基于迁移学习和半监督学习结合的跨领域分类软件 V1.0

文档说明

1 主要功能：

(1)输入模块

用户可以根据需要，自行设定算法的输入，包括源领域训练数据、目标领域训练数据以及目标领域的未标注数据。同时还可以输入文件所在的根目录，算法的迭代次数和实例选择的个数等。

(2) 算法模块

此模块主要是 TPTSVM 算法的核心模块，用来实现对大量源领域的训练数据、少量目标领域的训练数据和大量目标领域未标注数据的学习，训练生成跨领域分类器，对测试数据进行分类。并按照精确率、召回率的评价标准，给出精确率、召回率和 F 值的相关信息。

(3) 初始化模块

本模块主要将输入模块的数据进行初始化，包括对数据的拷贝，结果文件的删除以及计算 β 值（$\beta=\frac{1}{1+\sqrt{2\ln s/T}}$， s 是源领域训练数据的实例个数，T 是初算法迭代的次数）等操作。

(4) 输出模块

本模块主要对于训练和测试阶段的中间文件进行输出,包括新生成的训练数据和测试数据已经训练得到的模型文件，最后的结果文件等均通过该模块进行输出。

2. 安装过程说明

本软件是绿色软件，无须安装。运行该软件需要电脑有 VC++6.0 或以上版本的编译环境支持。

3. 使用说明

代码主要包括输入模块和输出模块，下面将从两个方面详细介绍软件的使用情况。

3.1 输入模块

用户输入模块主要包括四个部分：源领域标注数据、目标领域标注数据、目标领域未标注数据（即目标领域的测试数据）、中间结果输出路径。在 PTSVM.cpp 下 input 函数里进行初始化和设置。Input 函数的具体内容和使用说明如下：

3.1.1 input 函数说明

变量名称	使用说明
source_file_path_input	源领域的标注数据文件
train_file_path_input	目标领域标注数据文件
test_file_path_input	目标领域未标注数据文件
svm_pre_tenstep_path	中间结果输出路径
iter_N	算法的迭代次数
semi_add_M	每次迭代从未标注数据中选取的可信样例数
weight_test_max	权重向量初始化，默认是 1.0
tempDir	临时目录路径，默认值“D:\\PTSVMtemp”

3.2 输出模块

输出模块主要将训练和测试阶段的中间文件以及训练得到的模型和测试的结果文件进行输出并保存。

各输出文件具体内容说明如下：

输出文件	说明
source_test_data_file	源领域训练阶段生成的标注数据文件
train_test_data_file	目标领域训练阶段生成的标注数据文件
test_test_data_file	目标领域训练阶段生成的未标注数据文件
des_test_data_file	目标领域训练阶段生成的标注数据文件
semi_test_data_file	半监督学习输出的测试文件
whole_train_data_file	基于源领域和目标领域训练数据调整权重之后的训练数据文件
des_train_data_file	仅基于目标领域训练数据调整权重之后的训练数据文件
whole_model_file	基于源领域和目标领域训练数据训练获得的模型文件
des_model_file	仅基于目标领域训练数据训练获得的模型文件
source_result_by_whole_file	利用源领域和目标领域训练数据训练获得的模型对源领域数据分类后的结果文件
des_result_by_whole_file	利用源领域和目标领域训练数据训练获得的模型对目标领域训练数据分类后的结果文件
des_result_by_des_file	利用目标领域训练数据训练获得的模型对目标领域训练数据分类后的结果文件
test_result_by_whole_file	利用源领域和目标领域训练数据训练获得的模型对目标领域测试数据分类后的结果文件
test_result_by_des_file	利用目标领域训练数据训练获得的模型对目标领域测试数据分类后的结果文件