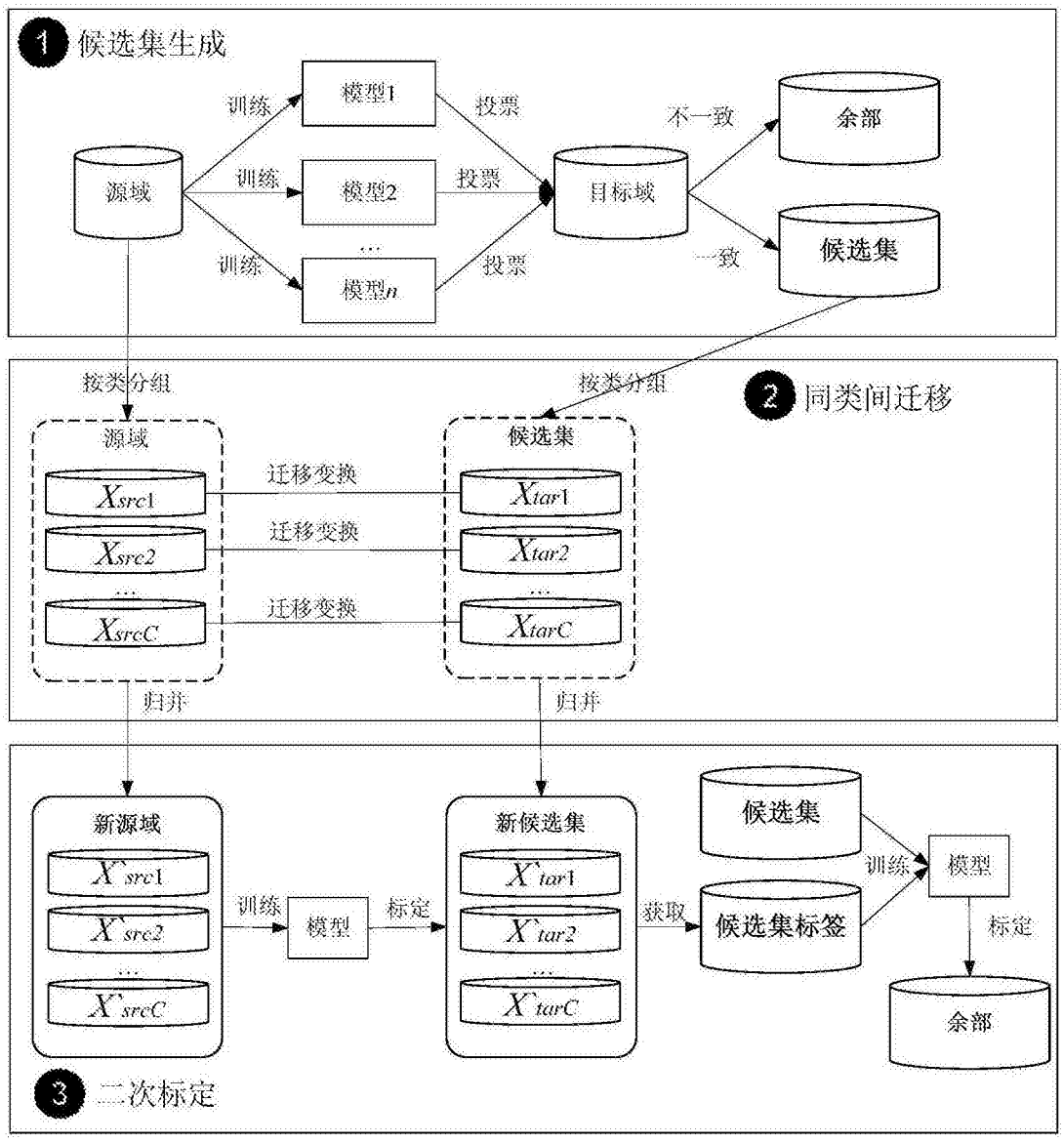

1 候选集生成
源域
训练
模型1
模型2
模型n
投票
目标域
不一致
余部
一致
候选集
按类分组
2 同类间迁移
迁移变换
归并
新源域
新候选集
候选集标签
模型
标定
获取
3 二次标定

