LIBSVM格式数据与PAI平台集成

前言：libSVM数据有固定的数据格式，把这个格式稍加修改，然后导入数据到MaxCompute中，可以使用PAI平台算法中的” 是否稀疏数据 k:v,k:v类型特征”功能，直接进行算法训练、预测，也可以使用KV to Table组件转换成Key和Value表数据。

1. 下载LIBSVM：<http://www.csie.ntu.edu.tw/~cjlin/libsvm/#download>

下载完成后，直接解压下载包，$install\_home\libsvm-3.22\windows添加到path路径。

熟悉libsvm使用，可忽略。

1. 训练和测试数据集下载：[***http://www.csie.ntu.edu.tw/%7Ecjlin/papers/guide/data/***](http://www.csie.ntu.edu.tw/%7Ecjlin/papers/guide/data/)

熟悉libsvm使用，可忽略。

1. LIBSVM数据格式：

训练数据和检验数据文件格式如下：

[label] [index1]:[value1] [index2]:[value2] …

[label] [index1]:[value1] [index2]:[value2] …

Label 就是说 class （属于哪一类） , 就是你要分类的种类，通常是一些整数。

index 是有順序的索引，通常是连续的整数。就是指特征编号，必须按照升序排列

value 就是特征值，用来 train 的数据，通常是一堆实数组成。

（注： 修改训练和测试数据的格式为程序可以识别的程序

目标值 第一维特征编号：第一维特征值 第二维特征编号：第二维特征值 …

目标值 第一维特征编号：第一维特征值 第二维特征编号：第二维特征值 …

……

目标值 第一维特征编号：第一维特征值 第二维特征编号：第二维特征值 …

例如： 2.3 1:5.6 2:3.2

表示训练用的特征有两维，第一维是 5.6 ，第二维是 3.2 ，目标值是 2.3

注意：训练和测试数据的格式必须相同，都如上所示。测试数据中的目标值是为了计算误差用 ）

1. LIBSVM数据格式修改导入MaxCompute
   1. 数据修改与导入

**把libSVM数据文件修改为如下格式（PAI算法要求，其中两个key:value之间必须使用逗号分隔，如**key1:value1,key2:value2,……**）**：label key1:value1,key2:value2,……。

创建MaxCompute表，包含两个字段：label\_col,value\_col。

把修改后的数据文件使用tunnel加载到新创建的MaxCompute表中。

* 1. 举例
     1. 格式修改

修改前测试数据格式（文件名test）：

0 1:4.23629800 2:2.19821001 3:-3.50379701 4:9.75216301

0 1:4.23629800 2:2.19821001 3:-3.50379701 4:9.75216301

0 1:1.65701300 2:8.06630101 3:2.12600601 4:6.86683401

0 1:2.80400100 2:2.23530001 3:3.41496101 4:8.09512501

0 1:2.83570100 2:3.02016001 3:3.85592301 4:9.48986101

0 1:1.01129900 2:3.63975101 3:-1.50978301 4:6.57294501

……

修改后测试数据格式（文件名test\_tra），以”\*”分隔label与value：

0\*1:4.23629800,2:2.19821001,3:-3.50379701,4:9.75216301

0\*1:4.23629800,2:2.19821001,3:-3.50379701,4:9.75216301

0\*1:1.65701300,2:8.06630101,3:2.12600601,4:6.86683401

0\*1:2.80400100,2:2.23530001,3:3.41496101,4:8.09512501

0\*1:2.83570100,2:3.02016001,3:3.85592301,4:9.48986101

0\*1:1.01129900,2:3.63975101,3:-1.50978301,4:6.57294501

……

* + 1. MaxCompute建表

create table libsvm\_test(lable\_col bigint, value\_col string);

* + 1. 数据导入

tunnel upload test\_tra libsvm\_test -fd '\*';

1. 登录阿里云数加平台->机器学习，创建PAI任务
   1. 选择读数据表组件，作为数据来源，来源表输入libsvm\_test;