图1

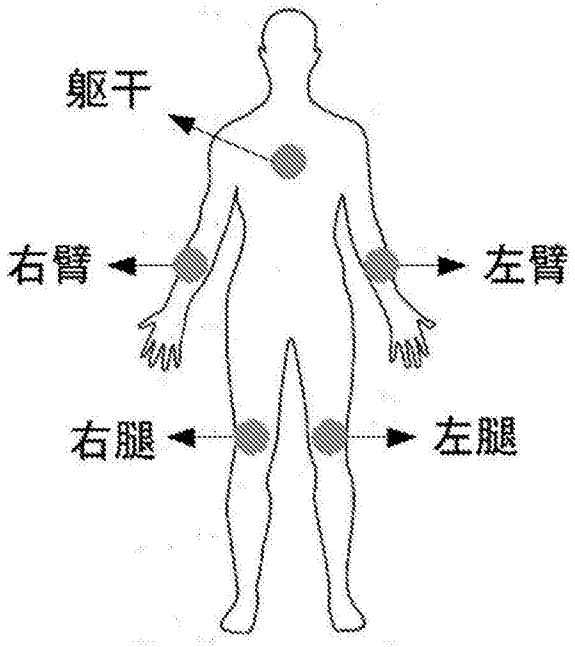

躯干
右臂
左臂
右腿
左腿

图2

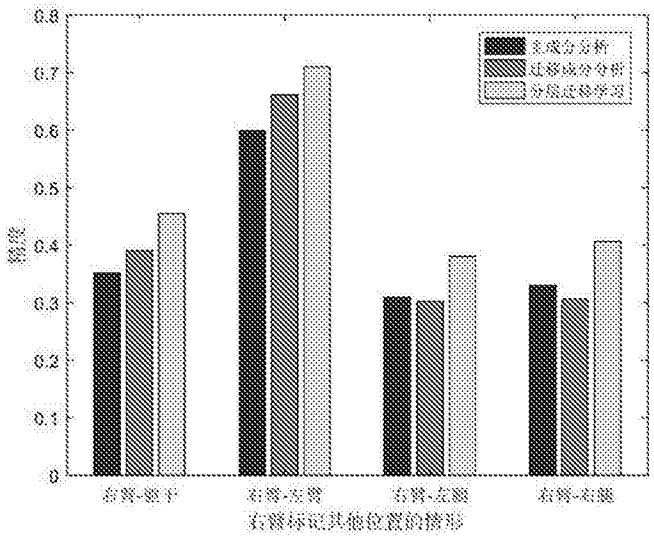

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
右臂-躯干
右臂-左臂
右臂-左腿
右臂-右腿
右臂标记其他位置的情形

图3(a)

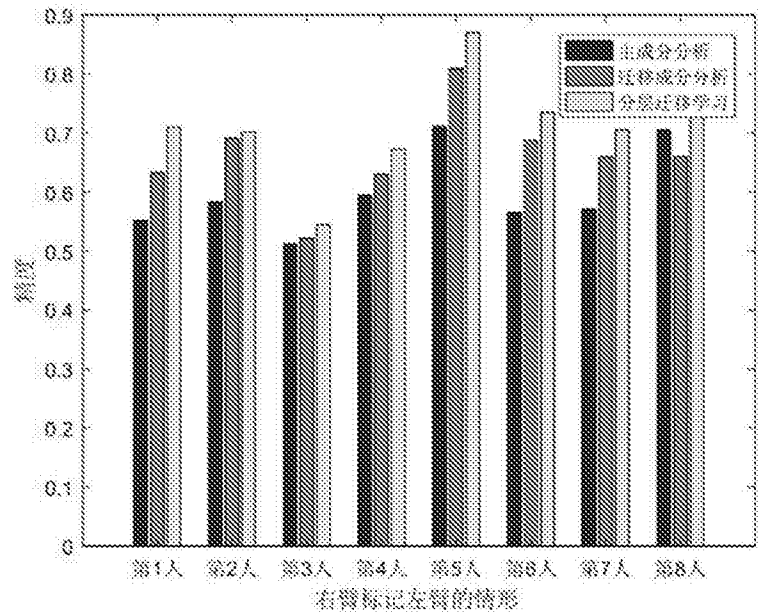

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
精度
主成分分析
近邻成分分析
分层迁移学习
第1人
第2人
第3人
第4人
第5人
第6人
第7人
第8人
右臂标记左臂的情形

图3 (b)

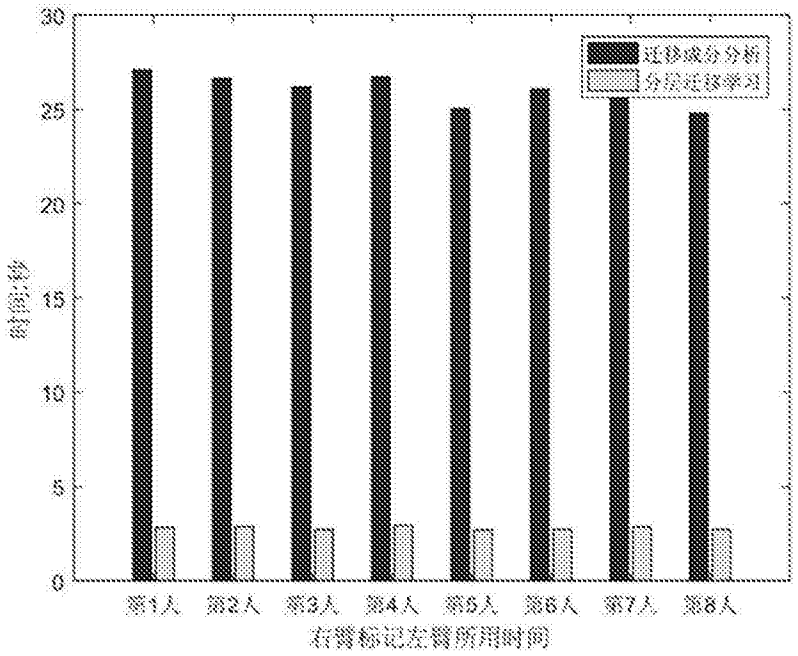

30
25
20
15
10
5
0
时间/秒
近邻成分分析
分层迁移学习
第1人
第2人
第3人
第4人
第5人
第6人
第7人
第8人
右臂标记左臂所用时间

图4 (a)

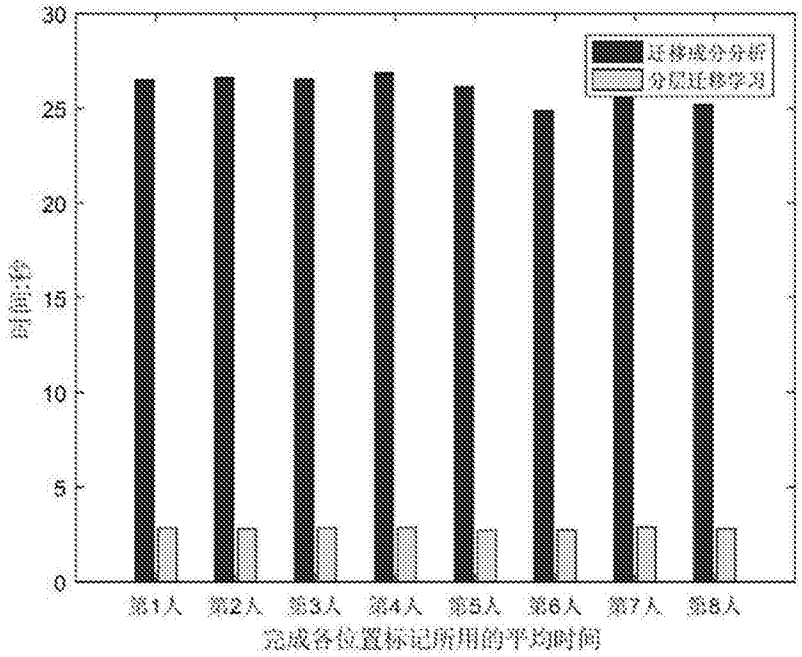

30
25
20
15
10
5
0
迁移成分分析
分区迁移学习
第1人
第2人
第3人
第4人
第5人
第6人
第7人
第8人
完成各位置标记所用的平均时间

图4 (b)