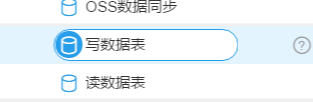
* 1. 选择拆分组件，80%数据用于模型训练，20%数据用于预测；



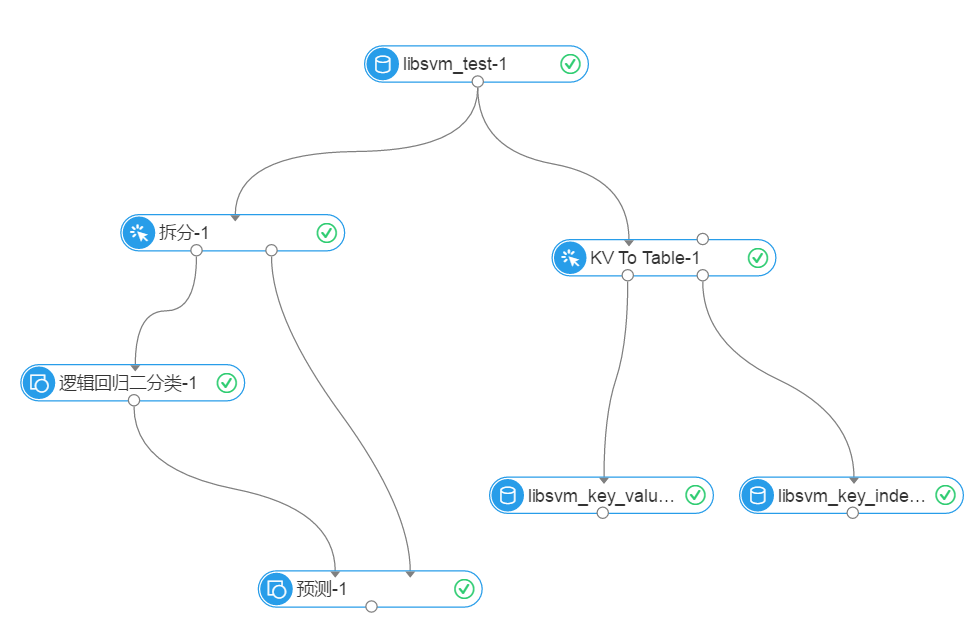
* 1. 选择逻辑回归组件，用于生成逻辑回归分类模型：特征列选择value\_col，目标列选择label\_col，同时选择”是否稀疏数据 k:v,k:v类型特征”;



* 1. 选择预测组件，用于生成预测结果数据；
  2. 选择KV to Table组件，用于Key:Value,Key:Value…字段的转换；
  3. 选择2个写数据表组件，用于输出key值和value值；



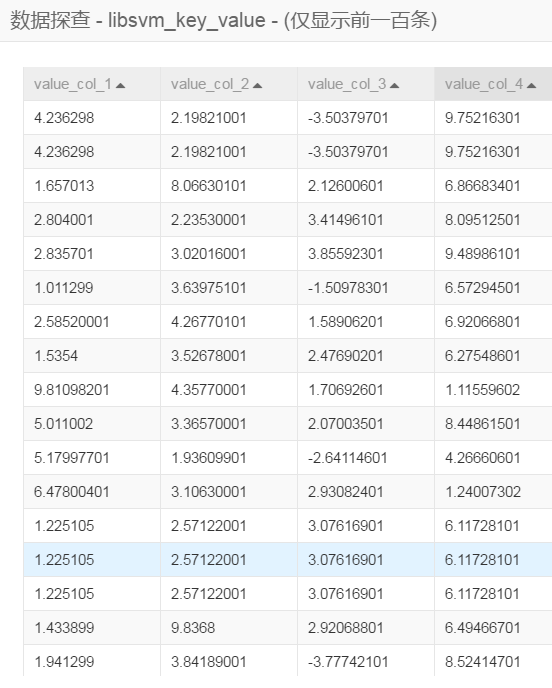
* 1. 整个PAI流程如下：



预测数据：



Value数据：



Key index数据：

